

JAMAevidence

THE RATIONAL CLINICAL EXAMINATION

EVIDENCE-BASED CLINICAL DIAGNOSIS



David L. Simel, MD, MHS • Drummond Rennie, MD

The Rational Clinical Examination

NOTICE

Medicine is an ever-changing science. As new research and clinical experience broaden our knowledge, changes in treatment and drug therapy are required. The authors and the publisher of this work have checked with sources believed to be reliable in their efforts to provide information that is complete and generally in accord with the standards accepted at the time of publication. However, in view of the possibility of human error or changes in medical sciences, neither the authors nor the publisher nor any other party who has been involved in the preparation or publication of this work warrants that the information contained herein is in every respect accurate or complete, and they disclaim all responsibility for any errors or omissions or for the results obtained from use of the information contained in this work. Readers are encouraged to confirm the information contained herein with other sources. For example and in particular, readers are advised to check the product information sheet included in the package of each drug they plan to administer to be certain that the information contained in this work is accurate and that changes have not been made in the recommended dose or in the contraindications for administration. This recommendation is of particular importance in connection with new or infrequently used drugs.

The Rational Clinical Examination

Evidence-Based Clinical Diagnosis

Editors

David L. Simel, MD, MHS

Chief, Medicine Service

Durham Veterans Affairs Medical Center

Professor of Medicine

Duke University School of Medicine

Durham, North Carolina

Drummond Rennie, MD

JAMA

Chicago, Illinois

Philip R. Lee Institute for Health Policy Studies

University of California, San Francisco

San Francisco, California

Editor, Education Guides

Sheri A. Keitz, MD, PhD

Chief, Medicine Service

Miami Veterans Affairs Medical Center

Associate Dean for Faculty Diversity and Development

University of Miami Miller School of Medicine

Miami, Florida



New York Chicago San Francisco Lisbon London
Madrid Mexico City Milan New Delhi San Juan
Seoul Singapore Sydney Toronto



JAMAevidence

Copyright © 2009 by the American Medical Association. All rights reserved. Manufactured in the United States of America. Except as permitted under the United States Copyright Act of 1976, no part of this publication may be reproduced or distributed in any form or by any means, or stored in a database or retrieval system, without the prior written permission of the publisher.

0-07-159031-5

The material in this eBook also appears in the print version of this title: 0-07-159030-7.

All trademarks are trademarks of their respective owners. Rather than put a trademark symbol after every occurrence of a trademarked name, we use names in an editorial fashion only, and to the benefit of the trademark owner, with no intention of infringement of the trademark. Where such designations appear in this book, they have been printed with initial caps.

McGraw-Hill eBooks are available at special quantity discounts to use as premiums and sales promotions, or for use in corporate training programs. For more information, please contact George Hoare, Special Sales, at george_hoare@mcgraw-hill.com or (212) 904-4069.

TERMS OF USE

This is a copyrighted work and The McGraw-Hill Companies, Inc. ("McGraw-Hill") and its licensors reserve all rights in and to the work. Use of this work is subject to these terms. Except as permitted under the Copyright Act of 1976 and the right to store and retrieve one copy of the work, you may not decompile, disassemble, reverse engineer, reproduce, modify, create derivative works based upon, transmit, distribute, disseminate, sell, publish or sublicense the work or any part of it without McGraw-Hill's prior consent. You may use the work for your own noncommercial and personal use; any other use of the work is strictly prohibited. Your right to use the work may be terminated if you fail to comply with these terms.

THE WORK IS PROVIDED "AS IS." McGRAW-HILL AND ITS LICENSORS MAKE NO GUARANTEES OR WARRANTIES AS TO THE ACCURACY, ADEQUACY OR COMPLETENESS OF OR RESULTS TO BE OBTAINED FROM USING THE WORK, INCLUDING ANY INFORMATION THAT CAN BE ACCESSED THROUGH THE WORK VIA HYPERLINK OR OTHERWISE, AND EXPRESSLY DISCLAIM ANY WARRANTY, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. McGraw-Hill and its licensors do not warrant or guarantee that the functions contained in the work will meet your requirements or that its operation will be uninterrupted or error free. Neither McGraw-Hill nor its licensors shall be liable to you or anyone else for any inaccuracy, error or omission, regardless of cause, in the work or for any damages resulting therefrom. McGraw-Hill has no responsibility for the content of any information accessed through the work. Under no circumstances shall McGraw-Hill and/or its licensors be liable for any indirect, incidental, special, punitive, consequential or similar damages that result from the use of or inability to use the work, even if any of them has been advised of the possibility of such damages. This limitation of liability shall apply to any claim or cause whatsoever whether such claim or cause arises in contract, tort or otherwise.

DOI: 10.1036/0071590307



Professional



Want to learn more?

We hope you enjoy this
McGraw-Hill eBook! If

you'd like more information about this book,
its author, or related books and websites,
please [click here](#).

CONTENTS

Contributors	ix
Foreword	xiii
Preface.....	xvii
Acknowledgments.....	xix
1. A Primer on the Precision and Accuracy of the Clinical Examination	1
UPDATE: Primer on Precision and Accuracy.....	9
2. Does This Patient Have Abdominal Aortic Aneurysm?	17
UPDATE: Abdominal Aortic Aneurysm	25
3. Is Listening for Abdominal Bruits Useful in the Evaluation of Hypertension?	29
UPDATE: Abdominal Bruits	35
4. Does This Patient Have an Alcohol Problem?	39
UPDATE: Problem Alcohol Drinking	47
5. Does This Adult Patient Have Appendicitis?	53
UPDATE: Appendicitis, Adult	61
6. Does This Patient Have Ascites?	65
UPDATE: Ascites	71
7. What Can the Medical History and Physical Examination Tell Us About Low Back Pain?	75
UPDATE: Low Back Pain.....	83
8. Does This Patient Have Breast Cancer?	87
UPDATE: Breast Cancer	99
9. Does This Patient Have a Clinically Important Carotid Bruit?	103
UPDATE: Carotid Bruit	107
10. Does This Patient Have Carpal Tunnel Syndrome?	111
UPDATE: Carpal Tunnel.....	121
11. Does This Patient Have Abnormal Central Venous Pressure?	125
UPDATE: Central Venous Pressure.....	133
12. Does This Patient Have Acute Cholecystitis?	137
UPDATE: Cholecystitis	145
13. Does the Clinical Examination Predict Airflow Limitation?	149
UPDATE: Chronic Obstructive Airways Disease.....	159
14. Does This Patient Have Clubbing?.....	163
UPDATE: Clubbing	171
15. Is This Patient Taking the Treatment as Prescribed?	173
UPDATE: Compliance and Medication Adherence	179
16(A). Can the Clinical Examination Diagnose Left-Sided Heart Failure in Adults?	183

16(B). Does This Dyspneic Patient in the Emergency Department Have Congestive Heart Failure?	195
UPDATE: Congestive Heart Failure	209
17. Is This Patient Dead, Vegetative, or Severely Neurologically Impaired? Assessing Outcome for Comatose Survivors of Cardiac Arrest	215
UPDATE: Death	225
18(A). Does This Patient Have Deep Vein Thrombosis? (1998)	227
18(B). Does This Patient Have Deep Vein Thrombosis? (2006)	235
UPDATE: Deep Vein Thrombosis	245
19. Is This Patient Clinically Depressed?	247
UPDATE: Depression	259
20. Does This Patient Have a Family History of Cancer?	265
UPDATE: Family History of Cancer	275
21. Does This Patient Have a Goiter?	277
UPDATE: Goiter	285
22. Does This Patient Have Hepatomegaly?	289
UPDATE: Hepatomegaly	299
23. Does This Patient Have Hypertension?	301
UPDATE: Hypertension	311
24. Is This Adult Patient Hypovolemic?	315
UPDATE: Hypovolemia, Adult	325
25. Is This Child Dehydrated?	329
UPDATE: Hypovolemia, Child	339
26. Does This Patient Have Influenza?	343
UPDATE: Influenza	355
27. Does This Patient Have a Torn Meniscus or Ligament of the Knee?	357
UPDATE: Knee Ligaments and Menisci	369
28. Is This Adult Patient Malnourished?	371
UPDATE: Malnourishment, Adult	379
29. Does This Patient Have a Mole or a Melanoma?	383
UPDATE: Melanoma	391
30. Does This Adult Patient Have Acute Meningitis?	395
UPDATE: Meningitis, Adult	403
31. Is This Woman Perimenopausal?	407
UPDATE: Menopause	415
32. Does This Patient Have Aortic Regurgitation?	419
UPDATE: Murmur, Diastolic	429
33. Does This Patient Have an Abnormal Systolic Murmur?	433
UPDATE: Murmur, Systolic	443
34. Does This Patient Have Myasthenia Gravis?	449
UPDATE: Myasthenia Gravis	459

35. Is This Patient Having a Myocardial Infarction?	461
UPDATE: Myocardial Infarction	471
36. Does This Woman Have Osteoporosis?	477
UPDATE: Osteoporosis	489
37. Does This Child Have Acute Otitis Media?	493
UPDATE: Otitis Media, Child	501
38. Does This Patient Have Parkinson Disease?	505
UPDATE: Parkinsonism	513
39. Is This Patient Allergic to Penicillin?	515
UPDATE: Penicillin Allergy	523
40. Does This Adult Patient Have Community-Acquired Pneumonia?	527
UPDATE: Community-Acquired Pneumonia, Adult	535
41. Does This Infant Have Pneumonia?	539
UPDATE: Pneumonia, Infant and Child	547
42. Is This Patient Pregnant?	551
UPDATE: Early Pregnancy	559
43. Does This Patient Have Pulmonary Embolism?	561
UPDATE: Pulmonary Embolus	571
44. Does This Patient Have an Instability of the Shoulder or a Labrum Lesion?	577
UPDATE: Shoulder Instability	589
45. Does This Patient Have Sinusitis?	593
UPDATE: Sinusitis	601
46. Does This Patient Have Splenomegaly?	605
UPDATE: Splenomegaly	611
47. Does This Patient Have Strep Throat?	615
UPDATE: Streptococcal Pharyngitis	623
48. Is This Patient Having a Stroke?	627
UPDATE: Stroke	641
49. Does This Patient Have Temporal Arteritis?	643
UPDATE: Temporal Arteritis	653
50. Does This Patient Have an Acute Thoracic Aortic Dissection?	659
UPDATE: Thoracic Aortic Dissection	671
51. Does This Woman Have an Acute Uncomplicated Urinary Tract Infection?	675
UPDATE: Urinary Tract Infection, Women	687
52. What Is Causing This Patient's Vaginal Symptoms?	691
UPDATE: Vaginitis	705
53. Does This Dizzy Patient Have a Serious Form of Vertigo?	709
UPDATE: Vertigo	715
Index	719

JAMAEvidence: Using Evidence to Improve Care

Founded around the *Users' Guides to the Medical Literature* and *The Rational Clinical Examination: Evidence-Based Clinical Diagnosis*, JAMAEvidence offers an invaluable online resource for learning, teaching, and practicing evidence-based medicine (EBM). Updated regularly, the site includes fully searchable content of the *Users' Guides to the Medical Literature* and *The Rational Clinical Examination* and features podcasts from the leading minds in EBM, interactive worksheets, question wizards, functional calculators, and a comprehensive collection of PowerPoint slides for educators and students.

www.JAMAEvidence.com

Please visit www.JAMAEvidence.com for subscription information.



CONTRIBUTORS

John Attia, MD, MSc, PhD

Academic Director of General Medicine
University of Newcastle
John Hunter Hospital
New Lambton, New South Wales, Australia
Meningitis, adult

Robert G. Badgett, MD

Clinical Professor of Medicine
University of Texas, San Antonio
San Antonio, Texas, USA
Congestive heart failure

Allan N. Barkun, MD, MSc

Professor of Medicine and Epidemiology,
Biostatistics and Occupational Health
McGill University
Montreal General Hospital
Montreal, Quebec, Canada
Splenomegaly

Mary B. Barton, MD

Scientific Director, US Preventive Services Task Force
Agency for Healthcare Research and Quality
Washington, DC, USA
Breast cancer

Lori A. Bastian, MD, MPH

Associate Professor of Medicine
Duke University
Durham Veterans Affairs Medical Center
Durham, North Carolina, USA
Early pregnancy

David Bates, MD

Professor of Medicine
Harvard University
Brigham and Women's Hospital
Boston, Massachusetts, USA
Knee ligaments and menisci

Stephen Bent, MD

Assistant Adjunct Professor
University of California, San Francisco
San Francisco Veterans Affairs Medical Center
San Francisco, California, USA
Urinary tract infection, women

Christopher M. Booth, MD

Medical Oncologist and Research Fellow
Queen's University
Kingston, Ontario, Canada
Death

Hayden Bosworth, PhD

Research Professor
Duke University
Durham Veterans Affairs Medical Center
Durham, North Carolina, USA
Compliance and medication adherence

David Cescon, MD

Resident in Medicine
University of Toronto
Toronto, Ontario, Canada
Murmur, diastolic; Murmur, systolic

Sanjeev D. Chunilal, MB, ChB

North Shore Hospital
Takapuna, Auckland, New Zealand
Pulmonary embolus

Cathleen Colón-Emeric, MD

Assistant Professor of Medicine
Duke University
Durham, North Carolina, USA
Osteoporosis

Richard A. Deyo, MD

Professor of Medicine
University of Washington, Seattle
Harborview Medical Center
Seattle, Washington, USA
Low back pain

Edward Etchells, MD

Associate Professor
University of Toronto
Sunnybrook and Women's College Health
Sciences Centre
Toronto, Ontario, Canada
Murmur, diastolic; Murmur, systolic

Jeff Ginsberg, MD

Professor of Medicine
Faculty of Health Sciences
McMaster University
Hamilton, Ontario, Canada
Pulmonary embolus

James Grichnik, MD, PhD

Associate Professor of Medicine
Duke University
Durham, North Carolina, USA
Melanoma

Steven A. Grover, MD, MPA

Professor of Medicine and Epidemiology and Biostatistics
McGill University
Montreal General Hospital
Montreal, Quebec, Canada
Splenomegaly

Rose Hatala, MD

Clinical Associate Professor
University of British Columbia
Vancouver, British Columbia, Canada
Meningitis, adult

Catherine P. Kamenetzky, MD, MPH

Associate in Medicine
Duke University
Durham Veterans Affairs Medical Center
Durham, North Carolina, USA
Shoulder instability

Jeffrey N. Katz, MD

Associate Professor of Medicine
Harvard University
Brigham and Women's Hospital
Boston, Massachusetts, USA
Knee ligaments and menisci

Sheri A. Keitz, MD, PhD

Chief, Medicine Service
Miami Veterans Affairs Medical Center
Associate Dean for Faculty Diversity and Development
University of Miami Miller School of Medicine
Miami, Florida, USA
The Rational Clinical Examination Education Guides

Michael Klompaas, MD

Research Fellow in Medicine
Harvard University
Massachusetts General Hospital and
Brigham and Women's Hospital
Boston, Massachusetts, USA
Thoracic aortic dissection

Frank A. Lederle, MD

Professor of Medicine
Minneapolis Veterans Affairs Medical Center
Minneapolis, Minnesota, USA
Abdominal aortic aneurysm

Catherine R. Lucey, MD

Clinical Professor of Medicine
Ohio State University
Columbus, Ohio, USA
Congestive heart failure

Kathryn Myers, MD

University of Western Ontario
St. Joseph's Health Care
London, Ontario, Canada
Clubbing

Daniel A. Ostrovsky, MD

Clinical Associate in Pediatrics
Duke University
Durham, North Carolina, USA
Pneumonia, infant and child

Joanne T. Piscitelli, MD

Associate Clinical Professor of Obstetrics and Gynecology
Duke University
Durham, North Carolina, USA
Vaginitis

James Rainville, MD

Assistant Clinical Professor of Physical Medicine
and Rehabilitation
Harvard Medical School
New England Baptist Hospital
Boston, Massachusetts, USA
Low back pain

Goutham Rao, MD

Associate Professor, Pediatrics
University of Pittsburgh School of Medicine
Pittsburgh, Pennsylvania, USA
Parkinsonism

Drummond Rennie, MD

JAMA
Chicago, Illinois
Philip R. Lee Institute for Health Policy Studies
University of California, San Francisco
San Francisco, California, USA
Editor, *The Rational Clinical Examination*

Russell Rothman, MD

Assistant Professor, Internal Medicine and Pediatrics
Vanderbilt University
Nashville, Tennessee, USA
Otitis media, child

Jonathan L. Schaffer, MD, MBA

Advanced Operative Technology Group
Orthopaedic Research Center
Cleveland Clinic
Cleveland, Ohio, USA
Knee ligaments and menisci

Katalin Scherer, MD

Assistant Professor of Neurology
University of Arizona
Tucson, Arizona, USA
Myasthenia gravis

Robert H. Shmerling, MD

Associate Professor of Medicine
 Harvard Medical School
 Beth Israel Deaconess Medical Center
 Boston, Massachusetts, USA
Temporal arteritis

Kaveh G. Shojania, MD

Assistant Professor of Medicine
 University of Ottawa
 Ottawa, Ontario, Canada
Cholecystitis

David L. Simel, MD, MHS

Chief, Medicine Service
 Durham Veterans Affairs Medical Center
 Professor of Medicine
 Duke University School of Medicine
 Durham, North Carolina, USA
 Editor, *The Rational Clinical Examination*
Abdominal bruits; Ascites; Carotid bruit; Carpal tunnel; Central venous pressure; Chronic obstructive airways disease; Community-acquired pneumonia, adult; Deep vein thrombosis; Family history of cancer; Goiter; Hepatomegaly; Hypertension; Hypovolemia, adult; Influenza; Malnourishment, adult; Melanoma; Myocardial infarction; Otitis media, child; Penicillin allergy; Primer on precision and accuracy; Problem alcohol drinking; Sinusitis; Streptococcal pharyngitis; Stroke; Vaginitis; Vertigo

Gerald W. Smetana, MD

Associate Professor of Medicine
 Harvard Medical School
 Beth Israel Deaconess Medical Center
 Boston, Massachusetts, USA
Temporal arteritis

Daniel H. Solomon, MD, MPH

Associate Professor of Medicine
 Harvard Medical School
 Brigham and Women's Hospital
 Boston, Massachusetts, USA
Knee ligaments and menisci

Michael J. Steiner, MD

Assistant Professor of Pediatrics and Medicine
 University of North Carolina, Chapel Hill
 Chapel Hill, North Carolina, USA
Hypovolemia, child

Ben Stern, MS, DPT

CEO and Director of Development
 Integrated Golf Schools
 Scottsdale, Arizona, USA
Low back pain

Robert Trowbridge, MD

Assistant Professor of Medicine
 University of Vermont College of Medicine
 Maine Medical Center
 Portland, Maine, USA
Cholecystitis

James Wagner, MD

Associate Professor
 University of Texas, Southwestern Medical Center
 Dallas, Texas, USA
Appendicitis, adult

John W. Williams, Jr, MD, MHS

Professor of Medicine
 Duke University
 Durham Veterans Affairs Medical Center
 Durham, North Carolina, USA
Depression; Sinusitis

Jeffrey G. Wong, MD

Professor of Medicine
 Associate Dean for Medical Education
 Medical University of South Carolina
 Charleston, South Carolina, USA
Meningitis, adult

This page intentionally left blank

FOREWORD

I remember my introduction to the medical history and clinical examination as the most exciting moments of my early career. As each item in the history and physical examination was explained and given meaning and significance, I believed that after the long preclinical years I had at last reached the threshold of becoming a physician. I could begin to hold more than a comforting conversation with a patient. I could use my ears, eyes, and hands to disclose the patient's problem and so begin to be of actual use to a real patient. As I polished my skills, it did not occur to me that the divination of all those signs and symptoms was anything but an art: the epitome of the art of medicine.

But, with time, I realized that many of the so-called pathognomonic symptoms and signs were so merely because someone, often the person whose name was attached to them, had declared that they were. Doubt started to overtake accepted wisdom as it became clear to me that little worthwhile evidence supported the artist's tools I thought I had mastered.

Towards the end of the 1980s, my friend David Sackett, then chief of medicine and clinical epidemiology and biostatistics at McMaster University, showed me a new way of thinking about all this. He equated items in the history and the physical examination with traditional diagnostic laboratory tests, each susceptible to evidentiary testing. So he and I began planning 2 series of articles on evidence-based medicine to appear in *JAMA*. One of these, the *Users' Guides to the Medical Literature*, was soon placed into the capable hands of Gordon Guyatt, also of McMaster University, and articles began to appear in *JAMA* in 1993. By 2002, they were printed in updated form in 2 books, an *Essentials* and a fuller *Manual*,^{1,2} both of which have been so successful that second editions^{3,4} have just been published.

The other series consisted of The Rational Clinical Examination articles and started appearing in 1992. With the first article, Sackett and I published an editorial.⁵ We reminded our readers of studies that showed that primary care providers usually establish the correct diagnosis at the end of a brief history and some subroutine of the physical examination. So on practical grounds alone, it made sense to improve our understanding of the parts of the history and examination that were useful, or useless, in pinning down, usually at an early stage of the disease, one diagnosis and ruling out others. We contrasted symptoms and signs with laboratory tests, which were subjected to rigorous testing before adoption, but which might have far less ability to narrow the diagnostic possibilities. As an example, we observed the overwhelming probability of coronary stenosis in a 65-year-old man who has smoked all his life when he tells you that he gets central chest tightness regularly on exertion, which forces him to stop and which disappears when he rests.^{6,7}

Perhaps most important, by encouraging research into the history and physical examination, we wanted to restore

respectability to a part of medicine that seemed to have been eroding as academic and financial rewards went to those who most resembled scientists relying on expensive diagnostic tests and least behaved as physicians relating to patients.

It is no coincidence that both Sackett and I, authors of the editorial launching the series, have served roles in the Cochrane Collaboration, an initiative that has had a massive effect on the way we see evidence and a profound influence on the methods and popularity of systematic review and meta-analysis. These sciences, as well as that of decision making, had grown up and spread to medicine during the 1970s and 1980s. Without them, both the Cochrane Collaboration and The Rational Clinical Examination series would have been impossible undertakings; indeed, the entire evidence-based movement would have grown far more slowly.

At the same time, because of the unfamiliarity of these techniques and the revolutionary approach we were taking, namely, a scientific examination of what most clinicians considered to be an ineffable art not susceptible to dissection, we published a primer on the precision and accuracy of the clinical examination. This laid out the approach to be taken and took the reader through the terms, methods, and calculations underpinning clinical diagnosis.⁸

Although each article's purpose could be worked out from its title, the full meaning of the concepts took time to sink in, as I discovered from comments sent in by many of the expert specialty peer reviewers to whom I sent the manuscripts as they came in to *JAMA*. Indeed, it was unfamiliar even to some prospective authors. David Sackett had a firm belief that the reviews would be done best by generalist physicians who had learned basic critical appraisal skills. As the editor, I learned that these generalist physicians were often speaking a different language from our specialist reviewers. Sackett was clearly correct, and it remains commonplace for specialty reviewers to ask that specialists be added to the writing team because, well, they are specialists. What has happened in our process is that both authors and reviewers learn from the editorial review process, with specialty reviewers ensuring that authors interpret the data in the proper context. In return, the specialists often learn that much of what they took for granted has no basis in evidence.

The Rational Clinical Examination book should not replace books on clinical diagnosis. But, somewhat as the Cochrane Database of Systematic Reviews provides a systematic evaluation of all studies on a particular intervention without becoming prescriptive, so articles in The Rational Clinical Examination series are careful systematic efforts to assess the accuracy of items from the patient's medical history and the clinical examination. In this sense, they are a revolutionary departure from what we have regarded as books on physical diagnosis, which, until the first articles in The Rational Clinical Examination series appeared,

had never taken that approach. Since then, however, such books have already started using the evidence as summarized in articles in the series.

In his preface to the eighth edition of *DeGowin's Diagnostic Examination*, Richard LeBlond writes:

References to articles from the medical literature are included in the body of the text. We have chosen articles which provide useful clinical information including excellent descriptions of disease and syndromes and, in some cases, photographs illustrating key findings. Evidence-based articles on the utility of the physical exam are included, mostly from The Rational Clinical Examination series published over the last decade in the *Journal of the American Medical Association*. They are included with the caveat that they evaluate the physical exam as a hypothesis-testing tool, not as a hypothesis generating task. ...⁹

Our series is indeed about testing tests (symptoms, signs) to separate the useful from the useless and so is about testing hypotheses. Books on physical diagnosis are hypothesis generating in that they are a compendium of instructions on how to elicit all symptoms and signs, typically presented in the absence of any certain disease consideration or context, typically organized by organ system (eg, "the cardiovascular examination"). In contrast, our articles are usually organized by a certain condition (eg, "Does this patient have systolic dysfunction?"). And, although there are a few articles in which the authors take a more hypothesis-generating tack (eg, those on splenomegaly and hepatomegaly), we always frame them in a clinical context.

An issue all along has been whether, and how much, to integrate the evidence on symptoms and signs with that provided by diagnostic tests. In general, we have had so much material to deal with, and there are so many good texts on diagnostic tests, that we have limited our approach as much as common sense would allow. Some articles do include assessments of a few basic laboratory and radiologic studies that are commonly available to the clinician and that can be interpreted only by the physician in the clinical context (eg, the sedimentation rate for temporal arthritis or vascular congestion on a chest radiograph for systolic dysfunction). Recently, we expanded the series to include "rational clinical procedures," because many procedures are actually part of the clinical examination and tightly linked to the presence of the history and physical examination findings.¹⁰

David Simel of Duke University had been immediately excited by the concept and was a coauthor of the first article in the series, "Does This Patient Have Ascites? How to Divine Fluid in the Abdomen."¹¹ At that time, 1992, Simel made it clear that he intended to devote his research career to investigating this crucial area of medicine, and soon after he took over as primary editor of the series. Since then, he has stimulated large numbers of authors to complete these systematic reviews. His personal involvement with authors has brought us many more articles than we could otherwise have expected and ensured a uniform presentation. He also made certain that every manuscript had been through review before submission

to *JAMA*, where I put each manuscript through rigorous external peer review, just as with all original submissions to *JAMA*.

Each review is a considerable undertaking, often requiring more than a year of unpaid and often unappreciated work, which explains why it has taken 15 years to produce what is now more than 70 articles in *JAMA*. As news of the series spread, volunteer authors suggested their own topics of interest. The appearance of fully fledged review articles depended on the skills and persistence of the authors and on the persuasive powers and analytic assistance of David Simel. Even then, more than a fifth of the proposed topics failed to result in publishable manuscripts, usually because the authors found insufficient evidence. It is for that reason that Simel and I published in 1995 a plea for support for a wide research agenda and the formation of collaborations to ensure that the wide gaps in our knowledge were filled.¹²

With the publication of this book, Simel has updated the first 51 published articles either alone or with the original authors. In addition, he has updated the primer⁸—essential for all readers of this book. David Simel's contributions to this series, and the transformation he has wrought in how we think about the clinical examination, have been immense, and working with him has been a privilege and a delight.

This is the first book in The Rational Clinical Examination series. Our plan is to keep soliciting and publishing in *JAMA* articles on fresh Rational Clinical Examination topics. We welcome volunteers with good ideas who are prepared to undertake the work. We will accumulate these articles, keeping them current with updates, and publish them as new chapters online and in succeeding editions of *The Rational Clinical Examination* book. *The Rational Clinical Examination* will be published online with a set of teaching/learning slides for each chapter and will be integrated with the *Users' Guides to the Medical Literature* and other online-only content and features in an extensive evidence-based medicine Web site called, JAMAevidence (<http://www.JAMAevidence.com>).

David Simel and I welcome Sheri Keitz (recently of the Durham Veterans Affairs Medical Center and Duke University, who has now moved to the University of Miami) as editor of *The Rational Clinical Examination Education Guides*. Sheri has many talents, including a fine critical eye. She has prepared or supervised development of all the teaching slides, and she has reviewed most of the Updates to the original manuscripts.

The series started with the encouragement of George Lundberg, then editor-in-chief of *JAMA* and the *Archives* journals. His successor, Cathy DeAngelis, has consistently and very strongly supported us, helping negotiate the complex path to publication. Annette Flanagan has been a tireless worker in this, as in so many other *JAMA* causes. This book would not have been possible without her.

We are grateful to Barry Bowlus for directing the publishing of this book and to Richard Newman for his advice and support. We are also grateful for the expertise of Jim Shanahan, Robert Pancotti, Helen Parr, and others at McGraw-Hill, as well as Peter Compitello at NewGen, and Holly Auten and her colleagues at Silverchair.

Publishing, like medicine, moves forward. During the last few years, the illustrations in *JAMA* have come under the care of Ronna Siegel and 2 medical illustrators, Cassio Lynn and Alison Burke. The series articles have benefited from their extraordinary skills, and improvements continue with the introduction of video images, as well as teaching clips. We also thank Cara Wallace and Angela Grayson for their expert editing and support.

The response to the articles published in *JAMA* tells us that this book will be useful. We also hope that readers will be stimulated to conduct research on aspects of the clinical examination. Perhaps readers will contact us if they believe they can undertake the sort of review that could constitute future articles in *JAMA* and chapters in the next book.

—Drummond Rennie, MD, FRCP, MACP

REFERENCES

1. Guyatt G, Rennie D, eds. *Users' Guides to the Medical Literature: A Manual for Evidence-Based Practice*. Chicago, IL: AMA Press; 2001.
2. Guyatt G, Rennie D, eds. *Users' Guides to the Medical Literature: Essentials of Evidence-Based Practice*. Chicago, IL: AMA Press; 2001.
3. Guyatt G, Rennie D, Meade MO, Cook DJ, eds. *Users' Guides to the Medical Literature: A Manual for Evidence-Based Clinical Practice*. 2nd ed. New York, NY: McGraw-Hill; 2008.
4. Guyatt G, Rennie D, Meade MO, Cook DJ, eds. *Users' Guides to the Medical Literature: Essentials of Evidence-Based Clinical Practice*. 2nd ed. New York, NY: McGraw-Hill; 2008.
5. Sackett DL, Rennie D. The science of the art of the clinical examination. *JAMA*. 1992;267(19):2650-2652.
6. Diamond GA, Forrester JS. Analysis of probability as an aid in the clinical diagnosis of coronary artery disease. *N Engl J Med*. 1979;300(24):1350-1358.
7. Sox HC, Hickam DH, Marton KJ, et al. Using the patient's history to estimate the probability of coronary artery disease: a comparison of primary and referral practices. *Am J Med*. 1990;89(1):7-14.
8. Sackett DL. The rational clinical examination: a primer on the precision and accuracy of the clinical examination. *JAMA*. 1992;267(19):2638-2644.
9. LeBlond RF. Preface. In: LeBlond RF, DeGowin RL, Brown DD, eds. *DeGowin's Diagnostic Examination*. 8th ed. New York, NY: McGraw-Hill; 2004.
10. Straus SE, Thorpe KE, Holroyd-Leduc J. How do I perform a lumbar puncture and analyze the results to diagnose bacterial meningitis? *JAMA*. 2006;296(16):2012-2022.
11. Williams JW Jr, Simel DL. The rational clinical examination: does this patient have ascites? how to divine fluid in the abdomen. *JAMA*. 1992;267(19):2645-2648.
12. Simel DL, Rennie D. The clinical examination: an agenda to make it more rational. *JAMA*. 1997;277(7):572-574.

This page intentionally left blank

PREFACE

I've never met a medical student who lacked passion for making a diagnosis. And, among all the diagnoses a student might make, clinching the case right at the bedside is the most treasured. The same holds true not only for physicians in practice but also for all those involved in caring for patients—physician assistants, nurses, and physical therapists must each constantly assess their patient and consider what's wrong. The Rational Clinical Examination series, published in *JAMA* since 1992 and collected in this book, should appeal to anyone who wonders about the meaning of a patient's symptoms and signs. Many indispensable textbooks instruct learners on "how" to elicit the medical history and perform the physical examination, but we suspect that, once the "how" is learned, clinicians only infrequently return to what was one of their favored textbooks during their training years. When I ask clinicians to recall the book they used for physical diagnosis class in medical school, there is no pause before they state DeGowin and DeGowin, Bates, Mosby, Schwartz, or another of a select few. We see *The Rational Clinical Examination* as an essential companion to, and not a replacement for, these time-honored texts of the "complete" medical history and physical examination.

Although standard textbooks might clearly describe several maneuvers for detecting ascites, for example, we identify those findings that work best. Although textbooks typically march from "head to toe" without regard to diagnoses when describing the complete physical examination, we start with clinical diagnostic questions and provide data that identify the most relevant symptoms and signs. Unlike physical examination textbooks, we also provide data on what does not work, derived from a thorough review of the literature that backs up our recommendations.

Please recognize that we can never replace a great textbook on the complete medical history and physical examination because we will never be complete in describing the rational clinical examination. There are many diagnoses we have not yet reviewed and many more to come. After more than 15 years of producing systematic reviews in *JAMA*, which included the article that launched the evidence-based medicine movement,¹ it was time for us to update and combine our work in one resource for learners and clinicians to enjoy.

Accordingly, this book is evidence based. We present the original Rational Clinical Examination article, followed by an Update. For each topic, we recreated the original literature search and evaluated the new literature dating from 1 year before the publication of the original article to the time we prepared the Update. If anything, we tried to be even more restrictive in applying our quality measures for including new research in the Updates. The Updates follow a format similar to that of the original articles: they open with a clinical scenario, present the results of the literature search, and summarize new information. Sometimes we discovered that we had

not reviewed the topic as thoroughly as we thought, so we also recount any improvements we made when we reanalyzed data. Simple tables display the new findings that we incorporate with the previously published data.

Because evidence-based guidelines for most diseases did not exist when we launched The Rational Clinical Examination series, we review the recommendations of the major federal agencies for each of the topics and highlight how our information supports or differs from those recommendations. Finally, we include a Make the Diagnosis section that gives a summary of the prior probability of the target disorder, the population for whom the target disorder should be considered, a table of likelihood ratio data for the best clinical findings, and a list of the accepted reference standards.

Some readers will want more data, so we provide a structured review of every article identified in our Update that met our inclusion criteria. These reviews are available online in an Evidence to Support the Update section, available at <http://www.JAMAEvidence.com>. JAMAEvidence is a Web site resource for learning, teaching, and practicing evidence-based medicine that includes the complete online content of *The Rational Clinical Examination* and the *Users' Guides to the Medical Literature*, along with other features, such as downloadable projection slides to enhance classroom or conference teaching and learning experience, an extensive evidence-based medicine glossary, functional calculators, question wizards, customizable worksheets, podcasts, and regular updates.

We hope that long-time readers of The Rational Clinical Examination series will recognize the painstaking care and preparation taken during the review of each topic. Every Update was reviewed by an author of the original article or a clinician who had no involvement with the original publication. Although this alone might seem reassuring and unlike typical medical textbooks, we went a step further.

For each topic, a slide presentation, called an Education Guide, has been prepared, primarily by Duke University Department of Medicine residents, or in a few cases by young clinical Duke University faculty members, all supervised by Sheri A. Keitz, MD, PhD. The Education Guides follow a similar format and have been "field-tested" among learners. The goal in preparing the Education Guides was to have the learners create a set of materials for their instructors that match how they, the learners, hope the topic would be taught. Just like the Updates themselves, the slides have also been reviewed. From this, we learned that trainees are among our most critical readers—they expect careful, accurate, and thoughtful presentation and exposition. The Education Guides slides are available online at <http://www.JAMAEvidence.com>.

For current students, *The Rational Clinical Examination* demonstrates the correct way to learn the medical history and physical examination, giving direction in interpreting

the results and answering questions that typical physical examination textbooks do not systematically address. For teachers, the Education Guides, amply supplemented with teacher's notes, allow you to teach physical diagnosis with an evidence-based approach. For established practitioners, perhaps far removed from their introductory physical examination course, we hope to challenge any cynicism that clinical examination is all "art." There *is* a science behind the art of clinical examination. We hope you discover that learning this

science not only validates your role as a clinician and improves your skills but also is fun.

—David L. Simel, MD, MHS

REFERENCE

1. Evidence-Based Medicine Working Group. Evidence-based medicine: a new approach to teaching the practice of medicine. *JAMA*. 1992;268(17): 2420-2425.

ACKNOWLEDGMENTS

To all those who supported me at the Durham Veterans Affairs Medical Center and Duke University, and to the many learners, trainees, and colleagues around the world who have encouraged this work and contributed to it, I give my thanks. Jack Feussner and Drummond Rennie have been my most important career mentors by allowing me to dabble in this seldom funded, but purely pleasurable, line of investigative work on the clinical examination.

Sheri Keitz and her mentorship of the Duke University Medical house staff in preparing the Education Guides that accompany the book have kept the writing very honest, elevating everything to a higher standard than I could have accomplished alone. Barry Bowlus, Annette Flanagin, and Cathy DeAngelis took charge of all the arrangements with our publisher, freeing me to concentrate on content because there are no shortcuts when the imprimatur of *JAMA* is attached to *The Rational Clinical Examination*. Cara Wallace and Angela Grayson helped with the editing, and Cassio Lynn, Alison Burke, and Ronna Siegel have provided us with illustrations that convey more information than words sometimes allow. Pete Compitello from NewGen, Helen Parr, Jim Shanahan, Robert Pancotti, and others from McGraw-Hill, and Holly Auten and her colleagues at Silverchair have taken what we put on paper and transformed the information to print and online formats in a way that met and then surpassed our vision.

Kenneth Goldberg and Rob Minton covered all my clinical and hospital tasks when I took time off, and David Matchar and Eugene Oddone made possible a 6-month escape from work. I especially thank Joanne, Lauren, Michael, and Brian for their love and for their continued forgiveness for the many times I am physically present while working on *The Rational Clinical Examination* though my mind is elsewhere. I owe Joanne a lot of Sunday morning bike rides.

—David L. Simel, MD, MHS

Having never worked on a book before, I had no way of knowing how it would affect my life or the lives of those around me. I was deeply embedded in editing for *The Rational Clinical Examination* when my 11-year-old son entered our study at home and said, “Mom, I think you need to take a break. Look,” he declared, pointing to my hands; I was shaking out the cramps after hours at the keyboard. “Isn’t that flick sign?” My poor children had spent so much time during the past year watching me display signs as I worked on each chapter that one of them could correctly identify flick sign. I responded, “See! Everyone flicks their wrists, whether or not they have carpal tunnel syndrome!”

I am particularly indebted to my family, who supported my work on many nights, weekends, and holidays in order to complete my portions of the book. I am also thankful for the great number of people who have been critical to the success of the project at the Durham Veterans Affairs Medical Center and Duke University. Beth Weast, Sarah Williams, and Phyllis Scott were a tremendous help in holding things together at work and helping me find time to move the book forward. My personal and professional mentors, Gene Oddone and Dave Simel, helped me develop my passion for including best evidence in my clinical practice and teaching. I want to give special thanks to Dave. As he has done for so many, he created this opportunity for me and provided steady support, guidance, and tolerance as I made my way through each chapter. Finally, this work was only possible through collaboration with many Duke house staff and faculty members whose constant drive for understanding improved the quality of our work and elevated our standards to the highest level.

—Sheri A. Keitz, MD, PhD

This page intentionally left blank

CHAPTER

1

A Primer on the Precision and Accuracy of the Clinical Examination

David L. Sackett, MD, MSc Epid, FRCPC

This background article will introduce and explain the terms and concepts that are being used in the series of overviews on the rational clinical examination that begins in this issue of THE JOURNAL. It includes definitions and explanations of certain key concepts, clinical examples, guides for reading clinical journals about a diagnostic test, and a blank “working table” that you can use to apply the concepts on your own.

Background articles in this series will discuss selected issues in the precision and accuracy of the clinical examination in greater detail or extend them to more complex diagnostic situations. Some of these issues are also discussed in clinical epidemiology textbooks.¹

Of course, the precision and accuracy of the clinical examination are not the only concerns in the clinical encounter, and their proper application provides only the starting point for decisions about how certain we need to be about a diagnosis before we act on it (the decision threshold) and how we ought to incorporate the concerns of both patients and society in deciding whether and how to act. Later background articles will discuss these additional considerations; this one will be confined to precision and accuracy.

Like others in the series, this background article will be introduced with a patient.

THE PATIENT

One of your patients, whom you have not seen for several years, is admitted to the orthopedic service after a packing crate has tipped over onto his leg, producing an unstable fracture of his distal tibia and fibula. You stop by to see him as he is being prepared for surgery. He is alert and hemodynamically stable but smells of alcohol (at 10 AM) and has 3 spider nevi on his upper chest (but no gynecomastia or asterixis). He is obese, and his belly is prominent. Among the questions that are raised in your mind, the following are of special significance:

1. Is this man an alcoholic? You would place the odds for this disorder at 50-50 (and the science of the art of how clinicians generate these odds will be the subject of a later background article). The answer to this diagnostic question is important in the long run and in protecting him from the complications of acute withdrawal during and after his operation.
2. Does he have ascites? You are much less sure here, but if he is alcohol dependent you would place the odds that the prominence of his belly represents ascites also at 50-50. Again, it would be important to know whether he has this manifestation of advanced alcoholic liver damage.

Your options for answering these questions are several. To explore his possible alcohol abuse or dependency, (1) you could take the time required for a thorough confrontation and

		Alcohol Abuse or Dependency		
		Yes	No	
No. of Positive Answers to the 4 CAGE Questions	3 Or 4	60 (True +)	1 (False +)	61
	2, 1, Or none	57 (False -)	400 (True -)	457
		a+c 117	b+d 401	a+b+c+d 518
		a c	b d	a+b c+d a+b+c+d

Figure 1-1 The CAGE Questions for Alcohol Abuse or Dependency

Characteristics: sensitivity, $a/(a + c) = 60/117 = 0.51$, or 51%; specificity, $d/(b + d) = 400/401 = 0.998$, or 99.8%. Predictions: positive predictive value or posttest probability of having the target disorder (alcohol abuse or dependency) for patients with 3 or 4 positive responses, $a/(a + b) = 60/61 = 0.98$, or 98%; negative predictive value or posttest probability of not having the target disorder for patients with 2 or fewer positive responses, $d/(c + d) = 400/457 = 0.88$ or 88%; posttest probability of having the target disorder for patients with 2 or fewer positive responses, $c/(c + d) = 57/457 = 0.12$, or 12%. Prevalence or pretest probability of having the target disorder (adapted from Bush et al⁵), $(a + c)/(a + b + c + d) = 117/518 = 23\%$. Abbreviation: CAGE, cut down, annoyed, guilty, eye opener.

interrogation about the amount of alcohol he consumes (and, in the process, risk alienating him, estranging the nursing staff, and exasperating yourself); (2) you could order 1 or more liver function tests; (3) you could even request one of the new, “hot” tests for platelet enzyme activity, reported to be elevated in persons with alcoholism²; or (4) you could ask him the 4 quick “CAGE” questions: Have you ever felt you should cut down on your drinking? Have people annoyed you by criticizing your drinking? Have you ever felt bad or guilty about your drinking? Have you ever had a drink first thing in the morning to steady your nerves or to get rid of a hangover (eye-opener)? This opening example in the series is all the more appropriate when we observe that the first report on the CAGE questionnaire in a general medical journal was by John Ewing³ and that it was accompanied by an editorial from a major supporter of this series, George Lundberg.⁴ To explore his possible ascites, (1) you could check him for shifting dullness, fluid wave, or even the puddle sign; (2) you could order an abdominal ultrasonographic examination; or (3) you could simply ask him whether he has ever had swollen ankles.

Stop for a moment and consider the implications, in terms of your time and somebody’s money, of the alternative ways of answering these 2 questions. Would it not be better if you could answer them both with just 5 quick questions (4 for CAGE and 1 about ankle swelling)?

As it happens, you might be able to do just that. If he answers yes to 3 or 4 of the CAGE questions, he is an alcohol-abusing or alcohol-dependent man (and this medical history is far more powerful than any laboratory tests you can

order). If he answers no to ankle swelling, you have pretty well ruled out clinically important ascites (you could double check the latter by testing for shifting dullness; like most such patients, he did not have a fluid wave, and as you will learn in a forthcoming overview on ascites, the puddle sign is not useful in him or anybody else). Thus, for both questions, a quick bedside examination has provided definitive diagnostic information, without the need for laboratory testing or diagnostic imaging.

How can we make such a bold statement about the power of these simple elements of the clinical history and physical examination? The answer lies in the science of the art of clinical diagnosis that underpins this series of overviews on the rational clinical examination. This first background article will introduce and illustrate the key elements of this science (and readers who want a more detailed discussion of what follows can consult a step-by-step discussion published elsewhere¹). The background articles also are intended to convey the fun and gratification physicians derive from making correct diagnoses with crispness and dispatch.

TAKING AN ALTERNATIVE HISTORY FOR ALCOHOLISM

Examine Figure 1-1. In it are shown the number of positive answers to the CAGE questions from 2 groups of patients admitted to the orthopedic or medical services of a community-based teaching hospital in Boston, Massachusetts.⁵ In the left-hand column are the responses from patients whose extensive evaluations (including, where indicated, detailed social histories, follow-ups, and liver biopsies) provided acceptable “proof” that they were alcohol abusers or alcohol dependent. In the right-hand column are patients whose evaluations showed that they were not alcohol abusers or dependent. These extensive confirmatory investigations often are referred to as criterion standards of diagnosis and typically consist of definitive findings at angiography, operation, autopsy, and the like.

This study is useful to clinicians because the CAGE history and the extensive (reference or criterion standard) investigations were carried out independently among a wide spectrum of well-described patients in whom it was clinically reasonable to inquire about alcohol abuse. It thus satisfies the first criterion of a valid, clinically useful article on diagnostic strategies that appears in Table 1-1 (has there been an independent, “blind” comparison with a criterion standard of diagnosis?). The readers’ guides in Table 1-1 have been used by the authors of this series on the rational clinical examination to “screen” articles for inclusion in their overviews of diagnostic approaches to specific clinical problems. Table 1-1 can be clipped and carried for easy reference when reading clinical articles that make claims about the usefulness of (especially new) diagnostic tests, and the reasoning behind its elements are described in detail elsewhere.¹

The study that generated Figure 1-1 also satisfied the second, commonsense guide, for it was carried out in a patient sample that included an appropriate spectrum of mild and severe, treated and untreated alcoholism, plus individuals

with different but commonly confused disorders. The setting for the study (a large, urban, general hospital) was described, satisfying the third readers' guide and permitting us to determine the applicability of the results to our own setting, and the term *normal* (the fifth guide) was clearly and sensibly defined as the absence of alcohol abuse or dependency (we shall return to the fourth guide of reproducibility later).

The authors of the CAGE study were not proposing that their questions be used as part of an extensive series ("cluster") of diagnostic tests (so the sixth guide does not apply), and the questions were presented with their exact wording in the article, satisfying the seventh guide and permitting their exact application in the reader's own practice. The final readers' guide (has the utility of the test been determined?) is satisfied to the extent that the CAGE questions recognized far more persons with alcoholism, especially alcohol abusers, than routine clinical diagnosis and made them candidates for treatment and counseling.

In summary, the CAGE study observed the methodologic standards required for a valid and clinically useful description of the clinical applicability of any diagnostic information, whether it comes from the clinical history, the physical examination, or the diagnostic laboratory.

THE PRECISION OF THE CLINICAL EXAMINATION

For an item of the clinical history or physical examination to be accurate, it first must be precise. That is, we need to have some confidence that 2 clinicians examining the same, unchanged patient would agree with each other on the presence or absence of the symptom (such as our patient's answer to one of the CAGE questions) or sign (such as the presence of spider nevi on our patient's chest). The precision (often appearing under the name of "observer variation" in the clinical literature) of such clinical findings can be quantitated.⁶

Suppose 2 clinicians recorded whether they found spider nevi when they independently examined the same 100 patients suspected of having liver disease and generated the data shown in Figure 1-2. The 2 clinicians agreed that 23 of the patients (cell a) had spider nevi and that 66 patients (cell d) did not; thus, they agreed on $(23 + 66)/100 = 89\%$ of the patients they examined. However, 6 patients (cell c) judged to have spider nevi by the first clinician were judged not to have nevi by the second, and 5 patients (cell b) judged to have spider nevi by the second clinician were judged not to have nevi by the first. How should we interpret this precision? Is this degree of clinical agreement good, or should we expect better?

We might begin by recognizing that some clinical agreement would occur by chance alone. For example, if the second clinician merely tossed a coin for each patient instead of carrying out an examination, reporting nevi if the coin came up "heads" and no nevi if it came up "tails," agreement would be 50%. We should begin, then, by determining how much of the observed agreement of 89% was because of chance, so that we can find out how much real clinical skill (agreement beyond chance) was being displayed by these clinicians.

Table 1-1 Readers' Guides for an Article About a Diagnostic Test

- Has there been an independent, "blind" comparison with a criterion standard of diagnosis?
- Has the diagnostic test been evaluated in a patient sample that included an appropriate spectrum of mild and severe, treated and untreated disease, plus individuals with different but commonly confused disorders?
- Was the setting for this evaluation, as well as the filter through which study patients passed, adequately described?
- Have the reproducibility of the test result (precision) and its interpretation (observer variation) been determined?
- Has the term *normal* been defined sensibly as it applies to this test?
- If the test is advocated as part of a cluster or sequence of tests, had its individual contribution to the overall validity of the cluster and sequence been determined?
- Have the tactics for carrying out the test been described in sufficient detail to permit their exact replication?
- Has the utility of the test been determined?

		First Clinician's Examination for Spider Nevi		
		Positive	Negative	
Second Clinician's Examination for Spider Nevi	Positive	23 (Expect 8)	5	28
	Negative	6	66 (Expect 51)	72
		<i>a</i>	<i>b</i>	<i>a+b</i>
		<i>c</i>	<i>d</i>	<i>c+d</i>
		<i>a+c</i>	<i>b+d</i>	<i>a+b+c+d</i>
		29	71	100

Figure 1-2 The Precision of the Clinical Examination for Spider Nevi

Observed agreement:

$$(a + d)/(a + b + c + d) = (23 + 66)/100 = 89\%$$

Expected agreement:

$$\text{For cell } a, [(a + b) \times (a + c)]/(a + b + c + d) = (28 \times 29)/100 = 8$$

$$\text{For cell } d, [(c + d) \times (b + d)]/(a + b + c + d) = (72 \times 71)/100 = 51$$

$$\text{Calculate expected agreement as } (\text{expected } a + \text{expected } d)/(a + b + c + d) = (8 + 51)/100 = 59\%.$$

$$\text{Agreement beyond chance} = \kappa = (\text{observed agreement} - \text{expected agreement}) / (100\% - \text{expected agreement}) = (89\% - 59\%) / (100\% - 59\%) = 0.73.$$

Conventional levels of κ : slight, 0.0–0.2; fair, 0.2–0.4; moderate, 0.4–0.6; substantial, 0.6–0.8; almost perfect, 0.8–1.0.

Adapted from Lundberg.⁴

Chance agreement can be calculated by the formal process of "marginal cross-products" shown in Figure 1-2, but it also can be thought of as a coin toss in which, for example, the first clinician's coin came up heads 29% of the time (based on $[a + c]/[a + b + c + d]$). Thus, 29% of the 28 patients judged to have spider nevi by the second clinician ($a + b$) would also be judged to have them by the first clinician, and 29% of 28 is 8 (the number of patients we would expect to

find in cell a by chance alone). Similarly, the first clinician's coin came up tails 71% of the time ($(b + d)/[a + b + c + d]$), and 71% of the 72 patients judged to be free of spider nevi by the second clinician ($c + d$) is 51 (the expected value for cell d). As a result, we would expect the 2 clinicians to agree $(8 + 51)/100$, or 59% of the time, on the basis of chance alone, and the remaining potential agreement beyond chance is therefore $100\% - 59\%$, or 41%.

How much of this 41% potential agreement beyond chance was achieved? This is determined by comparing it with the actual agreement beyond chance of 89% – 59%, or 30%, and 30%/41% comes to 0.73, which means that about three-fourths of the potential agreement beyond chance was achieved by our 2 clinicians. This measure of agreement goes by the name κ and is rather like a correlation coefficient.¹ It ranges from -1.0 (where 2 clinicians would be in perfect disagreement), through 0.0 (where only chance agreement was accomplished), to +1.0 (where 2 clinicians would be in perfect agreement). As you can see in the listing of "conventional levels of κ " that appears in the legend for Figure 1-2, the agreement between our 2 clinicians is considered "substantial," and this is the case for many "present/absent" aspects of the physical examination. As you might imagine, agreement is greater still when the 2 examinations are carried out by the same clinician.

Other items on the clinical examination do not fare as well. For example, in one study of the chest examination, the κ for cyanosis, tachypnea, and whispered pectoriloquy was 0.36, 0.25, and 0.11, respectively.⁷

No measure of clinical agreement is ideal, and κ is no exception. Its size is slightly affected by the frequency of the abnormal finding in the group of patients being examined (it is highest when half of the patients have the finding and tails off a bit when the finding is extremely common or uncommon). If your and our interests warrant, we shall come back to this in a subsequent background article.

But, of course, high precision is not enough, for examiners may be consistent but wrong in their assessments. All 5 members of my clinical team occasionally fail to detect a big liver or hear an important diastolic murmur. In other cases, clinicians may be neither precise nor accurate. For example, a group of iridologists was asked to examine the irises of a series of patients and distinguish those with gallstones from those who had sonographically empty gallbladders.⁸ Their clinical agreement was only "slight," with an average κ of 0.18 (about like whispered pectoriloquy). More important, however, their diagnostic accuracy was no better than chance: they missed about half the patients with gallstones (sensitivity, 54%) and diagnosed gallstones in about half the patients with negative sonogram results (specificity, 52%). To understand sensitivity and specificity, we must now shift from determining the precision of the clinical examination to defining the characteristics of its accuracy.

THE CHARACTERISTICS OF THE ACCURACY OF DIAGNOSTIC TESTS

Returning our attention to Figure 1-1, we can examine the accuracy characteristics of the CAGE questions. The 60

patients in cell a of Figure 1-1 answered yes to 3 or 4 of the CAGE questions and constitute 51%, or 0.51, of all the 117 patients ($a + c$) with a positive diagnosis of alcohol dependency or abuse. The shorthand term for this proportion of 0.51, or $a/(a + c)$, is *sensitivity*, and it is a useful measure of how well a diagnostic test (whether a symptom, sign, or laboratory test) detects a target disorder when it is present. The closer the sensitivity to 100%, the more "sensitive" the clinical or laboratory finding.

In the right-hand column are the responses from patients for whom the criterion standard ruled out the diagnosis of problem drinking. The 400 patients in cell d answered yes to 2, only 1, or none of the CAGE questions and constitute 99.8%, or 0.998, of all the 401 patients ($b + d$) who did not have alcohol dependency or abuse. The shorthand term for this proportion of 0.998, or $d/(b + d)$, is *specificity*, and it is a useful measure of how often a symptom, sign, or other diagnostic test is absent when the target disorder is not present. The closer the specificity to 100%, the more "specific" the clinical or laboratory finding. (Of course, clinicians are not interested in sensitivity and specificity as such but in their effect on the interpretation of positive and negative findings, and we shall get to that shortly. Sensitivity and specificity are properties that must be established beforehand, and that is why they are presented here.)

You will observe that the sensitivity of the CAGE questions is not impressive. The number of "true positives" in cell a is almost equaled by the number of "false negatives" in cell c, and the sensitivity of only 51% confirms that it "misses" about half the problem drinkers. On the other hand, the specificity of the CAGE questions is outstanding. The number of "true negatives" in cell d vastly outnumbers the number of "false positives" in cell b, and the specificity of 99.8% confirms that it almost never labels a patient as a problem drinker when this disorder is absent.

Now we can consider the "predictions" we make about our patient according to the foregoing characteristics. Because of the high specificity, virtually every patient in cell a who answered yes to 3 or 4 of the CAGE questions ($a + b$) has the target disorder, alcohol abuse or dependency, and the shorthand term for this proportion $a/(a + b)$, which is 60/61, or 98%, is the *positive predictive value* or posttest probability of having the target disorder (among patients with 3 or more positive answers). Moreover, despite the rather unimpressive sensitivity, most of the patients in cells c and d who answered yes to none, just 1, or 2 of the CAGE questions were in cell d and did not have the target disorder. The shorthand term for this proportion $d/(c + d)$, which is 400/457, or 88%, is the *negative predictive value* or posttest probability of not having the target disorder among those patients with 2 or fewer positive answers. The complement of this negative predictive value, or $c/(c + d)$, describes the posttest probability of having the disorder among those patients with 2 or fewer positive answers, and this other way of saying the same thing is found useful by some clinicians.

The reason that the negative predictive value looks relatively high, despite the low sensitivity, lies in the fact that the proportion of all patients in this study who had alcohol

dependency or abuse, $(a + c)/(a + b + c + d)$, or 117/518, was only 23% to begin with. That is, 100% – 23%, or 77%, of the patients were not alcohol dependent before they were asked any questions. The shorthand term for the previous knowledge contained in this $(a + c)/(a + b + c + d)$ is *prevalence* or, more usefully, the pretest probability of the target disorder (because this pretest probability is the starting point for making clinical use of the test characteristics, we will place it above the “predictions” entries in subsequent figures).

In contrast to this pretest probability of 23% in the clinical article describing the CAGE questions, in our patient, we judged that the pretest probability of alcohol abuse or dependency was 50%. How would the CAGE questions perform in patients like ours? If the patients in the study summarized in Figure 1-1 were like our own patient, we would expect the result shown in Figure 1-3.

As long as the patient “mix” and severity of disease in the CAGE study summarized in Figure 1-1 are similar to the patient mix and severity of disease in our practice, we would expect sensitivity and specificity to remain constant, despite changes from the study’s to our patient’s pretest probability of the target disorder. Thus, the sensitivity (51%) and specificity (99.8%) in Figure 1-3 are the same as those in Figure 1-1.

Notice, however, that the negative predictive value has decreased from 88% to 67% because predictive values must change with changes in the prevalence of the target disorder. One useful way to think about this is to carry through this concept of prevalence. After all, the predictive value of a positive test result is simply the prevalence of the target disorder among those patients with positive test results. Similarly, the negative predictive value is the prevalence of not having the target disorder among patients with a negative test result. No wonder, then, that predictive values must change with a change in the overall prevalence of the target disorder.

BACK TO THE PATIENT

Your patient readily admitted that he had cut down on his drinking, that his spouse and workmates had annoyed him by complaining about his drinking, and that he often needed an “eye opener” to get going in the morning. According to this quick medical history, and given your previous judgment (before you had any knowledge of his responses to any of these questions) that his chances of being alcohol dependent were 50-50 (ie, a pretest probability of 50%), you can follow his response through Figure 1-3 and conclude that his post-test probability of alcohol dependency is 99.6%, or about as certain as you ever can be about any diagnosis.

Your patient helps us make another general point: because he gave a positive response to a diagnostic history whose specificity was extremely high (99.8%), you “ruled in” the target disorder. A simple way of remembering this property of a powerful diagnostic test is the acronym SpPin: when specificity is extremely high, a positive test result rules *in* the target disorder.

Would the laboratory tests you were considering ordering have saved you some time and done a better job of determining this diagnosis? In fact, and in addition to delaying the diagnosis,

		Alcohol Abuse or Dependency		
		Yes	No	
No. of Positive Answers to the 4 CAGE Questions	3 Or 4	510 (True +)	2 (False +)	512
	2, 1, Or none	490 (False -)	998 (True -)	1488
		<i>a</i> <i>b</i>	<i>c</i> <i>d</i>	<i>a+b</i> <i>c+d</i>
			<i>a+c</i> <i>b+d</i>	<i>a+b+c+d</i>
		1000	1000	2000

Figure 1-3 The CAGE Questions for Alcohol Abuse or Dependency When the Pretest Probability Is 50%

Characteristics: sensitivity, $a/(a + c) = 510/1000 = 0.51$, or 51%; specificity, $d/(b + d) = 998/1000 = 0.998$, or 99.8%. Prevalence or pretest probability of having the target disorder, $(a + c)/(a + b + c + d) = 1000/2000 = 50\%$. Predictions: positive predictive value or posttest probability of having the target disorder for patients with 3 or 4 positive responses, $a/(a + b) = 510/512 = 0.996$, or 99.6%; negative predictive value or posttest probability of not having the target disorder for patients with 2 or fewer positive responses, $d/(c + d) = 998/1488 = 0.67$, or 67%; posttest probability of having the target disorder for patients with 2 or fewer positive responses, $c/(c + d) = 490/1488 = 0.33$, or 33%. Abbreviation: CAGE, cut down, annoyed, guilty, eye opener.

their accuracy is much worse. In the same investigation that studied the CAGE questions, the specificities for γ -glutamyl transpeptidase, mean corpuscular volume, and an entire liver function battery were only 76%, 64%, and 81%, respectively.³ Moreover, the hot new test of platelet enzyme activity has a specificity of only 73%.² Thus, in your patient, a simple medical history was not only quicker and easier but also far more specific.

What about his possible ascites? Given that you have established the diagnosis of alcohol dependency, you already can plan his perioperative and postoperative management to prevent, detect, and treat alcohol withdrawal syndromes. Nonetheless, you would like to know whether he has sufficient liver damage to affect his handling of the sorts of drugs he is likely to receive. Given his fractured ankle, the kneeling position required for eliciting the puddle sign is out of the question, and even a test for shifting dullness will cause him considerable pain. He has already been to radiology, and you do not want him to make the trip again for an abdominal ultrasonographic examination if you can avoid it. His uninvolved ankle is not swollen now, and he tells you he has never had ankle swelling in the past. Would this simple medical history for previous ankle swelling be of any use?

Figure 1-4 summarizes a study of 63 patients admitted to a general medical service in Durham, North Carolina.⁹ Of 15 patients with ascites on abdominal ultrasonographic examination (the criterion standard), 14 had a history of ankle swelling, for an impressive sensitivity of 93%. If we applied this sensitivity (93%) and specificity (66%) to our pretest probability for ascites of 50%, the result (shown in Figure 1-5) suggests

		Presence of Ascites on Abdominal Ultrasonography		
		Present	Absent	
History of Ankle Swelling	Yes	14	16	30
	No	1	32	33
		<i>a+c</i>	<i>b+d</i>	<i>a+b+c+d</i>
		15	48	63

Figure 1-4 Relationship Between a History of Ankle Swelling and Ascites

Characteristics: sensitivity, $a/(a + c) = 14/15 = 0.93$, or 93%; specificity, $d/(b + d) = 32/48 = 0.67$, or 67%. Prevalence or pretest probability of having the target disorder, $(a + c)/(a + b + c + d) = 15/63 = 24\%$. Predictions: positive predictive value or posttest probability of having the target disorder for patients with a history of ankle swelling, $a/(a + b) = 14/30 = 0.47$, or 47%; negative predictive value or posttest probability of not having the target disorder for patients with a negative history for ankle swelling, $d/(c + d) = 32/33 = 0.97$, or 97%; posttest probability of having the target disorder for patients with a negative history for ankle swelling (adapted from Simel et al¹⁹), $c/(c + d) = 1/33 = 0.03$, or 3%.

		Presence of Ascites on Abdominal Ultrasonography		
		Present	Absent	
History of Ankle Swelling	Yes	93	34	127
	No	7	66	73
		<i>a+c</i>	<i>b+d</i>	<i>a+b+c+d</i>
		100	100	200

Figure 1-5 Relationship Between a History of Ankle Swelling and Ascites When the Pretest Probability Is 50%

Characteristics: sensitivity, $a/(a + c) = 93/100$, or 93%; specificity, $d/(b + d) = 66/100 = 0.66$, or 66%. Prevalence or pretest probability of having the target disorder, $(a + c)/(a + b + c + d) = 100/200 = 0.5$, or 50%. Predictions: positive predictive value or posttest probability of having the target disorder for patients with a history of ankle swelling, $a/(a + b) = 93/127 = 0.73$, or 73%; negative predictive value or posttest probability of not having the target disorder for patients with a negative history for ankle swelling, $d/(c + d) = 66/73 = 0.90$, or 90%; posttest probability of having the target disorder for patients with a negative history for ankle swelling, $c/(c + d) = 7/73 = 0.10$, or 10%.

that the posttest probability of not having ascites is 90% when the patient denies ankle swelling. Again, this simple element of the clinical history provides powerful diagnostic information: when the *sensitivity* of a symptom or sign is high, a *negative response rules out* the target disorder, and the acronym for this property is SnNout.

However, you may have observed that this study included only 15 patients with ascites, and you may well inquire how confident we should feel about this sensitivity of 0.93. As it happens, the degree of confidence we ought to place in this (or any other) estimate of sensitivity (or specificity) can be calculated and expressed as a confidence interval, within which you can be confident that the true sensitivity resides, say, 95% of the time.¹ In this case, the 95% confidence interval on this sensitivity of 0.93 based on 15 patients runs all the way from 0.81 (not terribly sensitive) to 1.00 (or perfect sensitivity). If, on the other hand, this sensitivity of 0.93 were based on 100 patients with ascites, the 95% confidence interval would run from 0.88 to 0.98, and you would be justified in being more confident that a negative medical history rules out ascites. Thus, you should look for information on the 95% confidence interval for measures of accuracy such as sensitivity and specificity when you read about them.

A FASTER AND MORE POWERFUL APPROACH: THE LIKELIHOOD RATIO

Many of the overviews in this series will describe not only the sensitivity and specificity of specific symptoms and signs but also their likelihood ratios (LRs). This method of describing the accuracy of diagnostic information, once mastered, is much faster and more powerful than the sensitivity and specificity approach.¹ It is shown in Figure 1-6 for ankle swelling and ascites. In brief, an LR expresses the odds that a given finding on the medical history or physical examination would occur in a patient with, as opposed to a patient without, the target disorder. When a finding's LR is above 1.0, the probability of disease increases (because the finding is more likely among patients with than without the disorder); when the LR is below 1.0, the probability of disease decreases (because the finding is less likely among patients with than without the disorder); finally, when the LR is close to 1.0, the probability of disease is unchanged (because the finding is equally likely in patients with and without the disorder).

LRs are related to sensitivity and specificity but possess some advantages for clinicians. In a 2×2 table such as Figure 1-6, the LR for a positive history of ankle swelling is equal to sensitivity/ $(1 - \text{specificity})$ or $0.93/0.33$, or 2.8, indicating that a positive history is almost 3 times as likely to be obtained from a patient with, as opposed to a patient without, ascites. The LR for a negative history of ankle swelling is equal to $(1 - \text{sensitivity})/\text{specificity}$ or $0.07/0.67$, or 0.10, indicating that a negative history is only as likely to be obtained from a patient with, as opposed to a patient without, ascites (and confirming our earlier conclusion that this negative history permitted us to SnNout this diagnosis).

The first advantage of LRs is that the LR for a given finding, when applied to the pretest odds of the target disorder, generates the posttest odds for that disorder. Because the LR is expressed

as an odds, this may at first appear cumbersome, for it means that the pretest probability must also be expressed as an odds (although this is tedious to do by hand, later, we will show you how to avoid the calculations by using the nomogram shown in Figure 1-7). When done by hand, the pretest probability of the target disorder is converted into pretest odds by the formula:

$$\text{Pretest odds} = \frac{\text{Probability of having the target disorder}}{\text{Probability of not having the target disorder}}$$

In Figure 1-6, the pretest probability of ascites is 0.24, and the pretest probability of not having ascites is $1.00 - 0.24$, or 0.76. Therefore, the pretest odds of ascites are $0.24/0.76$, or 0.32, and this can be multiplied by 2.8 (generating a posttest odds of ascites of 0.90) when the history is positive for ankle swelling and by 0.10 (generating a posttest odds of 0.03) when this history is negative.

These posttest odds can then be converted back to probabilities by the formula:

$$\text{Posttest probability of the target disorder} = \frac{\text{Posttest odds}}{\text{Posttest odds} + 1}$$

Thus, the posttest odds of 0.90 following from a positive history of ankle swelling converts (by $0.90/1.90$) to 47%, and the posttest odds of ascites of 0.03 following from a negative history converts (by $0.03/1.03$) to 3%, and you will observe that these are the same values for the posttest probability of having ascites that we generated in Figure 1-4.

The necessity for converting probability to odds and back again can be obviated by using the nomogram shown in Figure 1-7, which has already carried out the conversions for us.¹ You can prove this to yourself as follows: anchor a straightedge at the left margin of the nomogram, at the pretest probability of 24%, and rotate the straightedge until it intersects the middle line of the nomogram at an LR of 2.8, corresponding to a positive history of ankle swelling. It will intersect the right margin of the nomogram at just below 50%. Similarly, rotate the straightedge until it intersects an LR of 0.10 for the negative history and observe that the posttest probability of ascites decreases to 3%.

The second advantage of LRs becomes apparent when we see that the nomogram permits us to determine the probability of ascites when the pretest probability changes from 24% in Figure 1-4 to 50% in Figure 1-5 without having to construct the latter. We can simply reanchor the straightedge at 50% and run it across the LRs of 2.8 and 0.10 as before, intersecting the posttest probability line at about 73% and 10%. The third advantage of LRs is that, unlike sensitivity and specificity (which limit the number of test results to just 2 levels, “positive” and “negative”), they can be generated for multiple levels of the diagnostic test result. At each level, the proportion of patients with the target disorder at this level is divided by the proportion of patients who do not have the target disorder at this same level; the result is the LR for this level. This is shown in Table 1-2, in which LRs for 4, 3, 2, and 1 and no positive responses to the CAGE questionnaire are shown (the awkward, infinitely high LR for 4 positive answers can be avoided if 3 and 4 positive answers are combined, generating an LR of 206 for the combination).

		Presence of Ascites on Abdominal Ultrasonography		
		Present	Absent	
History of Ankle Swelling	Yes	14 (0.93)	16 (0.33)	30
	No	1 (0.07)	32 (0.67)	33
		a+c	b+d	a+b+c+d
		15 (1.00)	48 (1.00)	63

Figure 1-6 Likelihood Ratios for a History of Ankle Swelling in Diagnosing Ascites

Characteristics: sensitivity/(1 – specificity) = likelihood ratio (LR) (of having the target disorder) for a positive test result = $(a/[a + c])/(b/[b + d]) = 0.93/0.33 = 2.8$; (1 – sensitivity)/specificity = LR (of having the target disorder) for a negative test result = $(c/[a + c])/(d/[b + d]) = 0.07/0.67 = 0.10$. Pretest probability: prevalence or pretest probability of having the target disorder, $(a + c)/(a + b + c + d) = 15/63 = 24\%$. Predictions: posttest probability of the target disorder (expressed as odds) = pretest probability of the target disorder (expressed as odds) \times LR for the test result. Positive history, $0.24/0.76 = 0.32 \times 2.8 = 0.90/1.90 = 47\%$. Negative history, $0.24/0.76 = 0.32 \times 0.10 = 0.03/1.03 = 3\%$. Adapted from Simel et al.⁹

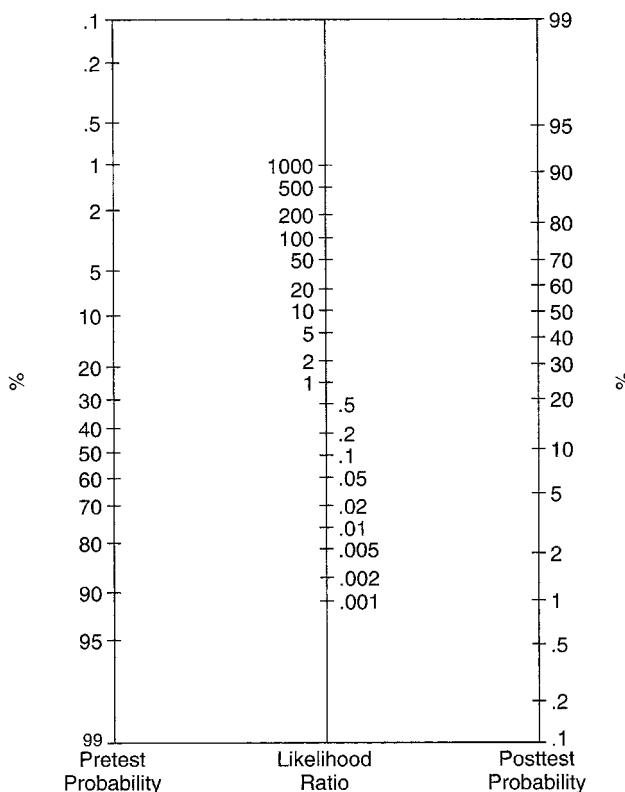


Figure 1-7 A Nomogram for Applying Likelihood Ratios

Adapted from Sackett et al.¹

Table 1-2 Multiple Levels of Responses to the CAGE Questions for Alcohol Abuse or Dependency^{a,b}

No. of Positive Answers to the 4 CAGE Questions	Alcohol Abuse or Dependency		Likelihood Ratios
	Yes	No	
4	23 (0.20)	0	∞
3	37 (0.32)	1 (0.002)	127
2	28 (0.24)	14 (0.03)	6.8
1	11 (0.09)	28 (0.07)	1.3
0	18 (0.15)	358 (0.89)	0.17
Total	117	401	

Abbreviation: CAGE, cut down, annoyed, guilty, eye opener.

^aAdapted from Bush et al.⁵^bNumbers in parentheses are proportions of the respective columns.

The fourth advantage of the LR strategy is that the posttest probability of the target disorder obtained from the first item of diagnostic information (say, a history of ankle swelling) is the pretest probability of that diagnosis for the next item of diagnostic information (say, the physical examination for ankle edema). This example also identifies the problem we always face when we combine diagnostic information from the medical history and physical examination (and chemistry laboratory, and radiology suite!): the results of the medical history and physical examination are not independent from each other. Thus, a patient with a positive history of swollen ankles is far more likely to have pedal edema than a patient with a negative history, and we must either use an LR that considers both of the 2 items as a pair or modify the LR for the second, according to the results of the first. This issue of independence, along with the consideration of the site (primary care or a tertiary hospital) where the examination is carried out, will be taken up in a subsequent background article in this series.

		Target Disorder		
		Present	Absent	
Diagnostic Test Result	Positive	a b		a+b c+d
	Negative	c d		c+d
		a+c b+d		a+b+c+d

Figure 1-8 Working Table for the Reader's Use**For accuracy****Sensitivity** = $a/(a + c)$; SnNout: when sensitivity is high, a negative test result rules *out* the target disorder**Specificity** = $d/(b + d)$; SpPin: when specificity is high, a positive test result rules *in* the target disorder.**Positive predictive value** or posttest probability of having the target disorder among patients with positive test results, $a/(a + b)$.**Negative predictive value** or posttest probability of not having the target disorder among patients with negative test results, $d/(c + d)$. Posttest probability of having the target disorder for patients with negative test results, $c/(c + d)$.**Prevalence or pretest probability** of having the target disorder, $(a + c)/(a + b + c + d)$.**Sensitivity/(1 - specificity) = likelihood ratio (LR) (of having the target disorder) for a positive test result** = $(a/[a + c])/(b/[b + d])$.**(1 - sensitivity)/specificity = LR (of having the target disorder) for a negative test result** = $(c/[a + c])/(d/[b + d])$.**Posttest probability** of the target disorder (expressed as odds) = pretest probability of the target disorder (expressed as odds) \times LR for the test result.**For precision (and κ)****Observed agreement:** $(a + d)/(a + b + c + d)$ **Expected agreement:**Expected cell a, $([a + b] \times [a + c])/(a + b + c + d)$ Expected cell d, $([c + d] \times [b + d])/(a + b + c + d)$

Calculate expected agreement as (expected a + expected d)/(a + b + c + d);

Agreement beyond chance = κ = (observed agreement - expected agreement)/(100% - expected agreement)**Conventional levels of κ :** slight, 0.0-0.2; fair, 0.2-0.4; moderate, 0.4-0.6; substantial, 0.6-0.8, almost perfect, 0.8-1.0.**CONCLUSION**

This first background article has described readers' guides for articles about diagnostic information and has shown how diagnostic data derived from the medical history and physical examination can be assessed for their precision and accuracy. It concludes with a working table (Figure 1-8) and glossary that can be photocopied or clipped. Kept handy, they can help readers study and understand the overviews published in this and subsequent issues of the series on the rational clinical examination.

Author Affiliations at the Time of the Original Publication

Departments of Medicine and Clinical Epidemiology and Biostatistics, Faculty of Medicine, McMaster University, and the Hamilton Civic Hospitals, Hamilton, Ontario, Canada.

Acknowledgments

I wish to thank Andreas Laupacis, MD, David Simel, MD, Richard Deyo, MD, Drummond Rennie, MD, James Nishikawa, MD, Charles Goldsmith, PhD, and the *JAMA* reviewers for their kind and helpful responses to earlier versions of this background article.

REFERENCES

1. Haynes RB, Sackett DL, Guyatt GH, Tugwell P. *Clinical Epidemiology: How to Do Clinical Practice Research*. 3rd ed. Philadelphia, PA: Lippincott Williams & Wilkins; 2005.
2. Tabakoff B, Hoffman PL, Lee JM, Saito T, Willard B, de Leon-Jones F. Differences in platelet enzyme activity between alcoholics and nonalcoholics. *N Engl J Med*. 1988;318(3):134-139.
3. Ewing JA. Detecting alcoholism: the CAGE questionnaire. *JAMA*. 1984; 252(14):1905-1907.
4. Lundberg GD. Ethyl alcohol: ancient plague and modern poison. *JAMA*. 1984;252(14):1911-1912.
5. Bush B, Shaw S, Cleary P, Delbanco TL, Aronson MD. Screening for alcohol abuse using the CAGE questionnaire. *Am J Med*. 1987;82(2):231-235.
6. Theodossi A, Knill-Jones RP, Skene A, et al. Inter-observer variation of symptoms and signs in jaundice. *Liver*. 1981;1(1):21-32.
7. Spiter MA, Cook DG, Clarke SW. Reliability of eliciting physical signs in examination of the chest. *Lancet*. 1988;331(8590):873-875.
8. Knipschild P. Looking for gall bladder disease in the patient's iris. *BMJ*. 1988;297(6663):1578-1581.
9. Simel DL, Halvorsen RA Jr, Feussner JR. Quantitating bedside diagnosis: clinical evaluation of ascites. *J Gen Intern Med*. 1988;3(5):423-428.

UPDATE : Primer on Precision and Accuracy

Prepared by David L. Simel, MD, MHS

Reviewed by Sheri Keitz, MD, PhD

UPDATED SUMMARY ON PRECISION AND ACCURACY OF THE CLINICAL EXAMINATION

Original Review

Sackett DL. A primer on the precision and accuracy of the clinical examination. *JAMA*. 1992;267(19):2638-2644.

WHAT IS THERE TO UPDATE?

Each of the updates in The Rational Clinical Examination systematically evaluates the newly published literature on the topic, except this one. Updating the Primer requires a different approach to fulfill the original promise that the series would address methodologic concerns beyond precision and accuracy. What we will do is take a very utilitarian approach, driven by the topic updates themselves. The updates and our own lectures on the rational clinical examination unearthed topics that we need to address. Rather than conducting a systematic review of quality measures, sensitivity, specificity, likelihood ratios (LRs), and a plethora of related topics, we instead provide background information and answers to questions that our own authors required when preparing their reviews and updates.

Of course, the basic premise for diagnosis has not changed since the Primer (or since Thomas Bayes figured it out more than 3 centuries ago):

$$\text{Prior odds} \times \text{LR} = \text{Posterior odds}$$

For the clinical examination, this means we (1) use information about the probability of a target disorder (frequently taken as the prevalence, which is then converted to the prior odds) and then (2) apply the results of symptoms or signs (in the form of an LR). After applying the LR associated with various symptoms and signs, we get the posterior odds of disease. The probability of disease increases when a clinical finding is more likely in a patient with the target disorder (reflected by an $\text{LR} > 1$). The probability of disease decreases when a clinical finding is more likely to occur in a patient without the target disorder (reflected by an $\text{LR} < 1$). The resultant probability becomes the “posterior” probability because the prior probability is established first and then modified with information from the medical history and physical examination quantitatively expressed in the form of the LR.* Keeping the simple equation

in mind focuses the goal of The Rational Clinical Examination series articles on providing all the data needed to solve the posterior odds equation.

Why LRs?

In the Primer, we emphasized the role of the univariate LR for clinicians. The term *univariate* means the results for 1 finding, without regard to the findings of other historical or clinical features. We chose this route for a variety of reasons, most important being its fundamental property that allows clinicians to apply the values to individual patients in a consistent pattern. LRs always convey the same information—they quantify the change in odds of disease for a particular test result. By tradition for dichotomous test results, we call the LR associated with a positive test the LR^+ (positive LR), whereas the LR associated with a negative test is the LR^- (negative LR). In either case, the actual LR value is related to the change in likelihood that the patient has the disease of interest. Thus, there can be no confusion, as is sometimes the case when physicians become overwhelmed with how to translate positive predictive value, true-positive rate, false-positive rate, negative predictive value, true-negative rate, or false-negative rate into a change in the likelihood of disease for an individual patient.

Many clinicians feel more comfortable with the terms *sensitivity* and *specificity*. However, these values in and of themselves have little application to the clinical setting. Sensitivity and specificity are values that apply to a screening test result before we know whether the patient has the target disorder. So which result do we use at the bedside? Sensitivity applies only to patients with disease, whereas specificity applies only to patients without disease. Because we use screening tests precisely because we do not know about the presence or absence of disease, how do we decide whether the value of

*Do not be confused by the transition between odds in the equation and our discussion of probability. The equation requires that we use the odds ratio, but clinicians find it easier to think in terms of probability. We can convert any probability of disease to the odds ratio by the equation odds = probability of disease/probability of no disease. After we convert the prior probability to odds and multiply it by the LR to get the posterior odds, we convert the result back to the probability of disease by the equation probability = odds/(1 + odds).

Table 1-3 Examples of Symptoms or Signs That Have Results Other Than Just “Present” or “Absent”

Example	Screening Test	Multilevel Outcome
A symptom reported by the patient	“Do you have trouble initiating your urine stream?”	“Always” “Frequently” “Sometimes” “Never”
A sign on the physical examination	Is a third heart sound present?	Abnormal Uncertain Normal
Ordinal ^a valued findings	Deep tendon reflexes	4+ 3+ 2+ 1+ 0

^aOrdinal means “ordered.” The results can be ranked, although the incremental value has no quantitative meaning. For example, deep tendon reflexes of 2+ are more pronounced but not twice as prominent as 1+ reflexes.

Table 1-4 Hypothetical Data to Demonstrate How to Describe the Results for a Finding With 3 Possible Outcomes

	LV Systolic Dysfunction Present	Normal LV Function
S3 definitely present	30	5
Uncertain	5	10
S3 definitely absent	10	50

Abbreviation: LV, left ventricular.

sensitivity or the value of specificity applies to our patient? The simple answer is that we do not know. If we do know which result applies to our patient, then, by definition, we know the disease status, and the results of screening tests lose relevance. The true value of an LR comes from its mathematical definition that combines the values of sensitivity and specificity, making it applicable to each patient before we know whether disease is present or absent.

When evaluated in combination, the sensitivity and specificity are the building blocks of the LR for tests that are dichotomous (eg, “positive” or “negative,” “present” or “absent”). The LR for a positive result is sensitivity/(1 – specificity), whereas the LR for a negative result is (1 – sensitivity)/specificity. But what happens when a screening test has more than 2 outcomes (Table 1-3)?

Traditional laboratory tests are measured on continuous scales, where the result intervals have a mathematical meaning, but the clinician could not possibly know the LR for every outcome. A clinical laboratory reports the raw result, along with a designator for whether the result is “high,” “normal,” or “low.” The report takes the raw value and transforms it to an ordinal scale, making it easier for clinicians to review a large amount of data. When there are more than 2 outcomes of a screening test, sensitivity and specificity cannot be

directly calculated, so the clinician must rely on LRs that are usually given for ordinal results.

A simple quantitative explanation helps explain why the sensitivity and specificity lose meaning when there are more than 2 screening test results. The presence of a third heart sound (S3) suggests left ventricular (LV) systolic dysfunction. Sometimes, the clinician is uncertain whether the sound is present. To illustrate this point, we can make up some data that might apply to the clinician’s interpretation of the S3 compared with a reference standard echocardiogram that quantified the LV function (Table 1-4).

We can describe the sensitivity of the S3 as $30/(30 + 5 + 10) = 0.68$ and the specificity as $50/(5 + 10 + 50) = 0.77$. Although this may seem straightforward, closer inspection reveals some problems with that interpretation. First, the treatment of the “uncertain” results lacks consistency. For calculating the sensitivity, we “count” an uncertain S3 as if it were actually absent. But the clinical reality was that the physician could not state with certainty whether it was present or absent. When we calculate the specificity, we do the exact opposite and count the “uncertain” outcomes as if they were “positive.” How can one “uncertain” finding be considered “positive” for sensitivity but “negative” as specificity? This dual treatment creates problems that become even more pronounced as the number of results increases beyond 3 outcomes.

Second, even if we believed that the sensitivity and specificity captured the meaning of an S3 that is either present or absent, how do we describe the results for “uncertain?” Sensitivity provides an inadequate definition because sensitivity is the value that describes the percentage of patients with an abnormal result among all those with disease and “uncertain” is neither abnormal nor normal. A similar argument applies to the specificity, so that neither sensitivity nor specificity offers a reasonable description of the value of an uncertain result. The constructs just do not apply to a test result that is neither completely normal nor completely abnormal. The LR provides a way to describe not only the positive and negative results but also those that are uncertain.

At a fundamental level, the LR takes a given screening test result and for that outcome tells us the ratio of those with disease to those without disease. So once we know which row of the table a patient belongs in according to their test result (S3 present, S3 uncertain, or S3 absent), the LR tells us the likelihood that the patient will come from the first column vs the second column. We can calculate an LR for every row of an $r \times 2$ table (where r represents the number of rows) (Table 1-5).

Thus, when we hear an S3 in the patient, we apply the value 8.7, which makes LV systolic dysfunction much more likely. When we feel confident that an S3 is absent, the likelihood of LV systolic dysfunction decreases. However, when we are “uncertain,” the LR we apply is 0.72, a value that approaches 1 and suggests that the “uncertain” result should not have a large effect on our estimate of the likelihood of disease. Oftentimes, it is useful to know that “uncertain” really means “not much information” with an LR approaching 1.

Isn't All the Information in the Patient's Medical History?

We now need to address a common belief that the physical examination is not particularly helpful and, at best, only confirms the historical findings and symptoms. Oftentimes, a clinician takes a patient's medical history and makes a diagnosis before performing a physical examination. This process, although sometimes successful, leads to the inference that the physical examination was unnecessary. For a simple reason, the inference is not true: the physical examination begins from the moment the clinician meets a patient and before the patient utters a word! We observe body language, the patient's gait, vital signs (eg, tachypnea), and physical deformities, and we judge the acuity of illness. These findings derived from visual observations may be hard to quantify (eg, a sense that the quiet, sullen patient might be depressed), although most clinicians recognize the huge amount of information they collect in the first few moments of a patient interaction. Because describing and measuring the influence of our overall observations is difficult, researchers often overlook the clinical gestalt.

One way of isolating the clinical gestalt is to evaluate whether we can make a diagnosis in the absence of directly observing a patient. A symptom checklist (but not the patient's medical history) can be obtained through a completed patient self-administered questionnaire. Sometimes, we can infer a diagnosis from such questionnaires with our impression uncontaminated by physical findings, but the diagnosis typically requires confirmation obtained through a patient interview or physical examination. The ability to disentangle the history from the physical examination findings is often an illusion, leading to the inference that the patient's medical history (symptoms) dominates the clinical diagnostic process over the physical examination (signs).

The Pretest Probability

The most important part of the clinical examination and the resulting diagnosis is typically not the symptoms or signs—it is the pretest probability, transformed to the prior odds, that dominates the equation. Simply put, if a condition is highly unlikely (or vice versa), then the presence or absence of any addition findings will typically not change things. As a corollary, when the probability of a target condition is not so certain, the effect of the signs and symptoms on the prior probability creates a potentially bigger effect.

So, where does the pretest probability come from? We establish the pretest probability in the course of our clinical examination, and that creates a bit of a problem (for both researchers and clinicians). In other words, as we learn more about the patient's medical history, symptoms, and signs, we orient our approach to a narrower spectrum of disease possibilities. This approach requires that we "waste" a few findings to establish the pretest probability. For example, most patients we examine do not have sinusitis, and we do not ask questions about symptoms related to sinusitis, nor do we transilluminate the sinuses during the course of a clinical examination unless we have a suspicion of the disease. We might constrain our evaluation for sinusitis to patients who claim nasal stuffiness, nasal discharge, or maxillary facial discomfort or who come right out and state, "I think I have a sinus infection." Each of these findings would prompt an

Table 1-5 A Likelihood Ratio Can Be Calculated for Each Row of an $r \times 2$ Table as Shown With These Hypothetical Data

	LV Systolic Dysfunction Present	Normal LV Function	LR ^a
S3 present	30	5	(30/45)/(5/65) = 8.7
S3 uncertain	5	10	(5/45)/(10/65) = 0.72
S3 absent	10	50	(10/45)/(50/65) = 0.29
Total	45	65	

Abbreviations: LR, likelihood ratio; LV, left ventricular.

^aBy convention, for LR values 0-1, we round off to the 100ths; for LR values 1-10, we round off to the tenths; and for LR > 10, we round off to the nearest integer.

appropriate evaluation for sinusitis and in a research study create the "entrance criteria." Thus, when we refer to the pretest probability of sinusitis, we most likely are referring to the prevalence of sinusitis among patients with any of those findings rather than to the prevalence of sinusitis among all patients in general. This pretest probability becomes the value we use in the equation and the anchor for applying other symptoms and signs we uncover during our clinical examination.

The establishment of the pretest probability is the problem most learners fear, representing their main "excuse" for not using the concepts in The Rational Clinical Examination. Frequently, learners claim "lack of experience." When existing studies adequately describe their study population, the pretest probability is not difficult to understand. Experience becomes more valuable when the literature is less clear, and perhaps this is part of the "art" of the clinical examination. Trainees may be quite good at estimating the pretest probability of common conditions. However, both trainees and experienced clinicians tend to overestimate the prior probabilities of less common diseases. Trainees express discomfort when estimating the prior probability because (1) they do not practice quantifying and then validating their clinical impression and (2) they may recall their own cases in which they pursued an unlikely diagnosis for a seemingly "classic" presentation, only to find that the disease was not present. Although the second reason emanates from overlooking the importance of prior probability, it requires a reassessment of the role of symptoms and signs.

What Is a "Good" Symptom or Sign?

The presence of a "good" symptom or sign creates a large effect on the probability, convincing the clinician that the target condition is much more likely to be present than the prior probability suggests. The suggestion that some prespecified LR threshold defines a good clinical finding for all disease is a myth so persistent that it represents a medical urban legend. Some researchers and clinicians define a "good" test result as that associated with an LR greater than 10 or an LR less than 0.1, but these results do not have intrinsic properties that are the sine qua non of high value. For example, a pretest probability of 10% and positive test with an LR = 10 generates a posttest probability of 53%; this is a big increase in the probability of disease but hardly an increase that

clinches the diagnosis. Furthermore, this is a similar posttest probability that follows from a disease with a pretest probability of 20% and a positive test with an $LR = 5$. Thus, although positive test results are increasingly powerful as the LR increases and negative results are increasingly valuable as the LR decreases, the efficiency of the finding in making a diagnosis depends on the pretest probability.

When considering that multiple symptoms and signs are interpreted together, individual findings with much less impressive LRs alone (eg, $LR+$, 2-5; or $LR-$, 0.25-0.50) could prove useful when used in combination. If no LR threshold automatically qualifies a result as good, is there a way to compare the efficiency of different clinical findings?

A positive clinical finding with the highest $LR+$ or a negative finding with the lowest $LR-$ will always have the greatest effect on posttest probability. Unfortunately, clinicians discover that a list of symptoms and signs for an individual patient sometimes simultaneously yields outcomes both suggesting (positive results) and pointing away from (negative findings) a target disorder. There is a way, though, to make sense of this. Rank ordering the $LR+$ associated with each result, along with the reciprocal of the $LR-$ ($1/LR-$), reveals the single “best” clinical finding for a target condition. The value with the highest $LR+$ or $1/LR-$ is the single best symptom or sign result. A single symptom or sign may be useful when present (high $LR+$) or absent (small $LR-$). Unfortunately, most symptoms and signs will not produce both the best findings when positive and also the best when it is negative. For example, a clinical sign may have a low $LR-$ when negative, whereas a positive result may have an $LR+$ that approaches 1. Creating a mental list of LR and $1/LR-$ for a variety of symptoms and signs is not easy. Some clinicians want to identify the single finding that overall is the most likely to give them the right answer (ie, positive when the patient has disease and negative when the patient is not affected).

The diagnostic odds ratio (DOR) creates a single measure of accuracy that tells us which symptom or sign is most likely to correctly classify a patient as having the target disorder or not.¹ The DOR is not difficult to calculate, as the $DOR = LR+/LR-$. The more accurate the symptom or sign, the higher the DOR. So when faced with a table of data on many clinical findings in which none distinguishes itself as the overwhelming favorite, the clinician should choose the finding with the highest DOR. Unfortunately, the DOR cannot be used like the LR for estimating the probability of a diagnosis, but it can help us choose the symptoms and signs of higher utility so that we can ignore those of lesser value. At this point, the skeptical reader might accept that there is a method for identifying better symptoms and signs in terms of their overall measurement properties (through the DOR) and better results applicable to individual patients (through the LR). However, a remaining question might be, How confident can I be that the symptoms and signs I think are the best really are the best?

The Confidence Interval

When The Rational Clinical Examination series began, we presented likelihood results as single point values as if they

completely described a clinical finding—they do not. Like all statistical parameters, an LR has an associated confidence interval (CI) that helps us decide whether the data are sufficient for us to infer usefulness. These CIs are important because they provide transparency. An optimistic LR suggests a promising clinical finding, but a broad CI dampens the enthusiasm by implying that a small sample size accounts for some certainty. We are particularly cautious when the 95% CI includes 1 because LR values of 1 add no information to the pretest probability. Broad CIs around $LR-$, even when they do not include 1, are a particular problem. Because the $LR-$ values are constrained between 0 and 1, a broad CI seems less of a problem than the broad CI around a high $LR+$. To compare the relative findings, the clinical reader can use the technique we described above (ie, taking the value $1/LR-$) for comparing the breadth of the CIs of negative to positive LRs.

Some readers will be surprised that there are different methods that yield slight (but clinically unimportant) differences in CIs. We prefer the easiest computational method that also works well in spreadsheets.² One situation presents problems for researchers and clinical readers alike: what do we do when one cell of the 2×2 table is 0? When any single cell has a 0 value (typically, the cells for false positive or false negatives), adding 0.5 to each cell of the 2×2 table allows calculation of useful CIs.³ A sensitivity of 100% yields an $LR-$ of 0, with the LR upper 95% CI obtained after adding 0.5 to each cell. A specificity of 100% yields an $LR+$ that is not calculable (∞), so we report both the $LR+$ and CI obtained after adding 0.5 to each cell. Although high-quality studies report both the sensitivity and specificity of clinical findings, not all of them calculate the LRs for us. When researchers provide the actual numbers of affected and unaffected patients, together with the sensitivity and specificity, we can generate the LRs and 95% CIs. Although it is sometimes easy to calculate CIs from individual research reports, meta-analysis offers us an even better way of describing the LRs of findings evaluated across several studies.

Meta-analysis

Meta-analysis of symptoms and signs combines the results described across several studies and summarizes them to get a single estimate and CI. Although some statisticians have a high degree of skepticism about the appropriateness of combining LRs, we take the position that summarizing results provides clarity for clinicians that at the very least allows them to assimilate data and decide whether a symptom or sign is useful, useless, or uncertain.

An important part of meta-analysis requires the investigator to make decisions about the appropriateness of combining data. Although statisticians often suggest a purely statistical approach (ie, studies that have statistically heterogeneous results should not be combined), we take a more pragmatic approach similar to that espoused by other clinical diagnosticians.⁴ First, we evaluate whether the universe of published studies represents the universe of patients for whom the target condition might be considered. When the

studies reflect the population of patients for whom the symptoms and signs apply, we prefer to try combining the LRs. On the other hand, when studies use various definitions of disease or different thresholds for the symptoms and signs, we cannot combine the results in a meaningful way. When we cannot combine the results, we present ranges for the LRs. Second, we consider our target audience to be clinical readers. For a condition that might have a very different LR among different populations of patients (eg, findings for appendicitis among children vs geriatrics patients), we avoid combining results or we at least show how they vary. Part of this approach requires common sense, and part of this is statistical, in which we examine the outlier results to deduce whether there is anything recognizable that accounts for the variant LR findings. Third, we examine the actual results with their CIs after we combine the data. We always use random-effects measures for generating the LR and CIs, rather than the fixed-effects approach. Random-effects measures generate broader CIs than the fixed effects, providing at least some assurance that we are not overstating the importance and confidence in our findings. If a study is a statistical LR outlier, we still include it in the combined data if it does not make a large clinical difference in the LRs. We suggest that the clinician use clinical judgment when deciding whether 2 LRs yield clinically important differences in the posttest probability. For example, for a pretest probability of 30%, an LR of 5.4 produces a posttest probability of 70%, whereas an LR of 3.5 produces a posttest probability of 60%. These LRs “look” different, but a clinician might take a similar action for a posttest probability of 70% vs 60%. Thus, the 2 LRs could be statistically different but provide clinically similar results. We always provide the results from each study, and astute readers can decide from the point estimates and CIs whether they believe a finding is useful or useless.

More statistically experienced readers may recognize that meta-analysis of LRs differs from what they expect. Statisticians, when they accept meta-analysis of diagnostic tests at all, prefer summarizing the DOR as a global measure of test performance. We take a different approach because summarizing the DOR gives clinicians a value that they cannot use for individual patients. Although we do sometimes provide summary measures of the DOR, the summary measures of the prevalence of disease (pretest probability) and the LR are the values needed for solving the equation for posttest probability. Sometimes, we encounter studies that only provide sensitivity data. What do we do with studies that are case series of patients with disease and that do not have specificity values?

Sensitivity-Only Studies

When conditions are less common, investigators recognize that enrolling consecutive patients at risk for the target disorder creates a study population overwhelmed by those without disease. This approach is costly and takes time, and the small number of patients with disease leads to broad CIs around the sensitivity and LR-. The alternate approach of studying only patients with disease so that sensitivity can be

defined is pragmatic, and it may be the best the investigator can do. These studies typically come from a narrow spectrum of diseased patients, and often, the clinical finding is recorded among patients when the clinician knows that disease is present. In addition to understanding the potential biases in the data, we must understand the inferences made from the sensitivity of symptoms and signs without specificity values. The goal of sensitivity studies is to identify a group of symptoms and signs that would unlikely all be negative in a patient with the target condition.

Symptoms and signs with high sensitivity are less likely to be negative in patients with disease. When presented with sensitivity data by itself, clinicians will count the number of absent findings in their patients and deduce that those with normal findings on multiple high-sensitivity symptoms and signs will be unlikely to have disease. For example, suppose we identify 2 symptoms and 1 sign, each of which has a sensitivity of 85% for the target condition. That means that each finding would be absent in 15% of patients with disease; all 3 would be absent in fewer than 1% of patients ($0.15 \times 0.15 \times 0.15$).

How Do We Use All the Symptoms and Signs?

Among several reasons for preferring LRs as our common statistical parameter, rather than the individual sensitivity and specificity values, the ability to multiply likelihood results from several findings is the most alluring. Unfortunately, a crucial assumption is not often fully addressed—sequentially multiplying LRs requires that the symptoms and signs be independent of one another.

Let us explain the independence concept with a simple example. Suppose you conduct a study of chest pain symptoms as a predictor of acute ischemia and you categorize words as having “physical” or “emotional” connotations. Words that describe location and radiation would be physical (eg, “center of the chest,” “in the neck”), whereas words that describe the interpretation of pain would be emotional (eg, “suffocating,” “crushing”). You decide to record whenever a patient refers to an “elephant” in describing their discomfort as emotional as in, “It felt like an elephant stepped on my chest.” We suspect it is obvious that a patient who is “elephant-positive” is experiencing crushing pain, but if they report they are having “crushing pain that feels like an elephant on my chest,” should we report the findings separately for “crushing positive” and “elephant positive?” Multiplying the LRs together for “crushing,” “elephant-like” discomfort probably overstates the importance, producing a posttest odds ratio that is too high because elephant-like pain is not independent of crushing pain. Although common sense might work as an initial judge of independence, common sense should not be the only arbiter of independence. What should you do when presented with an array of findings for many symptoms and sign without any assessment of independence?

To make teaching and performing the medical history and physical examination more efficient and accurate, we want parsimony. By “parsimony,” we mean the fewest number of symptoms and signs that yield the most accurate information. Parsimonious examinations force teachers to teach only

the most relevant parts of the examination, allowing students to spend more time learning what is important while eliminating wasteful maneuvers. Of course, some of this waste is in eliminating maneuvers that do not work well. For example, a Rinne test is interesting to teach, but it does not add useful diagnostic information to the symptom of “decreased hearing” reported by the patient.⁵ We eliminate additional wasted effort when we discard nonindependent findings.

A parsimonious examination should mathematically make us more accurate because a “complete” medical history and physical examination almost certainly produces nonindependent findings. “Positive” nonindependent findings confuse us and distort our probability estimates, typically making us infer a higher probability of disease than is justified. Most authors of The Rational Clinical Examination articles emphasize no more than 3 to 4 findings, even when additional symptoms and signs have useful LRs. Narrowing down the number of recommended findings requires “face validity,” by which we mean using common sense to recommend the items with the best, seemingly independent LRs. When we take this approach, experienced clinicians then use semi-quantitative reasoning and deduce that the more findings present, the more likely the patient has disease (or vice versa).

When clinicians want to incorporate the results of diagnostic studies into their decision making, they can take 3 approaches to prevent errors created by lack of independence.⁶ Performing the clinical examination and then using only one single history or physical examination finding to adjust the prior odds will guarantee there is no problem with independence. (Of course, it also guarantees that the clinician might be ignoring a lot of useful clinical information!) Typically, the clinician will want to use the single finding that has the greatest effect on the prior odds, or the “best” finding that we described earlier. The approach is not difficult since simple math allows you to rank the findings in order from most useful to least useful. Suppose you have 3 findings (A, B, and C) that can each be positive or negative, with the LRs associated with each result shown in Table 1-6. Is the finding

that “A” is present more diagnostically useful than “C’s” absence? To determine this, you can rank order these by comparing the LR for the positive results to 1/LR for the negative results. Table 1-6 shows the relative value each of the findings. If your patient had “A” absent, “C” present, and “B” present, then you would multiply the prior odds by the LR associated with the outcome for test “B” (LR = 5.0) because it had had the most useful outcome for that individual.

Although the above result removes any concerns with independence, the clinician must collect many data that ultimately are discarded. At the very least, it is not efficient, and at the worst, important information could be ignored. Not surprisingly, this approach lacks appeal because it ignores the way most clinicians incorporate many bits of information into their decision making.

Clinical researchers must analyze their data in a multivariate way to help clinicians. By “multivariate,” we mean that they must analyze combinations of findings so that there is less concern about independence. This can involve one of 2 general approaches. The easiest approach is to take the medical history and physical examination findings and perform logistic regression. Logistic regression takes a number of individual variables and determines their importance in predicting whether disease is present or absent. In the first strategy for assessing independence, logistic regression identifies variables that lack independence and that can be eliminated as redundant. In our example above, if all patients with wheezing were also dyspneic, then the finding on the “variable” dyspnea might be unimportant once we know the wheezing status. The logistic regression approach would identify this as being nonsignificant, and the investigator would suggest we concentrate our efforts at assessing for wheezing. Used as a “data-reduction” step to achieve parsimony, the clinician would use the simple, univariate LRs for any finding identified as being independently useful in the logistic model. This approach has a lot of appeal because it identifies the important and useful variables for the clinician, and it does not require that they understand the logistic model itself, because the univariate LRs are used. However, in using the simple, unadjusted LRs, we ignore the relationship between the various clinical findings in favor of simplicity.

The β parameters of a multivariate logistic analysis describe the relative importance of symptoms and signs. From algebra, you might remember the equation for a straight line is $y = mx + b$. The m in the equation is the slope, and it quantifies how a change in x affects y .^{*} A logistic model works similarly, except that now, rather than having 1 x , we have several symptoms and signs that we evaluate all at once. The equivalent of m in the logistic model now represents the β parameter, which is the odds ratio associated with each symptom or sign; the higher the β parameter, the more important the finding. When investigators provide us the actual multivariate models, we can put the results of our own patient’s clinical examination into the model, and the outcome is the individual patient’s actual probability of disease.

Table 1-6 The Findings With the Biggest Influence Can Be Found by Rank Ordering the LR+ and LR-^a

Finding	LR	LR for Values > 1 and 1/LR for Values < 1 ^b
A present	15	15
C absent	0.1	10
B present	5.0	5.0
C present	2.0	2.0
B absent	0.6	1.7
A absent	0.9	1.1

Abbreviations: LR, likelihood ratio; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

^aAdapted from Hollerman and Simel.⁶

^bFor LRs < 1.0 (usually the LR-), the reciprocal (1/LR) is used.

*For those who just cannot remember b , it is the intercept where the line crosses the y -axis.

The Fuss About Precision

The Primer states, “for an item of the clinical history or physical examination to be accurate, it first must be precise.” By precision, we imply that 2 or more observers agree on the presence or absence of a finding in a patient who experienced no clinical changes.*

When we measure precision, describing the percentage of time that 2 observers agree on a symptom or sign fails to consider simple luck. Instead of reporting simple agreement, investigators report precision as the agreement beyond that attributable to chance. For dichotomous findings (“yes” vs “no” or “present” vs “absent”) compared between 2 observers, we quantify this agreement beyond chance with the κ statistic.[†] The κ statistic varies from -1 (perfect disagreement) to 0 (chance agreement) to +1 (perfect agreement).

Suppose we are interested in whether a third heart sound identifies patients with LV systolic dysfunction. It is easy to imagine that a cardiologist might be better at identifying this correctly than a generalist internist, suggesting that a κ statistic might show lower agreement beyond chance than if we were comparing 2 general physicians. Should we conclude that a third heart sound is not a good test from the precision between a cardiologist and a general internist? The answer, of course, is no because test accuracy depends on the quality of the observation—the cardiologist might be a better observer than a less experienced clinician. These seemingly imprecise symptoms and signs are potentially useful when certain providers get consistently good results because they represent opportunities for improved performance and accuracy.

A second type of precision is more important for identifying inaccurate findings. Although a low κ between observers points to opportunities for improving, poor intraobserver agreement precludes high accuracy unless the problem can be eliminated. Intraobserver agreement describes whether a clinician gets the same result when assessing a symptom or sign on a patient who is clinically unchanged. For example, when a clinician inquires about unilateral headaches as a symptom for migraines but the patient changes his or her answer, the finding can never be accurate or precise. Although the natural assumption might be to blame the patient for inconsistency, part of poor intraobserver agreement may be attributable to poor technique that can be improved. This is true even when applied to symptoms as reported by the patient because different answers follow when the information is solicited differently (eg, asking the patient a leading question about unilateral headaches vs an open-ended question). But if clinicians cannot assure reliability on their own findings, they will never use the symptoms and signs accurately. If you cannot agree with yourself, the LR results will be random.

*To clarify further, some researchers use the word *reliability* or the term *observer variability* instead of *precision*. These are all terms that imply the same concept of similar results on repeated examinations, so we use them interchangeably.

[†]We use the weighted κ when we have findings that are not dichotomous. For example, a sign graded as 0, 1, or 2+ would have a disagreement between observers of “grade 1 and 2” weighted as less than a discrepancy between “grade 0 and 2.” When we have multiple observers, we use regression techniques to generate the intraclass correlation coefficient for describing the interobserver variability.

A Brief Word About Quality

Every article in The Rational Clinical Examination series and the updates in this book use a standard process for assessing the quality of data. Although the Primer focuses mostly on the sensitivity, specificity, and LR results, it should be clear that narrow CIs around the results do not assure methodologic rigor of the studies that generated the results. At the inception of The Rational Clinical Examination series, the evidence-based medicine movement was in its infancy. An early article in the series heralded its entry into the mainstream thoughts of clinical educators and investigators.⁷ Because standardized approaches had not been developed for assessing the quality of the medical history and physical examination, David L. Sackett, MD, and Charles H. Goldsmith, PhD, agreed on certain characteristics that they asked their reviewers to use when judging quality. The criteria were simplified and summarized in an early article of the series.⁸ Subsequently, several groups have published their criteria for the review of diagnostic accuracy studies, although none address the particular nuances of symptoms and signs.⁹⁻¹¹ Perhaps it is not surprising that many clinical investigators and epidemiologists have reported on a large number of quality measures that describe what seem like innumerable potential biases in diagnostic test studies. Despite the increasing complexity of rating systems and quality measures, the original criteria for reviewing articles have stood the test of time and pragmatism. If anything, we made the process easier and reduced the number of quality levels a reviewer might assign an article. We reviewed the recommendations for diagnostic test studies^{9,10} and adapted them specifically for studies of the clinical examination.¹² In the early articles appearing in The Rational Clinical Examination series, we assigned Grades for levels of evidence. However, this blurred the distinction between Levels 3, 4, and 5. Because no study accepts Level 5 evidence in making recommendations, we dropped the Grade designation and now report only the Levels as shown in Table 1-7.¹³

Table 1-7 Levels of Evidence^a

Level of Evidence	Grade	Definition
1	A	Independent blinded comparison of sign or symptom results with a criterion standard of diagnosis among a large number of consecutive patients suspected of having the target condition
2	B	Independent blinded comparison of sign or symptom with a criterion standard of diagnosis among a small number of consecutive patients suspected of having the target condition
3	C	Independent blinded comparison of sign or symptom with a criterion standard of diagnosis among nonconsecutive patients suspected of having the target condition
4	C	Nonindependent comparison of sign or symptom with a criterion standard of diagnosis among samples of patients who obviously have the target condition plus, perhaps, normal individuals
5	C	Nonindependent comparison of sign or symptom with a standard of uncertain validity

^aModified from Holleman and Simel.¹³

Table 1-8 Hypothetical Data in Which Only the Patients Who Received Neuroimaging Appear in the Published Report

Finding	Target Condition		
	Present	Absent	
Present	90	10	$LR+ = 9.0$
Absent	10	90	$LR- = 0.11$
Sensitivity = 0.90			
Specificity = 0.90			

Abbreviations: LR+, positive likelihood ratio; LR-, negative likelihood ratio.

Most of the important biases that compromise a study's results follow from the study population not being consecutive, prospective, or independently assessed with an appropriate blindly applied reference standard. By consecutive, we mean that the authors enrolled all patients for whom the target disorder was a reasonable consideration. *Independent* means that the symptom or sign under study was not used to select patients for the study. *Blind* means that the symptoms and signs were applied without knowledge of the presence of disease determined by the reference standard, but also that the reference standard was interpreted without knowledge of the study questions. The size of a study (level 1 vs level 2) for quality assessment depends on the disease under consideration. The authors of The Rational Clinical Examination evaluate sample sizes according to their review of the literature because there is no uniform number that determines quality; for example, a large study of thoracic aortic aneurysms might likely not have as many patients as a large study of urinary tract infection in women.

One particular bias, verification bias, deserves special consideration because it can be insidious and have a big effect on the LR. Verification bias occurs when all the potentially eligible patients fail to undergo confirmation of their disease status. Often, this is done for pragmatic reasons. An example might be a study of headache patients that seeks to describe whether asymmetric neurologic findings (eg, weakness) indicating serious intracranial abnormalities were discovered through neuroimaging. Because it would be expensive and impractical to have every patient with headaches undergo imaging, an investigator typically chooses to maximize the chance of finding something by including all patients with asymmetric muscle strength but only a sample of those who are normal. We can highlight the effect of verification bias on the sensitivity, specificity, and LRs, through examining tables of example data. Suppose an investigator reports the findings displayed in **Table 1-8**.

In the example, the finding looks excellent, with a sensitivity and specificity of 90%. However, because the investigator could not justify the reference standard (eg, neuroimaging on every patient with a headache), the investigative team referred only a sample of those with positive clinical findings (for illustrative purposes, 10%). Had the investigator been evaluating every patient, the findings might have been as shown in **Table 1-9**.

The data demonstrate that verification bias tends to overestimate sensitivity while underestimating specificity.* When

Table 1-9 Hypothetical Data, Adjusted for the Patients Who Did Not Receive Neuroimaging

Finding	Target Condition		
	Present	Absent	
Present	90	10	$LR+ = 43$
Absent	$10/0.10 = 100$	$90/0.10 = 900$	$LR- = 0.53$
Sensitivity = 0.47			
Specificity = 0.99			

Abbreviations: LR+, positive likelihood ratio; LR-, negative likelihood ratio.

the bias is left unadjusted, the investigator will not recognize that the presence of the finding is actually better than suggested (the adjusted LR+ should be higher), whereas the absence of the finding is not as good as suggested (the adjusted LR- should be closer to 1). Astute investigators will recognize that if they collect complete data on all the potentially eligible patients, the bias is one of the few in diagnostic test research that can be mathematically corrected.

REFERENCES FOR THE UPDATE

- Glas AF, Ligmer JG, Prins MH, Bonsel GJ, Bossuyt PMM. The diagnostic odds ratio: a single indicator of test performance. *J Clin Epidemiol*. 2003;56(11):1129-1135.
- Simel DL, Samsa GP, Matchar DB. Likelihood ratios with confidence: sample size estimation for diagnostic test studies. *J Clin Epidemiol*. 1991;44(8):763-770.
- Hasselblad V, Hedges LV. Meta-analysis of screening and diagnostic tests. *Psychol Bull*. 1995;117(1):167-178.
- Devillé WL, Buntinx F, de Vet R, Ligmer J, Montori V. Guidelines for conducting systematic reviews of studies evaluating the accuracy of diagnostic tests. In: Knotterus JA, ed. *The Evidence Base of Clinical Diagnosis*. London, England: BMJ Books; 2002.
- Bagai A, Thavendiranathan P, Detsky AS, Simel DL, Rennie D, eds. Does this patient have hearing impairment? *JAMA*. 2006;295(4):416-428.
- Holleman DR, Simel DL. Quantitative assessments from the clinical examination: how should clinicians integrate the numerous results? *J Gen Intern Med*. 1997;12(3):165-171.
- Guyatt G, Cairns J, Churchill D, et al; Evidence-Based Medicine Working Group. Evidence-based medicine: a new approach to teaching the practice of medicine. *JAMA*. 1992;268(17):2420-2425.
- Holleman DR, Simel DL. Does the clinical examination predict airflow limitation? *JAMA*. 1995;273(4):313-319.
- Bossuyt PMM, Reitsma JB, Bruns DE, et al; for the STARD Group. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *Clin Chem*. 2003;49:1-6.
- Bossuyt PMM, Reitsma JB, Bruns DE, et al. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. *Clin Chem*. 2003;49(1):7-18.
- Whiting P, Rutjes AWS, Reitsma JB, et al. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Med Res Tech*. 2003;3:25.
- Simel DL, Rennie D, Bossuyt PM. The STARD statement for reporting diagnostic accuracy studies: application to the history and physical examination. *J Gen Intern Med*. 2008;23(6):768-774.
- Holleman DR Jr, Simel DL. Does the clinical examination predict airflow limitation? *JAMA*. 1995;273(4):313-319.

*Verification bias can work in the opposite direction, although that is not usually the case.

CHAPTER 2

Does This Patient Have Abdominal Aortic Aneurysm?

Frank A. Lederle, MD

David L. Simel, MD, MHS

CLINICAL SCENARIOS

CASE 1 A 60-year-old man requests a physical examination because a friend recently died suddenly from a ruptured abdominal aortic aneurysm (AAA). Your examination reveals nothing abnormal. After reassuring the patient, you are left wondering whether you might have missed an AAA large enough to warrant surgical repair.

CASE 2 A thin 80-year-old woman observes that she can feel her abdomen pulsating against her belt. While examining her abdomen, you find an easily palpable, strongly pulsating aorta that you measure to be about 2 cm wide. You wonder whether you should order an ultrasonographic examination.

CASE 3 You are asked to see a 75-year-old man with 12 hours of right flank and abdominal pain, constipation, urinary frequency, urgency, dysuria, and leukocytosis and who is about to be sent home on treatment for pyelonephritis. Deep palpation of the abdomen is difficult, but you faintly discern a large pulsatile mass. You order computed tomography, which confirms an AAA with bleeding into the retroperitoneum, and the patient is taken to the operating room.

WHY IS PHYSICAL DIAGNOSIS OF AAA IMPORTANT?

Abdominal aortic aneurysms cause more than 10 000 deaths each year in the United States,¹ and many of these deaths should be preventable through timely diagnosis and treatment. AAAs usually remain asymptomatic while slowly enlarging during a period of years or even decades. About a third will eventually rupture, an event associated with a mortality rate of 80%.² Important risk factors for AAA include age, male sex, and smoking.³

Abdominal palpation was the original method of AAA detection. When ultrasonography and computed tomography became available, it was clear that they were more accurate than palpation, and these became the procedures of choice for confirming the diagnosis of AAA and for measurement of AAA diameter. A variety of studies have shown the sensitivity and specificity of ultrasonography and computed tomography to be close to 100%.⁴⁻⁸ Since then, the importance of abdominal palpation has been limited to identifying patients who should have confirmatory imaging studies. In one recent report, 31% of all AAAs diagnosed at a university hospital were originally detected by routine physical examination.⁹

The first scenario addresses the issues of screening (or case finding) to detect AAA and the subsequent management of asymptomatic AAA, 2 subjects of considerable debate in recent literature. Although most of the discussion of screen-

ing has focused on the use of ultrasonography, the only study to consider both methods found screening with abdominal palpation to be more cost-effective.¹⁰ In a review of the periodic physical examination, abdominal palpation for AAA was one of the few maneuvers recommended for older men.¹¹ The Canadian Task Force on the Periodic Health Examination observed that abdominal palpation of men older than 60 years was prudent,¹² but both the Canadian and the US Preventive Services Task Forces gave each AAA screening method a C rating (poor evidence to include or exclude from the periodic health examination), and some authors have judged the accuracy of abdominal palpation for AAA to be insufficient for screening.¹³

Management is based on observations that the risk of AAA rupture (and hence the need for elective repair) increases with the diameter of the aneurysm. The diameter of asymptomatic AAA above which repair should be offered to good surgical candidates is the topic of ongoing clinical trials,¹⁴ and current recommendations range from 4.0 to 6.0 cm, with 5.0 cm as the cutoff point most commonly used.¹⁵ Patients with AAAs that do not yet warrant repair are followed up with ultrasonography once or twice a year to detect enlargement that might warrant repair.

The second scenario represents what has been termed the *students' aneurysm*.¹⁶ Realization that these symptoms and physical findings are normal allows the physician to provide immediate reassurance to the patient and makes further testing unnecessary.

In the third scenario, abdominal palpation may have been lifesaving. Physical examination should not be relied on to rule out the diagnosis of ruptured AAA, and any patient in whom the diagnosis is considered should undergo ultrasonography or computed tomography. However, there are patients whose clinical likelihood of having a ruptured AAA lies below the physician's threshold for obtaining an imaging study and for whom physical examination may therefore be decisive. Many physicians are unfamiliar with the varied presentations of ruptured AAAs, so palpation of a widened aorta may be the first suggestion of the diagnosis.¹⁷

The importance of the physical examination in these settings depends largely on its accuracy. In this article, the accuracy of physical diagnosis of an AAA is assessed by review and analysis of the available literature. In 1905, Osler¹⁸ observed that "no pulsation, however forcible, no thrill, however intense, no bruit, however loud—singly or together—justify [sic] the diagnosis of an aneurysm of the abdominal aorta, *only the presence of a palpable expansile tumour*." Accordingly, most of the literature on physical examination to detect AAA has dealt with abdominal palpation to measure the width of the pulsatile mass representing the aneurysmal aorta, but several other physical signs have been considered. In one study, abdominal and femoral bruits and absent femoral pulses had no predictive value.⁸ Another study found that location of the pulsation more than 3.0 cm caudad of the umbilicus was not predictive of AAA.¹⁹ In 1975, Guarino²⁰ stated that the pulsatile mass of AAA could be distinguished by its being moveable laterally but not cephalad or caudad. This observation was not studied, however, and in the current era of readily avail-

able ultrasonography, there may be little value in further increasing the specificity of physical examination once a widened aorta is felt. We are aware of no other putative signs of AAA for which published information is available, so the remainder of this article will be limited to the consideration of abdominal palpation in detecting a widened aorta. Attempts to measure precisely the AAA diameter by abdominal palpation (as opposed to simply differentiating abnormal from normal) have also been studied^{4,5,21-23} but are of limited importance now that AAA measurements are routinely obtained more accurately from follow-up imaging studies and so will not be considered further.

METHODS

We searched MEDLINE for articles from 1966 to August 1998, using a search strategy previously developed for The Rational Clinical Examination series that combined 10 exploded MeSH headings ("physical examination," "medical history taking," "professional competence," "sensitivity and specificity," "reproducibility of results," "observer variation," "diagnostic tests, routine," "decision support techniques," "Bayes theorem," "mass screening") and 2 text word categories ("physical exam\$" and "sensitivity and specificity"), and then we took the intersection of this set with aortic aneurysm (exploded). The resulting set, plus articles in our files, references cited by these articles, and references in textbooks, was reviewed for information pertinent to the clinical examination of AAA. Unpublished information was obtained from the authors of some studies.

Series with fewer than 10 patients and those published before 1966 were not considered. No other exclusions (eg, language, publication type) were applied. We assigned each study to a level of evidence according to a system previously developed for this series.²⁴ Level 1 studies are independent, blind comparisons of sign or symptom results with a criterion standard among a large number (sufficient to have narrow confidence limits on the resulting sensitivity, specificity, or likelihood ratio) of consecutive patients suspected of having the target condition. Level 2 studies are independent, blind comparisons of sign or symptom results with a criterion standard among a small number of consecutive patients suspected of having the target condition. Level 3 studies are independent, blind comparisons of signs and symptoms with a criterion standard among nonconsecutive patients suspected of having the target condition. Level 4 studies are nonindependent comparisons of signs and symptoms with a criterion standard among convenience samples of patients who obviously have the target condition plus, perhaps, healthy individuals. Level 5 studies are nonindependent comparisons of signs and symptoms with a standard of uncertain validity (which may even incorporate the sign or symptom result in its definition) among convenience samples of patients and, perhaps, healthy patients.

Abdominal aortic aneurysm, to provide consistency in data extraction, was defined as an abdominal aortic diameter of 3.0 cm or greater. There is no widely accepted method of defining

the cutoff point between a normal aorta and an AAA. Imaging studies done in clinical practice are often interpreted according to arterial shape (eg, distal widening), but epidemiologic studies have generally used the simpler measure of unadjusted infrarenal aortic diameter, which has been shown to be associated with rupture risk.²⁵ An infrarenal aortic diameter of 3.0 cm is a commonly used but somewhat controversial cutoff point in published articles, whereas a diameter of 4.0 cm or larger is clearly diagnostic of an AAA. Adjustment of the cutoff point for such factors as age, sex, and body size has been suggested but appears to have little practical value.²⁶

An *a priori* decision was made to consider intermediate findings on palpation as negative when the uncertainty was due to the aorta's being impalpable²⁷⁻³⁰ and positive when the findings were considered suggestive of an AAA (as opposed to definite).^{8,31}

Sensitivity was calculated as the proportion of affected patients with positive findings, specificity as the proportion of unaffected patients with negative findings, and a positive predictive value as the proportion of patients with positive findings who were affected. Likelihood ratios were also calculated; the positive likelihood ratio (LR_+) is defined as sensitivity/(1 – specificity) and expresses the increase in the odds of having the disease when the finding is positive (LR_+ values are ≥ 1), and the LR_- is defined as (1 – sensitivity)/specificity and expresses the decrease in the odds of having the disease when the finding is negative (LR_- values are 0-1). Values for true positives, false positives, true negatives, and false negatives were increased by 0.5 when likelihood ratios were computed to avoid division by 0.³² CIs for likelihood ratios from individual studies were computed using the method of Simel et al.³³

The studies of AAA screening were judged to be of sufficient quality and similarity of design to assess for statistical similarity. The χ^2 tests for heterogeneity of the sensitivity data were not significant (all $P > .10$), supporting the decision to pool these data.³⁴ However, assessments of heterogeneity of the effectiveness scores (a measure of the effect size of a diagnostic test result) were of borderline significance

(pooled effectiveness, 1.7; $P = .04$ using a cutoff of 3.0 cm; pooled effectiveness, 2.1; $P = .06$ using a cutoff of 4.0 cm).³² Therefore, a random-effects measure was used as a conservative method for pooling the results of these studies, and CIs for the pooled likelihood ratios were calculated by using the method of Eddy and Hasselblad.³⁴

RESULTS

Abdominal Palpation for Ruptured AAA

Several studies have reported the sensitivity of abdominal palpation in patients with ruptured AAA (Table 2-1).^{17,35-42} In these studies, it is not clear how often the physical findings suggested the diagnosis of AAA as opposed to being elicited after the diagnosis was made by other methods. The sensitivities tended to be higher when patient selection was limited to those diagnosed antemortem (including operative series). Three series included masses that were described as not pulsatile, and sensitivities with these masses included are reported separately in Table 2-1. Compared with asymptomatic AAAs, ruptured AAAs tend to be larger, which would be expected to increase sensitivity,⁴³ but rupture may also be associated with guarding, intestinal distention caused by compromised circulation, and loss of integrity of the AAA, which could have the opposite effect.

Abdominal Palpation for Asymptomatic AAA

Some studies have reported the sensitivity of abdominal palpation in patients with known asymptomatic AAAs (range of sensitivities, 65%-100%).^{4-7,22,23,36,39,44-49} Most of these studies involved patients undergoing preoperative evaluation for elective repair of large AAAs, and many patients were originally identified by physical examination before referral to the study group. The lack of blinding and the preponderance of large AAAs likely resulted in higher sensitivities than would be achieved in most clinical settings.

Table 2-1 Sensitivity of Abdominal Palpation in Series of Patients With Ruptured Abdominal Aortic Aneurysm^a

Source, y	No. of AAAs	Sensitivity of Palpation (%) ^b	Patient Selection
Pryor, ³⁵ 1972	44	45 (82)	All
Williams et al, ³⁶ 1972	79	97	Operated on
Ottinger, ³⁷ 1975	40	75 (100)	Diagnosed antemortem
McGregor, ³⁸ 1976	41	44 (51)	Unoperated or at autopsy
Gordon-Smith et al, ³⁹ 1978	83	90	Operated on
Gaylis and Kessler, ⁴⁰ 1980	105	87	Diagnosed antemortem
Donaldson et al, ⁴¹ 1985	81	91	Not stated
Walsh et al, ⁴² 1992	55	64	All
Lederle et al, ¹⁷ 1994	23	52	Presented to internist

Abbreviation: AAA, abdominal aortic aneurysm.

^aAll studies provide level 4 evidence (see "Methods" section).

^bNumbers in parentheses represent the sensitivity if nonpulsatile masses are included.

Other studies have reported the positive predictive value of clinical suspicion for AAAs in a series of patients referred for imaging studies (range of positive predictive values, 15%-91%).^{6,13,21,31,48-53} The wide range of values may reflect possible inclusion in some studies of patients with previous diagnostic imaging studies before their referral to the study group (falsely increasing positive predictive value) and of patients referred for ruling out AAA according to indications other than palpation of a widened aorta (potentially falsely increasing or decreasing positive predictive value). Two studies provide results by age and sex, indicating that the highest positive predictive values are obtained in men older than 60 years, with low values (<15%) obtained in women and younger men.^{13,53}

The best evidence available for assessing the performance of abdominal palpation in detecting AAAs comes from series of patients not previously suspected of having AAAs who were screened by abdominal palpation and ultrasonography (Table 2-2).^{8,19,27-30,54-62} In all 15 of these studies, screening was limited to patients at increased risk for AAAs, usually older men with hypertension or vascular disease. Blinding of the examiner was ensured when physical examination preceded ultrasonography; this was stated to have occurred in 8 of these 15 studies^{8,19,27-30,55,59} and was implied to have occurred in the others. No study stated whether the ultrasonographer was blinded to the physical examination findings.

The low level of disease prevalence in these screening studies and the resulting low expectation of disease by the examiner have the advantage of reflecting most clinical settings. A disadvantage is that the small number of AAAs, particularly larger AAAs, limits the precision of the estimates from individual studies. We addressed this problem by pooling data from all studies.

In the pooled analysis, the sensitivity of abdominal palpation increased significantly with the AAA's diameter ($P < .001$, χ^2 for trend), illustrating the previously described effect of disease severity on sensitivity.⁴³ As seen in Table 2-2, the pooled sensitivity values range from 29% for AAAs of 3.0 to 3.9 cm to 50% for AAAs of 4.0 to 4.9 cm and to 76% for AAAs of 5.0 cm or greater. As would be expected, these sensitivities are lower than those observed in the series of previously known (and presumably larger) AAAs mentioned above.

The high LR+ indicates that the finding of a widened aorta greatly increases the odds that an AAA is present, whereas the LR- indicates that the absence of this finding is only moderately effective in ruling out an AAA. Not surprisingly, the likelihood ratios also indicate that abdominal palpation is a more effective diagnostic test for larger AAAs (LR+, 16; LR-, 0.51, using a cutoff point for AAA of ≥ 4.0 cm vs LR+, 12; LR-, 0.72 using a cutoff point for AAA of ≥ 3.0 cm).

Factors That Affect Abdominal Palpation for AAA

The sensitivities shown in Table 2-2 apply only to abdominal palpation directed at AAA detection and not to routine abdominal palpation (abdominal palpation as it is routinely done in practice, not necessarily specifically directed at mea-

suring aortic width). Several studies have compared routine physical examination with abdominal palpation directed at AAA detection. In one of the screening studies listed in Table 2-2, all 5 patients with AAAs considered definite at the study's physical examination and subsequently confirmed by ultrasonography had been missed on routine physical examination of the abdomen in the previous 6 months.⁸ Another study found that 95 of 188 patients with AAAs considered palpable on physical examination before surgery had been missed on at least 1 physical examination in the 12 months preceding the initial diagnosis.⁴⁷ In a third study, 19 of 37 patients with previously undiagnosed but easily palpable ruptured AAAs (all 6-10 cm in diameter) had undergone physical examination in the previous 24 months, but the diagnosis had been missed.⁶³ Abdominal aortic aneurysms enlarge at a mean rate of 0.2 to 0.5 cm/y,^{25,64} so the interval was unlikely to have been an important confounder in these studies.

Obesity also appears to limit the effectiveness of abdominal palpation. In one study, patients with AAAs missed at palpation had significantly greater mean abdominal girth than patients with AAAs detected at palpation (111 vs 96 cm; $P < .01$), and when abdominal girth was less than 100 cm, 6 of 6 AAAs were detected at palpation compared with 3 of 12 AAAs that were detected when abdominal girth was 100 cm or more ($P < .01$).⁸ Another study observed that 23% of the patients "were too obese for us to feel the aortic pulse."³⁰ We are aware of no reports discussing whether AAA is ruled out more reliably when the aorta is palpable and considered to be normal than when the aorta cannot be adequately palpated.

How to Perform Abdominal Palpation to Detect AAA

Abdominal palpation should be conducted while the patient is supine, with his or her knees raised while the abdomen relaxes. The examiner first feels deeply for the aortic pulsation, usually found a few centimeters cephalad of the umbilicus (the umbilicus marks the level of the aortic bifurcation) and slightly to the left of midline. The examiner then positions both hands on the abdomen with palms down, placing an index finger on either side of the pulsating area to confirm that it is the aorta (each systole should move the 2 fingers apart) and to measure the aortic width. A generous amount of abdominal skin should be included between the 2 index fingers, and it is often easier, initially, to probe for one side of the aorta at a time.

It is the width, and not the intensity, of the aortic pulsation that determines the diagnosis of an AAA; a normal aorta is often readily palpable in thin patients or those with loose abdominal muscles. The aorta is normally less than 2.5 cm (1 in) in diameter, and aortas larger than this (after allowing for skin thickness) warrant further investigation, usually with ultrasonography. Physical examination to detect AAAs is rarely warranted in persons younger than 50 years because of the low frequency of disease in this group.

There are no known risks associated with palpation of the abdominal aorta. We found no reports of AAA rupture attributed to physical examination, and a textbook author observed that he was "unaware of rupture during examina-

Table 2-2 Abdominal Palpation in Populations Screened for Asymptomatic Abdominal Aortic Aneurysm^a

Source, y	Range of Patient Age, y	Women, %	No. Screened	No. of AAAs Diagnosed by Ultrasonography and Sensitivity of Abdominal Palpation								Positive Predictive Value of Palpation, %	Likelihood Ratios				
				≥3.0 cm (All)		3.0-3.9 cm		4.0-4.9 cm		≥5.0 cm			Cutoff Point: AAA ≥3.0 cm		Cutoff Point: AAA ≥4.0 cm		
				AAA	Sensitivity, %	AAA	Sensitivity, %	AAA	Sensitivity, %	AAA	Sensitivity, %		LR+ (95% CI)	LR- (95% CI)	LR+ (95% CI)	LR- (95% CI)	
Cabellon et al, ²⁷ 1983	43-79	33	73	9 ^b	22	NA	NA	NA	...	NA	...	67	11 (1.6-73)	0.77 (0.54-1.1)	
Ohman et al, ⁵⁴ 1985	50-88	0	50	3	0	2	0	1	0	0	12 (0.3-528)	0.88 (0.61-1.3)	25 (0.6-968)	0.76 (0.34-1.7)	
Twomey et al, ⁵⁵ 1986	>50	0	200	14	64	7	43	3	100	4	75	64	21 (8.7-53)	0.38 (0.19-0.74)	18 (8.9-39)	0.20 (0.05-0.83)	
Allen et al, ⁵⁶ 1987	>65	43	168	3	0	2	0	0	...	1	0	0	1.6 (0.1-23)	0.95 (0.65-1.4)	3.3 (0.3-39)	0.81 (0.36-1.8)	
Allardice et al, ⁵⁷ 1988	39-90	25	100	15	33	10	0	3	100	2	100	100	59 (3.4-1018)	0.66 (0.46-0.94)	176 (11-2823)	0.08 (0.01-1.2)	
Lederle et al, ⁸ 1988	60-75	0	201	20	45	10	40	5	20	5	80	35	4.7 (2.5-9.0)	0.61 (0.41-0.90)	4.5 (2.2-9.1)	0.56 (0.31-1.0)	
Collin et al, ¹⁹ 1988	65-74	0	426	23 ^d	35	NA	NA	NA	...	NA	...	36	9.9 (4.7-21)	0.67 (0.50-0.90)	
Shapira et al, ⁵⁸ 1990	31-83	36	101	4	0	2	0	0	...	2	0	...	20 (0.4-890)	0.90 (0.68-1.2)	33 (0.8-1415)	0.84 (0.50-1.4)	
Andersson et al, ⁵⁹ 1991	38-86	42	288	14	29	NA	NA	NA	...	NA	...	31	8.7 (3.2-23)	0.73 (0.52-1.0)	
Spiridonov and Omirov, ⁶⁰ 1992	17-67	13	163	10	70	3	0	4	100	3	100	26	5.1 (2.9-9.1)	0.37 (0.15-0.87)	7.2 (4.6-11)	0.07 (0-1.0)	
MacSweeney et al, ²⁸ 1993	NA	36	200	55	24	33	0	16	44	6	100	72	6.4 (2.5-16)	0.79 (0.68-0.92)	19 (7.8-47)	0.43 (0.26-0.69)	
Karanjia et al, ⁶¹ 1994	55-82	41	89	9	100	2	100	5	100	2	100	82	31 (9.0-105)	0.05 (0-0.77)	17 (6.9-43)	0.07 (0-0.97)	
Molnar et al, ⁶² 1995	65-83	53	411	7	43	2	50	3	33	2	50	33	27 (9.1-81)	0.57 (0.31-1.0)	23 (6.9-74)	0.59 (0.30-1.2)	
al Zahrani et al, ²⁹ 1996	60-80	29	392	7	57	1	0	4	50	2	100	57	62 (18-208)	0.44 (0.20-3.0)	71 (22-231)	0.36 (0.13-0.97)	
Arnell et al, ³⁰ 1996	55-81	0	96	1	100	1	100	0	...	0	...	14	11 (3.7-33)	0.27 (0.02-3.0)	6.5 (0.8-52)	0.54 (0.08-3.8)	
Pooled results	...	26	2955	194	39	75	29	44	50	29	76	43	12 (7.4-19)	0.72 (0.65-0.81)	16 (8.6-28)	0.51 (0.38-0.67)	

Abbreviations: AAA, abdominal aortic aneurysm; CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio; NA, data not available.

^aIncludes unpublished information received from authors. All studies used ultrasonography and provide level 2 evidence. The pooled results for numbers are sums and for functions are from a random-effects measure and provide level 1 evidence (see "Methods" section). Abdominal aortic aneurysm is defined as at least 3.0 cm by ultrasonography.

^bNo information was given on AAA diameter.

^cEllipses indicate values cannot be calculated.

^dAbdominal aneurysms less than 3 cm are included.

tion of any of several thousand AAAs seen over four decades.”⁶⁵

We are aware of no educational studies examining methods of learning AAA palpation. In our experience, however, accurate palpation is readily learned through practice and feedback. We have found that physicians can become proficient after comparing their findings with ultrasonographic measurements in a few patients with AAAs and a few controls.

Bottom Line

The only physical examination maneuver of demonstrated value for the diagnosis of an AAA is abdominal palpation to detect a widened aorta. Palpation of AAA appears to be safe and has not been reported to precipitate rupture.

Positive findings on abdominal palpation greatly increase the likelihood that an AAA, particularly a large AAA, is present. Even so, the positive predictive value of 43% (Table 2-2) indicates that less than half of all high-risk patients (and fewer low-risk patients, such as most women and young men) suspected of having an enlarged aorta on abdominal palpation will be found to have an AAA. However, this may not be of great concern because ultrasonography provides a safe and relatively inexpensive confirmatory test.

Abdominal palpation will detect most AAAs large enough to warrant surgery, but it cannot be relied on to rule out the diagnosis. The sensitivity of palpation appears to be reduced by abdominal obesity and by routine abdominal examination not specifically directed at measuring aortic width. When a ruptured AAA is suspected, imaging studies such as ultrasonography or computed tomography should be performed regardless of physical findings.

Author Affiliations at the Time of the Original Publication

Departments of Medicine, Minneapolis Veterans Affairs Medical Center, University of Minnesota, Minneapolis (Dr Lederle), and Durham Veterans Affairs Medical Center, Duke University, Durham, North Carolina (Dr Simel)

Acknowledgments

The authors thank Andreas Laupacis, MD, and Kavita Nanda, MD, for their helpful reviews of the article.

REFERENCES

- Gillum RF. Epidemiology of aortic aneurysm in the United States. *J Clin Epidemiol*. 1995;48(11):1289-1298.
- Ingoldby CJH, Wujanto R, Mitchell JE. Impact of vascular surgery on community mortality from ruptured aortic aneurysms. *Br J Surg*. 1986; 73(7):551-553.
- Lederle FA, Johnson GR, Wilson SE, et al; Aneurysm Detection and Management (ADAM) Veterans Affairs Cooperative Study Group. Prevalence and associations of abdominal aortic aneurysm detected through screening. *Ann Intern Med*. 1997;126(6):441-449.
- Hertzler NR, Beven EG. Ultrasound aortic measurement and elective aneurysmectomy. *JAMA*. 1978;240(18):1966-1968.
- Graeve AH, Carpenter CM, Wicks JD, Edwards WS. Discordance in the sizing of abdominal aortic aneurysm and its significance. *Am J Surg*. 1982;144(6):627-634.
- Nusbaum JW, Freimanis AK, Thomford NR. Echography in the diagnosis of abdominal aortic aneurysm. *Arch Surg*. 1971;102(4): 385-388.
- Lee KR, Walls WJ, Martin NL, Templeton AW. A practical approach to the diagnosis of abdominal aortic aneurysms. *Surgery*. 1975;78:195-201.
- Lederle FA, Walker JM, Reinke DB. Selective screening for abdominal aortic aneurysms with physical examination and ultrasound. *Arch Intern Med*. 1988;148(8):1753-1756.
- Kiev J, Eckhardt A, Kerstein MD. Reliability and accuracy of physical examination in detection of abdominal aortic aneurysms. *Vasc Surg*. 1997;31(2):143-146.
- Frame PS, Fryback DG, Patterson C. Screening for abdominal aortic aneurysm in men ages 60 to 80 years. *Ann Intern Med*. 1993;119(5):411-416.
- Oboler SK, LaForce FM. The periodic physical examination in asymptomatic adults. *Ann Intern Med*. 1989;110(3):214-226.
- Canadian Task Force on the Periodic Health Examination. Periodic health examination, 1991 update, 5. *CMAJ*. 1991;145(7):783-789.
- Beede SD, Ballard DJ, James EM, Ilstrup DM, Hallet JW. Positive predictive value of clinical suspicion of abdominal aortic aneurysm. *Arch Intern Med*. 1990;150(3):549-551.
- Lederle FA, Wilson SE, Johnson GR, et al; ADAM VA Cooperative Study Group. Design of the Abdominal Aortic Aneurysm Detection and Management (ADAM) Study. *J Vasc Surg*. 1994;20(2):296-303.
- Ballard DJ, Etchason JA, Hilborne LH, et al. *Abdominal Aortic Aneurysm Surgery: A Literature Review and Ratings of Appropriateness and Necessity*. Santa Monica, CA: RAND; 1992.
- Fowler NO. Diseases of the aorta. In: Wyngaarden JB, Smith LH, eds. *Cecil Textbook of Medicine*. 17th ed. Philadelphia, PA: WB Saunders Co; 1985:345-353.
- Lederle FA, Parenti CM, Chute EP. Ruptured abdominal aortic aneurysm: the internist as diagnostician. *Am J Med*. 1994;96(2):163-167.
- Osler W. Aneurysm of the abdominal aorta. *Lancet*. 1905;2:1089-1096.
- Collin J, Araujo L, Walton J, Lindsell D. Oxford screening programme for abdominal aortic aneurysm in men aged 65 to 74 years. *Lancet*. 1988;332(8611):613-615.
- Guarino JR. Abdominal aortic aneurysm. *J Kans Med Soc*. 1975;76(5): 108, 15A.
- McGregor JC, Pollock JG, Anton HC. The value of ultrasonography in the diagnosis of abdominal aortic aneurysm. *Scott Med J*. 1975;20 (3):133-137.
- Brewster DC, Darling RC, Raines JK, et al. Assessment of abdominal aortic aneurysm size. *Circulation*. 1977;56(suppl 2):164-169.
- Buxton B, Buttery B, Buckley J. The measurement of abdominal aortic aneurysms. *Aust NZ J Surg*. 1978;48(4):387-389.
- Hollerman DR, Simel DL. Does the clinical examination predict airflow limitation? *JAMA*. 1995;273(4):313-319.
- Nevitt MP, Ballard DJ, Hallett JW. Prognosis of abdominal aortic aneurysms: a population-based study. *N Engl J Med*. 1989;321(15):1009-1014.
- Lederle FA, Johnson GR, Wilson SE, et al; ADAM VA Cooperative Study Investigators. Relationship of age, gender, race, and body size to infrarenal aortic diameter. *J Vasc Surg*. 1997;26(4):595-601.
- Cabellon S, Moncrief CL, Pierre DR, Cavanaugh DG. Incidence of abdominal aortic aneurysms in patients with atherosomatous arterial disease. *Am J Surg*. 1983;146(5):575-576.
- MacSweeney ST, O'Meara M, Alexander C, O'Malley MK, Powell JT, Greenhalgh RM. High prevalence of unsuspected abdominal aortic aneurysm in patients with confirmed symptomatic peripheral or cerebral arterial disease. *Br J Surg*. 1993;80(5):582-584.
- Zahrani HA, Rawas M, Maimani A, Gasab M, Aba al Khail BA. Screening for abdominal aortic aneurysm in the Jeddah area, western Saudi Arabia. *Cardiovasc Surg*. 1996;4(1):87-92.
- Arnell TD, de Virgilio C, Donayre C, Grant E, Baker JD, White R. Abdominal aortic aneurysm screening in elderly males with atherosclerosis: the value of physical exam. *Am Surg*. 1996;62(10):861-864.
- Robicsek F, Daugherty HK, Mullen DC, Tam W, Scott WP. The value of angiography in the diagnosis of unruptured aneurysms of the abdominal aorta. *Ann Thorac Surg*. 1971;11(6):538-550.
- Hasselblad V, Hedges LV. Meta-analysis of screening and diagnostic tests. *Psychol Bull*. 1995;117(1):167-178.
- Simel DL, Samsa GP, Matchar DB. Likelihood ratios with confidence: sample size estimation for diagnostic test studies. *J Clin Epidemiol*. 1991;44(8):763-770.
- Eddy DM, Hasselblad V. *Fast*Pro v1.8: Software for Meta-analysis by the Confidence Profile Method*. San Diego, CA: Academic Press; 1992:91-92.
- Pryor JP. Diagnosis of ruptured aneurysm of abdominal aorta. *BMJ*. 1972;3(5829):735-736.

36. Williams RD, Fisher FW, Dickey JW Jr. Problems in the diagnosis and treatment of abdominal aortic aneurysms. *Am J Surg.* 1972;123(6):698-701.
37. Ottinger LW. Ruptured arteriosclerotic aneurysms of the abdominal aorta: reducing mortality. *JAMA.* 1975;233(2):147-150.
38. McGregor JC. Unoperated ruptured abdominal aortic aneurysms. *Br J Surg.* 1976;63(2):113-116.
39. Gordon-Smith IC, Taylor EW, Nicolaides AN, et al. Management of abdominal aortic aneurysm. *Br J Surg.* 1978;65(12):834-838.
40. Gaylis H, Kessler E. Ruptured aortic aneurysms. *Surgery.* 1980;87(3):300-304.
41. Donaldson MC, Rosenberg JM, Bucknam CA. Diagnosis of ruptured abdominal aortic aneurysm. *Conn Med.* 1985;49(1):3-6.
42. Walsh JA, Dohnalek JA, Doley AJ, Wiadrowski TP. Ruptured abdominal aortic aneurysms. *Med J Aust.* 1992;156(2):138.
43. Lofgren RP. The dynamic nature of sensitivity and specificity. *J Gen Intern Med.* 1987;2(6):452-453.
44. Friedman SA, Hufnagel CA, Conrad PW, Simmons EM, Weintraub A. Abdominal aortic aneurysms: clinical status and results of surgery in 100 consecutive cases. *JAMA.* 1967;200(13):1147-1151.
45. Bergan JJ, Yao JST, Henkin RE, Quinn JL. Radionuclide aortography in detection of arterial aneurysms. *Arch Surg.* 1974;109(1):80-83.
46. Volpetti G, Barker CF, Berkowitz H, Roberts B. A twenty-two year review of elective resection of abdominal aortic aneurysms. *Surg Gynecol Obstet.* 1976;142(3):321-324.
47. Chervu A, Clagett GP, Valentine RJ, Myers SI, Rossi PJ. Role of physical examination in detection of abdominal aortic aneurysms. *Surgery.* 1995;117(4):454-457.
48. Robicsek F. The diagnosis of abdominal aneurysms. *Surgery.* 1981;89(2):275-276.
49. Roberts A, Johnson N, Royle J, Buttery B, Buxton B. The diagnosis of abdominal aortic aneurysms. *Aust N Z J Surg.* 1974;44(4):360-362.
50. Brewster DC, Retana A, Waltman AC, Darling RC. Angiography in the management of aneurysms of the abdominal aorta: its value and safety. *N Engl J Med.* 1975;292(16):822-825.
51. Lee TG, Henderson SC. Ultrasonic aortography: unexpected findings. *Am J Roentgenol.* 1977;128(2):273-276.
52. Karp W, Eklof B. Ultrasonography and angiography in the diagnosis of abdominal aortic aneurysm. *Acta Radiol.* 1978;19(6):955-960.
53. Kahn CE Jr, Quiroz FA. Positive predictive value of clinical suspicion for abdominal aortic aneurysm. *J Gen Intern Med.* 1996;11(12):756-758.
54. Ohman EM, Fitzsimons P, Butler F, Bouchier-Hayes D. The value of ultrasonography in the screening for asymptomatic abdominal aortic aneurysm. *Ir Med J.* 1985;78(5):127-129.
55. Twomey A, Twomey E, Wilkins RA, Lewis JD. Unrecognised aneurysmal disease in male hypertensive patients. *Int Angiol.* 1986;5(4):269-273.
56. Allen PI, Gourevitch D, McKinley J, Tudway D, Goldman M. Population screening for aortic aneurysms [letter]. *Lancet.* 1987;2(8561):736.
57. Allardice JT, Allwright GJ, Wafula JM, Wyatt AP. High prevalence of abdominal aortic aneurysm in men with peripheral vascular disease: screening by ultrasonography. *Br J Surg.* 1988;75(3):240-242.
58. Shapira OM, Pasik S, Wassermann JP, Barzilai N, Mashiah A. Ultrasound screening for abdominal aortic aneurysms in patients with atherosclerotic peripheral vascular disease. *J Cardiovasc Surg (Torino).* 1990;31(2):170-172.
59. Andersson AP, Ellitsgaard N, Jorgensen B, et al. Screening for abdominal aortic aneurysm in 295 outpatients with intermittent claudication. *Vasc Surg.* 1991;25(7):516-520.
60. Spiridonov AA, Omirov ShR. Selective screening for abdominal aortic aneurysms through clinical examination and ultrasonic scanning [in Russian]. *Grud Serdechnosudistaia Khir.* 1992;9-10:33-36.
61. Karanjia PN, Madden KP, Lobner S. Coexistence of abdominal aortic aneurysm in patients with carotid stenosis. *Stroke.* 1994;25(3):627-630.
62. Molnar LJ, Langer B, Serro-Azul J, Wanjgarten M, Cerri GG, Lucarelli CL. Prevalence of abdominal aneurysm in the elderly [in Portuguese]. *Rev Assoc Med Bras.* 1995;41(1):43-46.
63. Craig SR, Wilson RG, Walker AJ, Murie JA. Abdominal aortic aneurysm: still missing the message. *Br J Surg.* 1993;80(4):450-452.
64. Cronenwett JL, Sargent SK, Wall WH, et al. Variables that affect the expansion rate and outcome of small abdominal aortic aneurysms. *J Vasc Surg.* 1990;11(2):260-269.
65. Joyce JW. Examination of the patient with vascular disease. In: Loscalzo J, Creager MA, Dzau VJ, eds. *Vascular Medicine.* Boston, MA: Little Brown & Co; 1992:401-418.

This page intentionally left blank

Prepared by Frank Lederle, MD

Reviewed by Ed Etchells, MD

CLINICAL SCENARIO

You are performing a physical examination on an obese 65-year-old man. You have been thorough with abdominal palpation and allowed the abdominal muscles to relax enough so that you can feel the aortic pulsation. You estimate it to be 2 cm wide, which is normal. Because you have heard that abdominal palpation is less accurate in obese patients, you wonder whether the examination findings exclude abdominal aortic aneurysm (AAA).

UPDATED SUMMARY ON ABDOMINAL AORTIC ANEURYSM

Original Review

Lederle FA, Simel DL. Does this patient have abdominal aortic aneurysm? *JAMA*. 1999;281(1):77-82.

UPDATED LITERATURE SEARCH

We reviewed all citations listed under “exp aortic aneurysm” in MEDLINE, from 1998 to July 2004. The search yielded 7590 titles. We also searched personal files maintained on the topic since the original publication. We reviewed titles and abstracts to identify new studies that met the original inclusion and exclusion criteria, focusing on large studies that included information on the sensitivity or specificity of the physical examination for abdominal aneurysms in the general population. The review identified only 1 article that met our inclusion criteria.

NEW FINDINGS

- The interobserver variability for detecting aneurysms is good.
- The sensitivity of the examination is better for smaller patients than for larger patients. However, the sensitivity in larger patients is still good when the aorta can be palpated.
- When the patient cannot “relax” the abdomen, clinicians should be aware that they are more likely to “miss” an aneurysm.

Details of the Update

Abdominal palpation continues to be an important method for diagnosing AAA. In a recent study from a UK district general hospital, 48% of all AAAs were diagnosed by physical examination¹ compared with 31% in reference 9 of the original Rational Clinical Examination article.

A study published after the original review evaluated patient factors such as abdominal obesity, girth, and tightness and the effect of a palpable aorta, which might have an effect on the accuracy of the clinical evaluation. In addition, the investigators provided information on interobserver variability in abdominal palpation for AAA.² The only pragmatic way to conduct such an evaluation is through the evaluation of patients with and without an aneurysm. In this study of 200 subjects, 99 with and 101 without AAA, the interobserver pair agreement for AAA vs no AAA between the first and second examination was 77% ($\kappa = 0.53$). The sensitivity of the examination improves with increasing size of the aneurysm. For aneurysms 5 cm or larger, the sensitivity was 82%. Not surprisingly, the examiners also had better sensitivity in thinner subjects (abdominal girth less than 100 cm [40-in waistline]) than in more obese subjects (sensitivity, 91% vs 53% for girth of 100 cm or more). Even when girth was 100 cm or more, if the aorta was palpable, sensitivity was 82%. Physicians sometimes have trouble palpating the abdominal aorta when patients cannot “relax” their abdomen. This study confirmed that the examiners’ assessment that the abdomen was not tight improved their accuracy in detecting aneurysms (odds ratio, 2.7; 95% confidence interval, 1.2-6.1).

In another study, 125 subjects with AAA and 39 without AAA underwent abdominal palpation with a vascular surgeon, a nurse, and the patient.³ The vascular surgeon and nurse knew of the high prevalence of AAA in the sample, but they did not know an individual patient’s diagnosis. For vascular surgeons, sensitivity was 57% for AAAs less than 4.0 cm but more than 97% for AAAs larger than 4.0 cm. The accuracy of nurses and patients was similar to that of the surgeons, which is surprising because the patients used palpable pulsation as the only criterion for diagnosing AAA. The κ value for agreement between surgeons and nurses was high, at 0.92, and agreement of either with the patient was nearly as high. Factors independently associated with false negatives were smaller AAA diameter and higher body mass index. The extremely high sensitivities, presumably related to the exam-

Table 2-3 The More Certain the Examiner Feels About the Findings, the More Likely They Are Correct

Clinical Impression	LR+ (95% CI)
Examination "definite" for aneurysm	4.8 (2.7-8.8)
Examination "suggestive"	1.4 (0.92-2.1)
Examination "normal"	0.43 (0.35-0.54)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio.

iners' knowledge of the high prevalence of AAA, raise questions about the study's generalizability.

The largest sensitivity study to date was recently reported from Brazil.⁴ The first 3000 subjects to call in response to an advertising campaign were scheduled for screening. The study group consisted of 2756 subjects who responded to an advertising campaign, were older than 50 years, had no previous diagnosis of AAA, and had an adequate ultrasonographic examination. Each subject underwent abdominal palpation with a vascular surgeon and ultrasonography. It is unclear whether palpation was blinded to ultrasonographic findings. There were 64 AAAs 3.0 cm or larger identified by ultrasonography. Sensitivity and positive predictive value of a positive abdominal palpation result were 31% and 33%, respectively. This sensitivity was somewhat lower than in previous studies, possibly reflecting reduced examiner vigilance resulting from the size of the study.

Several other studies since the original review added useful information but did not meet our inclusion criteria. A pulsatile mass may be present after endovascular repair of AAA, potentially leading to diagnostic confusion.⁵ A cohort study from the Medical Research Council Thrombosis Prevention Trial examined the result of abdominal palpation of the aorta by general practitioners in 4171 men from 1992 to 1994.⁶ Abdominal aortic aneurysm was suspected in 60 men and confirmed in 25 (positive predictive value, 42%). By mid-1996, 6 men died of ruptured AAA who had not been suspected of having AAA on palpation, suggesting that sensitivity of palpation to detect clinically important AAA was less than 81%.

In an older study addressing predictive value, only 1 of 29 consecutive patients presenting to the Massachusetts General Hospital emergency department in the 1970s with tender pulsatile mass without hypovolemia actually had AAA.⁷

IMPROVEMENTS IN THE DATA PRESENTED IN THE ORIGINAL PUBLICATION

A new study allows us to assess the likelihood of an aneurysm according to clinicians' confidence in their examination findings and the accuracy of the examination related to various patient factors such as obesity (see Table 2-3).

Whereas the original publication observed that 5 cm was the threshold most commonly used for considering surgery, 2 large randomized trials show no benefit of repair for aneurysms with a diameter of less than 5.5 cm.^{8,9}

CHANGES IN THE REFERENCE STANDARD

There are no changes in the reference standard.

RESULTS OF LITERATURE REVIEW

Univariate Findings for AAA

The efficiency of the examination depends on the confidence in your findings.

EVIDENCE FROM GUIDELINES

Four trials of screening for abdominal aneurysms with ultrasonography have been conducted since the original US Preventive Services Task Force and Canadian Task Force recommendations. The US Preventive Services Task Force now recommends one-time screening for AAA by ultrasonography in men aged 65 to 75 years who have ever smoked.¹⁰

CLINICAL SCENARIO—RESOLUTION

Although it is true that abdominal palpation is less accurate in obese patients (roughly those with a waist circumference of more than 40 in), the fact that you could palpate the aorta improves the accuracy. The sensitivity for detecting an AAA 3.0 cm or larger is 82%, and your finding that the aorta was normal confers a negative likelihood ratio of 0.30. You are able to reassure the patient that, given your examination findings, the likelihood that he has an AAA is low.

ABDOMINAL AORTIC ANEURYSM—MAKE THE DIAGNOSIS

PRIOR PROBABILITY

Abdominal aortic aneurysms occur in 4% to 8% of older men. The prevalence in older women is less than 2%.

POPULATION FOR WHOM AAA SHOULD BE CONSIDERED

- Age older than 50 years
- History of ever smoking
- Male sex
- White race
- Family history of AAA

DETECTING AN ABDOMINAL AORTIC ANEURYSM

The size of an aneurysm affects the clinician's ability to detect it ([Table 2-4](#)).

Table 2-4 Likelihood Ratios Vary With the Size of the Aneurysm

Ability to Detect an Asymptomatic Aneurysm According to Size	LR+ (95% CI)	LR- (95% CI)
Aneurysm > 4.0 cm (n = 12 studies)	16 (8.6-29)	0.51 (0.38-0.67)
Aneurysm > 3.0 cm (n = 15 studies)	12 (7.4-20)	0.72 (0.65-0.81)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

Clinicians can detect asymptomatic AAAs. The ability to detect the aneurysm relates, in part, to patient characteristics. The examination should focus on the width of the palpated abdominal aorta. Fortunately, the examination results are just as good for the obese as for the nonobese patient when the clinician detects an aneurysm. However, the examination is not as efficient at ruling out an aneurysm in obese patients or in those who cannot relax their abdomen to facilitate the examination.

REFERENCE STANDARD TESTS

Imaging studies (ultrasonography or computed tomography).

REFERENCES FOR THE UPDATE

1. Karkos CD, Mukhopadhyay U, Papakostas I, Ghosh J, Thomson GJ, Hughes R. Abdominal aortic aneurysm: the role of clinical examination and opportunistic detection. *Eur J Vasc Endovasc Surg.* 2000;19(3):299-303.
2. Fink HA, Lederle FA, Roth CS, Bowles CA, Nelson DB, Haas MA. The accuracy of physical examination to detect abdominal aortic aneurysm. *Arch Intern Med.* 2000;160(6):833-836.^a
3. Venkatasubramaniam AK, Mehta T, Chetter IC, et al. The value of abdominal examination in the diagnosis of abdominal aortic aneurysm. *Eur J Vasc Endovasc Surg.* 2004;27(1):56-60.
4. Puech-Leao P, Molnar LJ, Oliveira IR, Cerri GG. Prevalence of abdominal aortic aneurysms: a screening program in Sao Paulo, Brazil. *Sao Paulo Med J.* 2004;122(4):158-160.^a
5. Lachat M, Pfammatter T, Moehrlen U, Hilfiker P, Hoerstrup SP, Turina MI. Abdominal pulsatile tumor after endovascular abdominal aortic aneurysm repair. *Vasa.* 1999;28(1):55-57.
6. Zuhrie SR, Brennan PJ, Meade TW, Vickers M. Clinical examination for abdominal aortic aneurysm in general practice: report from the Medical Research Council's General Practice Research Framework. *Br J Gen Pract.* 1999;49(446):731-732.
7. Kadir S, Athanasoulis CA, Brewster DC, Moncure AC. Tender pulsatile abdominal mass: abdominal aortic aneurysm or not? *Arch Surg.* 1980;115(5):631-633.
8. Lederle FA, Wilson SE, Johnson GR, et al. Immediate repair compared with surveillance of small abdominal aortic aneurysms. *N Engl J Med.* 2002;346(19):1437-1444.
9. The United Kingdom Small Aneurysm Trial Participants. Long-term outcomes of immediate repair compared with surveillance of small abdominal aortic aneurysms. *N Engl J Med.* 2002;346(19):1445-1452.
10. US Preventive Services Task Force. Screening for abdominal aortic aneurysm: recommendation statement. *Ann Intern Med.* 2005;142(3):198-202.

^aFor the Evidence to Support the Update for this topic, see <http://www.JAMAevidence.com>.

This page intentionally left blank

EVIDENCE TO SUPPORT THE UPDATE:

2

Abdominal Aortic Aneurysm

TITLE The Accuracy of Physical Examination to Detect Abdominal Aortic Aneurysm.

AUTHORS Fink HA, Lederle FA, Roth CS, Bowles CA, Nelson DB, Haas MA.

CITATION *Arch Intern Med.* 2000;160(6):833-836.

QUESTION How well do commonly used maneuvers work for detecting abdominal aortic aneurysm (AAA)?

DESIGN Each participant underwent physical examination of the abdomen by 2 internists.

SETTING Minneapolis Veterans Affairs Medical Center.

PATIENTS Two hundred participants (aged 51-88 years), 99 with and 101 without AAA as determined by previous ultrasonography.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

The internists were blinded to one another's findings and to the ultrasonographic diagnosis.

MAIN OUTCOME MEASURES

κ , Mean pair agreement, sensitivity, specificity, likelihood ratios, independent predictors of correct diagnosis. The unit of analysis was the examination.

MAIN RESULTS

Interobserver pair agreement for AAA vs no AAA between the first and second examinations was 77% ($\kappa = 0.53$). Sensitivity increased with AAA diameter, from 61% for AAAs of 3.0 to 3.9 cm, to 69% for AAAs of 4.0 to 4.9 cm, 72% for AAAs of 4.0 cm or larger, and 82% for AAAs of 5.0 cm or larger. Sensitivity in subjects with an abdominal girth less than 100 cm (40-in waistline) was 91% vs 53% for girth of 100 cm or greater ($P < .001$). When girth was 100 cm or greater and the aorta was palpable, sensitivity was 82%. When girth was less than 100 cm and the AAA was 5.0 cm or

larger, sensitivity was 100% (12 examinations). Factors independently associated with correct examination findings included AAA diameter (odds ratio [OR], 1.95 per centimeter increase; 95% confidence interval [CI], 1.1-3.6), abdominal girth (OR, 0.90 per centimeter increase; 95% CI, 0.87-0.94), and the examiner's assessment that the abdomen was not tight (OR, 2.7; 95% CI, 1.2-6.7).

The authors provided us data for each examiner according to their degree of confidence in their examination. As expected, these data indicate that an examination "suggestive" of aneurysm conveys considerably less certainty than an examination "definite" for aneurysm (see Table 2-5).

Table 2-5 The Efficiency of the Examination Depends on the Confidence in Your Findings (n = 3 Examiners)

Level of Certainty in Findings	LR+ (95% CI)
Examination "definite" for aneurysm	4.8 (2.7-8.8)
Examination "suggestive"	1.4 (0.92-2.1)
Examination "normal"	0.43 (0.35-0.54)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio.

CONCLUSION

LEVEL OF EVIDENCE Level 3.

STRENGTHS This study was the first to involve sufficient numbers of AAA to examine the effect of patient factors such as obesity, girth, and abdominal tightness and the effect of a palpable aorta. Because previous work indicated that abdominal palpation was insensitive when girth was 100 cm or greater, the authors sought to determine whether subgroups of patients with large girth could be identified in whom abdominal palpation might be reliable. Those with a palpable aorta and large girth had sensitivity of 82%.

LIMITATIONS One likely reason for the increased sensitivities was increased diagnostic vigilance owing to the high prevalence of AAA.

Unlike previous studies that used consecutive patients with relatively low prevalence of AAA, this study included a large

number of patients with AAA to provide power to look at the value of various patient and examination factors. It was also the first study to look at interobserver variability in abdominal palpation for AAA. The mean pair agreement (77%) and κ (0.53) for AAA vs no AAA are considered moderate. Abdominal palpation has only moderate overall sensitivity for detecting AAA but appears to be sensitive for diagnosis of AAAs large enough to warrant elective intervention in patients who do not have a large girth. Abdominal palpation has good sensitivity, even in patients with a large girth, when the aorta is palpable.

Reviewed by Frank A. Lederle, MD

TITLE Prevalence of Abdominal Aortic Aneurysms: A Screening Program in São Paulo, Brazil.

AUTHORS Puech-Leao P, Molnar LJ, Oliveira IR, Cerri GG.

CITATION *Sao Paulo Med J.* 2004;122(4):158-160.

QUESTION How accurate is abdominal palpation for detecting abdominal aortic aneurysm (AAA)?

DESIGN Each subject underwent abdominal palpation with a vascular surgeon and ultrasonography.

SETTING University Hospital, São Paulo, Brazil.

PATIENTS The first 3000 subjects to call in response to an advertising campaign were scheduled for the screening clinic. The study group consisted of 2756 subjects who were older than 50 years, without previous diagnosis of AAA, and for whom ultrasonography was adequate.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

The description of palpation precedes that of ultrasonography in the “Methods,” but we are not told explicitly that palpation was performed before, or blinded to, ultrasonography.

MAIN OUTCOME MEASURES

Palpation result was recorded as positive, negative, or impossible.

AAA was defined as aortic diameter of 3.0 cm or more by ultrasonography. See Table 2-6 for the results of palpation for this study.

MAIN RESULTS

Table 2-6 Results of Palpation in a Large Screening Setting

Palpation	N	No. of AAAs by Ultrasonography
Positive	60	20
Negative	2398	41
Impossible	298	3

Abbreviation: AAA, abdominal aortic aneurysm.

Sensitivity: 20/64 = 31%. Specificity: 2652/2692 = 98%. Positive predictive value of positive examination result: 20/60 = 33%.

CONCLUSION

LEVEL OF EVIDENCE Level 3.

STRENGTHS This is by far the largest study of the sensitivity of palpation to date, comprising nearly as many patients as all previous studies combined. The sensitivity of 31% is somewhat lower than the pooled sensitivity of 39% reported in our original Rational Clinical Examination article, which could result from a greater attenuation of any increased examiner vigilance resulting from study participation.

LIMITATIONS It is not clear from the article that examiners were blinded to the ultrasonographic results, though the low sensitivity would suggest that they were. Although the authors have information on age, sex, and AAA diameter, the effect of these factors on palpation is not described.

Reviewed by Frank A. Lederle, MD

CHAPTER 3

Is Listening for Abdominal Bruits Useful in the Evaluation of Hypertension?

Jeffrey M. Turnbull, MD, FRCP

Toward the end of an unusually busy clinic, a clinical clerk greets the final patient of the day, a man with a recently documented increase of blood pressure. With all the enthusiasm that remains after 4 years of medical training, she compulsively listens for abdominal bruits. Almost surprised, she hears a soft systolic-diastolic epigastric bruit and is faced with the inevitable question: so what?

WHY IS THIS AN IMPORTANT QUESTION TO ANSWER WITH A CLINICAL EXAMINATION?

As we have gained insight into the origin and meaning of vascular bruits, detailed auscultation of the abdomen has become more common. Once detected, an abdominal bruit often is characterized according to pitch, timing, amplitude, and location in an effort to detect and document pathologic states, such as renovascular disease, splenic enlargement, hepatic cirrhosis, carcinoma of the pancreas and liver, splenic and hepatic vascular abnormalities, intestinal vascular insufficiency, and aortic disease. More recently, abdominal bruits have been documented in a substantial percentage of healthy individuals.

Although the search for an abdominal bruit has become part of the general physical examination, it also has been recommended as a key element of the examination of the hypertensive patient, in whom the presence of an abdominal bruit is considered to be an important feature of renovascular hypertension.¹⁻³

To be of value, a diagnostic investigation (such as eliciting an abdominal bruit in the setting of hypertension) must reliably predict the presence or absence of disease (in this case, renovascular hypertension). This process should influence the course of management or prognosis. With this in mind, the reliability and accuracy of auscultating for an abdominal bruit in a patient with hypertension will be examined.

THE ANATOMIC AND PHYSIOLOGIC ORIGIN OF THE ABDOMINAL BRUIT

Whereas turbulent flow within a vessel is the physiologic basis for a bruit, the pitch and radiation are a function of the flow and direction of the turbulent stream. Intrinsic or extrinsic abnormalities can produce turbulence, and although these abnormalities usually arise from within the abdomen, they can also arise from the inguinal area, retroperitoneum, or thorax.

PREVALENCE OF ABDOMINAL BRUITS

The prevalence of bruits in different groups is summarized in Table 3-1. In “normal” populations (individuals without hypertension), the presence of any abdominal bruit has been detected in 6.5% to 31% of patients, with a predilection for the younger age groups (Figure 3-1). Among normal individ-

uals older than 55 years, the prevalence was 4.9%. It is generally believed that the short, faint, and midsystolic bruit heard in these asymptomatic patients is “innocent.”⁷

In patients with angiographically proven renal artery stenosis, bruits have been documented in 77.7% to 86.9% of cases, with higher prevalence than the 28% observed among unselected patients referred for hypertension.^{5,8,9} In a study by Grim et al,¹⁰ the systolic-diastolic bruit was never detected in 379 normal subjects and was found in 1 of 199 patients with essential hypertension.

Eppier et al¹¹ distinguished the presence of abdominal bruits in fibromuscular hyperplasia of the renal artery from that in atherosclerotic lesions. Their retrospective medical record review of 87 patients with surgically treated renal artery stenosis revealed a bruit in 77% of patients with fibromuscular disease and in 35% of patients with atherosclerotic disease.

Table 3-1 The Prevalence of Abdominal Bruits

Reference, y	Age, y	No. and Study Group	Prevalence, %
General Population			
Edwards et al, ⁴ 1970	17-30	200 healthy volunteers	6.5
Julius and Stewart, ⁵ 1967	Unknown	170 volunteers	16
Rivin, ⁶ 1972	16-85	426 patients without cardiovascular or intra-abdominal disease	18
Watson and Williams, ⁷ 1973	13-71	161 psychiatric patients	31
	13-78	200 patients referred with gastrointestinal complaints	27
Patients With Hypertension			
Julius and Stewart, ⁵ 1967		155 patients referred with hypertension	28
Patients With Angiographically Proven Renal Stenosis			
Hunt et al, ⁸ 1974	6-63	100 patients referred for investigation of hypertension	87
Perloff et al, ⁹ 1961	17-72	54 patients referred with sustained hypertension	78

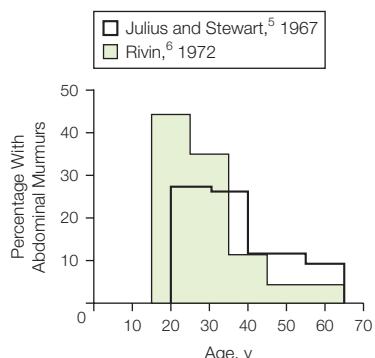


Figure 3-1 The Prevalence of Bruits Varies With Age in Normal Populations

HOW TO EXAMINE FOR ABDOMINAL BRUITS

The patient should be relaxed in a supine position, with the room quiet and with the examiner initially auscultating in the epigastrium, with moderate pressure applied to the diaphragm of the stethoscope. All 4 quadrants should be auscultated anteriorly. The auscultation should continue over the spine and flanks in the areas between T12 and L2 to rule out bruits that may be heard best posteriorly. However, no data exist that would support the routine auscultation of the back for abdominal or retroperitoneal bruits. Once detected, bruits can be correlated to the cardiac cycle by palpation of the carotid upstroke, with the systolic-diastolic bruit being more prolonged and extending into diastole.

Because the kidneys lie retroperitoneally and the renal arteries leave the aorta in the area cephalad to the umbilicus, attention should be given to auscultation in the epigastric area for the bruit of renovascular disease, a pancreatic neoplasm, or an innocent bruit (Figure 3-2). The bruit of a hepatic carcinoma has been heard in the right upper quadrant, whereas that of a splenic arteriovenous fistula has been described in the left upper quadrant. Perumbilical bruits are at times heard in the setting of mesenteric ischemia, and venous hums are from portosystemic hypertension. Finally, in the older population, an abdominal bruit may be associated with an abdominal aortic aneurysm. Estes,¹² in a study of 102 patients with abdominal aortic aneurysms, demonstrated the presence of an associated bruit in 28% of cases.

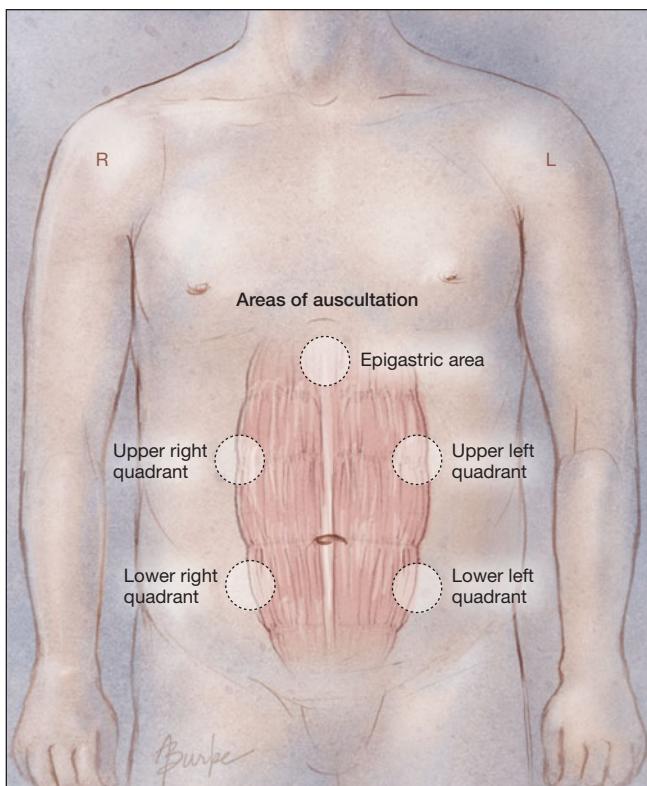


Figure 3-2 Appropriate Areas of Auscultation

THE PRECISION OF ABDOMINAL AUSCULTATION FOR BRUITS

Neither intraobserver nor interobserver variations in the way we elicit this sign have been evaluated in detail. However, Watson and Williams⁷ reported 92% (149/161) agreement when patients with celiac artery compression were prospectively examined by 2 examiners for the presence of an abdominal bruit. With standardization, auscultation of the abdomen can be performed with the appropriate degree of precision.

THE ACCURACY OF ABDOMINAL AUSCULTATION IN RENOVASCULAR HYPERTENSION

This discussion will concentrate on abdominal bruits in fibromuscular and atherosclerotic renovascular disease. Because abdominal bruits occur in healthy individuals and in those with the nonrenovascular conditions listed in Table 3-2, they may occasionally yield false-positive findings in hypertensive patients.

Many studies describe the accuracy of the abdominal bruit in detecting renovascular disease in patients referred for hypertension, but only 3 demonstrate sufficient methodologic rigor (Table 3-3). These reports were of sufficient size and uniform clinical assessment, and the angiogram was the criterion standard. A further study by Julius and Stewart⁵ reported a sensitivity of 20%; however, specificity could not be estimated.

PRESENCE OF ABDOMINAL BRUITS

The most useful study¹⁰ of the accuracy of abdominal auscultation assembled a consecutive series of patients referred to a university medical center for hypertension. All patients healthy enough for surgery underwent careful abdominal auscultation, with positive findings confirmed by a second examiner, plus other tests for renovascular hypertension, including arteriography. Of 64 patients with renovascular hypertension (an abnormal angiogram result and a renal vein renin ratio >1.5), 25 had combined systolic-diastolic abdominal bruits, for a sensitivity of 39% (95% confidence interval [CI], 27%-51%). Of 199 hypertensive patients with normal arteriogram results, 2 had systolic-diastolic bruits, for a specificity of 99% (95% CI, 98%-100%). Thus, although the absence of a systolic-diastolic bruit did not rule out renovascular hypertension, the presence of a systolic-

diastolic bruit helped to rule it in, with a likelihood ratio (LR) of 39 (95% CI, 9.4-160).

A second study recorded any epigastric or flank bruits in a series of hypertensive patients undergoing arteriography.²⁴ Not surprising, the sensitivity of 63% (95% CI, 45%-81%) for any bruit was higher than in the previous study, whereas the specificity for any bruit was somewhat lower, at 90% (95% CI, 84%-96%). Consequently, the presence of any systolic bruit confers a lower LR for renovascular hypertension (LR = 6.4; 95% CI, 3.2-13). Thus, the systolic-diastolic abdominal bruit is less sensitive ($P = .04$; $\chi^2_1 = 4.36$) and more specific ($P < .01$; $\chi^2_1 = 13.5$) than the combination of both isolated systolic and combined systolic-diastolic bruits.

Other than these studies and that by Perloff et al,⁹ additional studies of the accuracy of abdominal bruits in patients with hypertension are less rigorous and are not reported.

In summary, there is a substantial prevalence of systolic bruits in young, healthy patients, which increases in hypertensive patients, especially those with documented renovascular disease. In instances when the accuracy of the abdominal bruit has been rigorously assessed in evaluating patients with renovascular disease, the sensitivity has been reported to be between 20% and 78%, whereas the specificity has been between 64% and 90%. Systolic-diastolic bruits are seldom

Table 3-2 Reported Nonrenovascular Causes of an Abdominal Bruit^a

Reference, y	Condition
Arida, ¹³ 1977	Splenic arteriovenous fistula
Bloom, ¹⁴ 1950	Hepatic cirrhosis
Clain et al, ¹⁵ 1966	Alcoholic hepatitis, hepatoma
Estes, ¹² 1950	Abdominal aortic aneurysm
Goldstein, ¹⁶ 1968	Celiac artery compression syndrome
Lee, ¹⁷ 1967	Bacterial gastroenteritis
Matz and Spear, ¹⁸ 1969	Unilateral renal hypertrophy
McLoughlin et al, ¹⁹ 1975	Celiac artery stenosis
Sarr et al, ²⁰ 1980	Chronic intestinal ischemic
Serebro and W'srand, ²¹ 1965	Pancreatic neoplasia
Shumaker and Waldhausen, ²² 1961	Hepatic arteriovenous fistula
Smythe and Gibson, ²³ 1963	Tortuous splenic arteries

^aNo data exist that would permit the listing of these disorders by prevalence.

Table 3-3 Accuracy of the Abdominal Bruit in Renovascular Hypertension

Reference, y	Type of Bruit	Sensitivity, % (95% CI ^a)	Specificity, %	LR	
				If Present	If Absent
Grim et al, ¹⁰ 1979	Systolic and diastolic abdominal bruit	25/64 = 39 (27-51)	197/199 = 99 (98-100)	39	0.6
Fenton et al, ²⁴ 1966	Any epigastric or flank bruit, including isolated systolic bruit	17/27 = 63 (45-81)	82/91 = 90 (84-96)	6.4	0.4
Perloff et al, ⁹ 1961	Systolic bruit	78	64	2.1	0.35

Abbreviations: CI, confidence interval; LR, likelihood ratio.

^aCI obtained with the use of normal approximation method.

heard in healthy people or in patients with essential hypertension, but they are more common in individuals with renovascular disease. In patients with fibromuscular disease, there is an increased prevalence for all types of bruits.

AUSCULTATORY CHARACTERISTICS OF BRUITS

Although many bruits have been characteristically described as having a certain pitch, intensity, and location, the data to support this have been questioned.^{11,19} Moser and Caldwell²⁵ demonstrated a slightly increased prevalence of high-pitched bruits in association with renal artery disease (87%) when compared with the prevalence of medium-pitched or low-pitched bruits (57%). This finding supports the results of Julius and Stewart,⁵ who reported an increased prevalence (64%) of high-pitched bruits in these patients.

In the study by Moser and Caldwell,²⁵ the intensity of the bruit described in patients with renovascular disease was less discriminatory, with 80% (17/21) of cases having loud bruits and 55% (16/29) having quiet bruits. These same authors described their results in predicting the localization of the stenosis. In their study, of the 13 patients in whom renovascular disease was isolated to 1 vessel, stenosis was correctly localized beforehand in 6 (46%). Eppier et al¹¹ reported slightly better results because the site of the renovascular lesion was correctly localized in 70% of patients with fibromuscular disease and 43% of patients with atherosclerotic renovascular disease. Julius and Stewart⁵ directly auscultated the renal artery by using a sterile stethoscope at the time of renovascular surgery, demonstrating that, of 18 patients with bruits, in 9 the bruits were confined to the correct renal artery and in 7 the renal artery bruits were combined with additional vascular bruits. In 2 patients (11%), the bruits heard before surgery were secondary to other vascular abnormalities, and there were no bruits associated with the renal artery.

PROGNOSIS OF PATIENTS WITH HYPERTENSION AND BRUITS

Finally, the importance of identifying the location, pitch, and intensity of a bruit is questionable, and this issue awaits further clarification with larger prospective studies. Two reports have linked the presence of bruits to the outcome of renovascular surgery but with conflicting results. Eppier et al¹¹ found that 84% of patients with systolic-diastolic bruits had favorable surgical results, compared with 55% of patients with only systolic bruits or no bruits. This result was replicated in patients whose renal artery stenoses were due to atherosclerosis, but the presence of diastolic bruits and the recent onset of hypertension correlated with favorable surgical outcomes in patients with both fibromuscular and atherosclerotic vascular disease. In contrast, Simon et al²⁶ were unable to attach prognostic importance to abdominal bruits in patients with fibromuscular or atherosclerotic renovascular disease.

THE BOTTOM LINE

In view of the high prevalence (7%-31%) of innocent abdominal bruits in the younger age groups, if a systolic abdominal bruit is detected in a young, normotensive, asymptomatic individual, no further investigations are warranted. In view of the low sensitivity, the absence of a systolic bruit is not sufficient to rule out the diagnosis of renovascular hypertension. In view of the high specificity, the presence of a systolic bruit (in particular a systolic-diastolic bruit) in a hypertensive patient is suggestive of renovascular hypertension. Subsequent investigation should take into consideration the pretest likelihood of renovascular disease and full cost and potential benefits of any management decision. In view of the lack of evidence to support characterizing bruits as to pitch, intensity, and location, bruits should be reported only as systolic or systolic/diastolic. Existing information does not permit a definitive statement pertaining to the prognostic implication of a renal bruit.

In summary, the critical review of the literature pertaining to the abdominal bruit would suggest that the routine auscultation of the abdomen for the presence or absence of an abdominal bruit in the healthy asymptomatic population is of little value in view of the high prevalence of benign bruits. However, for our troubled clinical clerk, the presence of a systolic-diastolic bruit would provide supportive evidence of an underlying diagnosis of renovascular disease and should lead her to more aggressive investigation for this disorder.

Author Affiliation at the Time of the Original Publication

From the Department of Medicine, Ottawa (Ontario) Civic Hospital, Ottawa, Ontario, Canada.

Acknowledgments

The author and editors thank E. K. M. Smith, MD, for his helpful review of the manuscript.

REFERENCES

1. Gifford RW, Poutasse EF. Renal artery disease: diagnosis and therapy. *Geriatrics*. 1963;8:761.
2. Maronde RF. The hypertensive patient. *JAMA*. 1975;233(9):997-1000.
3. Havey RJ, Krumlovsky F, delGreco F, Martin HG. Screening for renovascular hypertension. *JAMA*. 1985;254(3):388-393.
4. Edwards AJ, Hamilton JD, Nichol WD, et al. Experience coeliac axis compression syndrome. *BMJ*. 1970;1(5692):342-345.
5. Julius S, Stewart BH. Diagnostic significance of and murmurs. *N Engl J Med*. 1967;276(21):1175-1178.
6. Rivin AU. Abdominal vascular sounds. *JAMA*. 1972;221(7):688-690.
7. Watson WC, Williams PB. Epigastric bruits in patients without celiac axis compression. *Ann Intern Med*. 1973;79(2):211-215.
8. Hunt JC, Sheps SG, Harrison EG Jr, Strong CG, Bernatz PE. Renal and renovascular hypertension: a reasoned approach to diagnosis and management. *Arch Intern Med*. 1974;133(6):988-999.
9. Perloff D, Sokolow M, Wylie EJ, et al. Hypertension secondary to renal artery occlusive disease. *Circulation*. 1961;24:1286-1304.
10. Grim CE, Luft FC, Myron H, et al. Sensitivity and specificity of screening tests for renal vascular hypertension. *Ann Intern Med*. 1979;91(4):617-622.
11. Eppier DF, Gifford RW, Stewart BH, et al. Abdominal bruits in renovascular hypertension. *Am J Cardiol*. 1976;37(1):48-52.
12. Estes E. Abdominal aortic aneurysm. *Circulation*. 1950;2(2):258-264.

13. Arida EJ. Splenic arteriovenous fistula with hypertension, varices, and ascites. *N Y State J Med.* 1977;77(6):987-990.
14. Bloom HJG. Venous hums in hepatic cirrhosis. *Br Heart J.* 1950;12(4): 343-350.
15. Clain D, Wartnaby K, Sherlock S. Abdominal arterial murmurs in liver disease. *Lancet.* 1966;2(7462):516-519.
16. Goldstein LT. Enlarged, tortuous arteries and hepatic bruit. *JAMA.* 1968;206(11):2518-2520.
17. Lee R. Provision of health services: past, present, future. *N Engl J Med.* 1967;277(13):682-686.
18. Matz R, Spear P. Abdominal bruit due to unilateral renal hypertrophy. *Lancet.* 1969;1(7589):310-311.
19. McLoughlin MJ, Colapinto RJ, Hobbs BB. Abdominal bruits. *JAMA.* 1975;232(12):1238-1242.
20. Sarr MG, Dickson ER, Newcomer AD. Diastolic bruit in chronic intestinal ischemia. *Dig Dis Sci.* 1980;25(10):761-762.
21. Serebro H, W'srand MB. Diagnostic sign of carcinoma of the body of the pancreas. *Lancet.* 1965;1:85-86.
22. Shumaker HB, Waldhausen JA. Intrahepatic arteriovenous fistula of hepatic artery and portal vein. *Surg Gynecol Obstet.* 1961;112:497-501.
23. Smythe C, Gibson DB. Upper-quadrant bruit due to tortuous artery. *N Engl J Med.* 1963;2:1308-1309.
24. Fenton S, Lytle JA, Pantridge JF. Diagnosis and results of surgery in renovascular hypertension. *Lancet.* 1966;2(7455):117-121.
25. Moser RJ, Caldwell JR. Abdominal murmurs. *Ann Intern Med.* 1962;56: 471-483.
26. Simon N, Franklin SS, Bleifer KH, Maxwell MH. Clinical characteristics of renovascular hypertension. *JAMA.* 1972;220(9):1209-1218.

This page intentionally left blank

Prepared by David L. Simel, MD, MHS

Reviewed by Lori Orlando, MD

CLINICAL SCENARIO

A 55-year-old, white, male smoker has had hypertension for 10 years. It has always been well controlled, with systolic measures of lower than 35 mm Hg. He is receiving a diuretic and a β -blocker. Recently, the systolic pressure has typically been 140 to 150 mm Hg. He is a bit overweight (body mass index, 26.5). There has been no evidence for atherosclerotic disease. His serum creatinine level is unchanged, at 0.11 $\mu\text{mol/L}$. The serum cholesterol level is 5.95 mmol/L. Your suspicion is that the increased blood pressure is a manifestation of essential hypertension, but you decide to auscultate for an abdominal bruit. You hear none. You would like to add an angiotensin-converting enzyme inhibitor, but you wonder whether you have ruled out renal artery stenosis as a cause of the recent upward trend in his pressure.

UPDATED SUMMARY ON ABDOMINAL BRUITS

Original Review

Turnbull JM. Is listening for abdominal bruits useful in the evaluation of hypertension? *JAMA*. 1995;274(16):1299-1301.

UPDATED LITERATURE SEARCH

Our literature search crossed the text words “renal artery,” “auscultation,” “bruit,” and “hypertension,” published in English from 1994 to 2004. We also searched on the subject heading “renal artery obstruction/di.” The search yielded 86 articles for which the titles and abstracts were reviewed. One article that included sensitivity and specificity data on the abdominal bruit as a sign for renal artery stenosis was retrieved.

NEW FINDINGS

- A large study of patients with hypertension that is difficult to control confirmed the usefulness of finding an abdominal bruit, even those heard only during systole.
- Available data do not allow us to make conclusions about the prevalence or importance of finding an abdominal bruit in black patients.

Details of the Update

Many normal individuals have abdominal bruits. The presence of an abdominal bruit becomes potentially important in hypertensive patients, especially those with certain characteristics. Abdominal bruits may be the harbinger of renal artery stenosis, and the diagnosis should be suspected in hypertensive patients who had their disease onset at a young age or who have blood pressures that are seemingly resistant to medical treatment. It may be therapeutically useful to identify patients with renal artery stenosis because balloon angioplasty may be a useful treatment intervention for controlling blood pressure, especially when medications fail.¹

One study, identified in the original Rational Clinical Examination article, found the highest diagnostic utility for an abdominal bruit that had both a systolic and diastolic component. The effect of an abdominal bruit with both components compared with an abdominal bruit with only a single systolic component has not been evaluated. In our updated literature review, we found 1 large, prospective cohort study of patients with hypertension that is difficult to control who were systematically evaluated for renal artery stenosis. The importance of a systolic bruit in this population of patients (predominantly white) was similar to that found in previous work that we reviewed in the original publication.²

A study of 85 consecutive patients with hypertension, diabetes, and normal renal function provides useful information about ethnicity and renal artery stenosis as it includes a higher proportion of black patients than previous studies.³ The odds ratio for Afro-Caribbean patients vs other patients (white or Asian) was 0.70 (95% confidence interval [CI], 0.19-2.5). We can combine the data with those from Krijnen et al² to find a summary odds ratio of 0.37 (95% CI, 0.12-1.1) for black ethnicity, suggesting that perhaps black patients are less likely than other patients to get renal artery stenosis. However, the broad CIs suggest that the currently available data do not allow us to conclude this with certainty. Unfortunately, data were not provided on the frequency of abdominal bruits, so we do not know whether the finding of an abdominal bruit in black patients has the same significance as in other patients.

IMPROVEMENTS IN THE DATA PRESENTED IN THE ORIGINAL PUBLICATION

CIs were not provided in the original publication. A typographic error in the negative likelihood ratio (LR-) for a bruit was found for Table 3-3. The LR- for the study by Perloff et al⁴ should have been 0.35, as is now shown. We reconfigured Table 3-3 from the original publication, providing the CIs and summary estimates for the presence of a bruit (Table 3-4).

CHANGES IN THE REFERENCE STANDARD

The reference standard remains arteriography. However, noninvasive tests have replaced arteriography in offering a less risky screening approach for appropriate patients.⁵ At possible treatment (ie, balloon angioplasty), all patients undergo arteriography to ensure proper technique.

RESULTS OF LITERATURE REVIEW

Multivariate Findings for Renal Artery Stenosis

A clinical prediction model can be used in white patients with hypertension that is difficult to control.² The model can be downloaded to a computer (the DRASTIC [Dutch Renal Artery Stenosis Intervention Cooperative] spreadsheet; <http://www2.eur.nl/fgg/mgz/software.html>, accessed May 16, 2008). The model has not been validated prospectively or in a population of blacks.

Table 3-4 Univariate Findings for Renal Artery Stenosis

Finding	LR+ (95% CI)	LR- (95% CI)
Bruit		
Systolic and diastolic		
Grim et al ⁶	39 (10-145)	0.62 (0.49-0.73)
Systolic with/without diastolic component ^a		
Krijnen et al ²	6.7 (3.7-12)	0.76 (0.66-0.84)
Fenton et al ⁷	6.4 (3.2-12)	0.41 (0.24-0.62)
Perloff et al ⁴	2.2 (1.5-3.2)	0.35 (0.20-0.57)
Summary systolic bruit	4.3 (2.3-8.0)	0.52 (0.34-0.78)
History of atherosclerotic disease ²	2.2 (1.8-2.8)	0.52 (0.40-0.66)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

^aDid not distinguish between individuals with systolic-only bruits vs systolic and diastolic.

EVIDENCE FROM GUIDELINES

The Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure (JNC 7) suggests that physicians auscultate for abdominal bruits in patients with hypertension.⁸ The suggestion is not accompanied with data but is an expert's opinion. The report specifically recommends considering renal artery stenosis for certain hypertensive patients.

CLINICAL SCENARIO—RESOLUTION

Patients with hypertension frequently need treatment with additional medications as they get older. The patient has none of the more obvious findings to suggest renovascular hypertension from renal artery stenosis. According to expert recommendations, you listened for abdominal bruits and heard none. The proper technique must be used, and you must be listening in a quiet room. Often, physicians do not apply enough pressure with the diaphragm of the stethoscope. Had you heard a bruit, you would have attempted to see whether the bruit extends into diastole. This can be done by palpating the carotid while listening to see whether the bruit prolongs beyond the carotid upstroke.

The LR data for the presence or absence of systolic bruits apply only to patients with resistant hypertension. With just 2 medications, you should not assume that he has resistant hypertension. Thus, the LR for the absence of bruit cannot be applied to this patient. You might resort to a clinical decision model (referenced above). Given his age, smoking status, sex, body weight, absence of a bruit, long history of hypertension, and normal creatinine and cholesterol levels, you would find that his predicted probability of renovascular stenosis is 10%. Two caveats apply to this model—it was also developed with data from patients with resistant hypertension, so his probability of renal artery stenosis is probably even lower. Second, had your patient been black, you would have needed to recognize that the accuracy of the model would be unknown.

RENAL ARTERY STENOSIS—MAKE THE DIAGNOSIS

Patients without hypertension should not have auscultation for asymptomatic renal artery bruits because bruits frequently are a normal finding. The search for renal artery stenosis should be confined to certain patient populations (see below). When present in these populations, an abdominal bruit is the most useful physical examination finding for assessment of renal artery stenosis.

PRIOR PROBABILITY OF RENOVASCULAR DISEASE

Approximately 1% to 5% of the general population has renovascular disease. Approximately 20% of white patients with medically refractory hypertension have renal artery stenosis.

POPULATION FOR WHOM RENAL ARTERY STENOSIS SHOULD BE CONSIDERED

- Onset of hypertension before 30 years of age
- Patients with an arterial bruit and hypertension, especially if there is a diastolic component
- Accelerated hypertension
- Hypertension that becomes resistant to medication
- Flash pulmonary edema
- Renal failure, especially in the absence of proteinuria or an abnormal urine sediment result
- Acute renal failure precipitated by angiotensin-converting enzyme inhibitors or angiotensin-receptor blockers

DETECTING THE LIKELIHOOD OF RENAL ARTERY STENOSIS IN PATIENTS WITH REFRACTORY HYPERTENSION

See Table 3-5.

Table 3-5 Clinical Examination Findings for Renal Artery Stenosis

Finding (No. of Studies)	LR+ (95% CI)	LR- (95% CI)
Systolic-diastolic bruit (n = 1)	39 (10-145)	0.62 (0.49-0.73)
Systolic bruit (n = 3)	4.3 (2.3-8.0)	0.52 (0.34-0.78)
History of atherosclerotic disease (n = 1)	2.2 (1.8-2.8)	0.52 (0.40-0.66)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

REFERENCE STANDARD TESTS

Moderate-risk and high-risk patients are subjected to a noninvasive screening test (ultrasonography, computed tomography, magnetic resonance imaging). The type of imaging modality for screening (eg, contrast-enhanced ultrasonography vs gadolinium-enhanced computed tomography or magnetic resonance angiography) may be operator dependent, and physicians will need to rely on their local radiologists' expertise. All patients have their disease status confirmed with arteriography as part of a therapeutic procedure.

REFERENCES FOR THE UPDATE

1. Nordmann AJ, Woo K, Parkes R, Logan AG. Balloon angioplasty or medical therapy for hypertensive patients with atherosclerotic renal artery stenosis: a meta-analysis of randomized controlled trials. *Am J Med*. 2003;114(1):44-50.
2. Krijnen P, van Jaarsveld BC, Steyerberg EW, Man in 't Veld AJ, Schalekamp MA, Habbema JD. A clinical prediction rule for renal artery stenosis. *Ann Intern Med*. 1998;129(9):705-711.^a
3. Valabhji J, Robinson S, Poulter C, et al. Prevalence of renal artery stenosis in subjects with type 2 diabetes and coexistent hypertension. *Diabetes Care*. 2000;23(4):539-543.
4. Perloff D, Sokolow M, Wylie EJ, et al. Hypertension secondary to renal artery occlusive disease. *Circulation*. 1961;24:1286-1304.
5. Vasbinder GB, Nelemans PJ, Kessels AG, Kroon AA, de Leeuw PW, van Engelshoven JM. Diagnostic tests for renal artery stenosis in patients suspected of having renovascular hypertension: a meta-analysis. *Ann Intern Med*. 2001;135(6):401-411.
6. Grim CE, Luft FC, Myron H, et al. Sensitivity and specificity of screening tests for renal vascular hypertension. *Ann Intern Med*. 1979;91(4):617-622.
7. Fenton S, Lytle JA, Pantridge JF. Diagnosis and results of surgery in renovascular hypertension. *Lancet*. 1966;2(7455):117-121.
8. Chobanian AV, Bakris GL, Black HR, et al. The seventh report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure. *JAMA*. 2003;289(19):2560-2572.

^aFor the Evidence to Support the Update for this topic, see <http://www.JAMAEvidence.com>.

This page intentionally left blank

EVIDENCE TO SUPPORT THE UPDATE:

Abdominal Bruits

TITLE A Clinical Prediction Rule for Renal Artery Stenosis.**AUTHORS** Krijnen P, van Jaarsveld BC, Steyerberg EW, Man in 't Veld AJ, Schalekamp MA, Habbema JD.**CITATION** *Ann Intern Med.* 1998;129(9):705-711.**QUESTION** Do clinical data identify patients likely to have renal artery stenosis?**DESIGN** Prospective data collected as part of a cohort study.**SETTING** Multiple internal medicine departments in the Netherlands.**PATIENTS** One thousand one hundred thirty-three patients, aged 18 to 75 years, with normal serum creatinine levels and referred for hypertension evaluations. Most patients had hypertension that was difficult to control.**DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD**

Patients were assigned to 1 of 2 treatment protocols. Those who had a mean diastolic blood pressure of 95 mm Hg or higher at follow-up, or those who experienced an increase in serum creatinine level when treated with angiotensin-converting enzyme inhibitor, underwent digital subtraction angiography, and underwent other noninvasive tests of the renal arteries.

The clinical data were collected prospectively. The presence of "abdominal bruit" was recorded before the reference standard tests.

MAIN OUTCOME MEASURES

Renal artery stenosis ($\geq 50\%$) identified by arteriography.

MAIN RESULTS

From a population of 1133 patients, 477 required renal artery stenosis evaluation for either blood pressure that is difficult to control or an increase in serum creatinine level when treated with an angiotensin-converting enzyme inhibitor. One hundred seven patients had renal artery stenosis (22%).

Table 3-6 Likelihood Ratio of Findings for Renal Artery Stenosis

Test	Sensitivity	Specificity	LR+ (95% CI)	LR- (95% CI)
Abdominal bruit	0.27	0.96	6.7 (3.7-12)	0.76 (0.66-0.84)
Atherosclerotic disease	0.63	0.72	2.2 (1.8-2.8)	0.52 (0.40-0.66)

Abbreviations: CI, confidence interval; LR, likelihood ratio.

Abdominal bruit or atherosclerotic disease (femoral or carotid bruit, angina, claudication, myocardial infarction, cerebrovascular accident, or vascular surgery) were the variables with the best accuracy (Table 3-6). A clinical prediction model included the additional terms of age, smoking history, recent onset of hypertension, obesity, hypercholesterolemia, and the serum creatinine level. The model can be downloaded via the Internet (the DRASTIC [Dutch Renal Artery Stenosis Intervention Cooperative] spreadsheet; <http://www2.eur.nl/fgg/mgz/software.html>, accessed May 16, 2008). The model had an area under the receiver operating characteristic curve (a measure of accuracy) of 0.84 (95% confidence interval, 0.79-0.89).

CONCLUSIONS**LEVEL OF EVIDENCE** Level 1.

STRENGTHS Prospective data collection in the relevant population of patients with hypertension that is difficult to control. The prediction model was subjected to internal validation.

LIMITATIONS "Abdominal bruit" is not defined. The study population had almost no patients of black ethnicity. The prediction rule was not externally validated in a separate population of patients.

This is a large study in the population of patients for whom renovascular hypertension and renal artery stenosis might be considered. The presence of any abdominal bruit was recorded by examiners and showed excellent specificity with a sufficiently high positive likelihood ratio. A patient's history that indicates previous atherosclerotic vascular disease also has diagnostic utility.

A problem for some clinicians is that the patients were almost all whites.¹ Given the low prevalence of renovascular hypertension in blacks, US physicians cannot be certain that the results will generalize well.

Reviewed by David L. Simel, MD, MHS

REFERENCE FOR THE EVIDENCE

1. Wilcox C. Screening for renal artery stenosis: are scans more accurate than clinical criteria? *Ann Intern Med.* 1998;129(9):738-740.

CHAPTER 4

Does This Patient Have an Alcohol Problem?

James M. Kitchens, MD, FRCPC

CLINICAL SCENARIO

A 58-year-old man was admitted to the hospital for an elective cholecystectomy. At the time of admission, he smelled of alcohol, although he was not obviously intoxicated. On questioning, he said that he had come from a business lunch where he had “a drink.” When questioned about his alcohol history, he became angry and defensive. He said that he was “offended by the implications of these questions.” On the day after the surgery, he was found to be diaphoretic, tremulous, and hallucinating and was judged to be in alcohol withdrawal. Could other interviewing techniques have identified this man as one who was alcohol dependent and at risk of withdrawal?

WHY IS THIS AN IMPORTANT QUESTION TO ANSWER WITH A CLINICAL EXAMINATION?

It is estimated that more than 100 million Americans drink alcohol and that about 10% of those who drink have alcohol problems that adversely affect their lives and the lives of their families.¹ Alcohol is involved in 10% of all deaths in the United States. The mortality rate in those who drink 6 or more drinks per day is 50% higher than the rate in matched controls.² Alcohol is a major factor in suicides, homicides, violent crimes, and fatal motor vehicle crashes. Alcohol abuse and dependence are common in both partners where spouse and child abuse occur.^{3,4} There is a 4-fold increased risk of alcohol dependence in the children of alcohol-dependent parents.⁵

Alcohol is primarily or secondarily implicated in a large number of medical problems such as cirrhosis, alcoholic hepatitis, portal hypertension, gastritis, nutritional deficiencies, cardiomyopathy, dysrhythmias, cognitive dysfunction, seizures, neuropathies, myopathies, low birth weight, fetal alcohol syndrome, and a variety of head and neck cancers.¹

Alcohol abuse and alcohol dependence are common problems. A history of alcohol abuse has been found in one-fifth to one-third of patients attending inner-city ambulatory medical clinics, and one-third of these patients report an active drinking problem. In some of these settings, the prevalence of abuse has been as high as two-thirds in men.⁶⁻⁸ Unfortunately, physicians recognize only about half of the problem drinkers that they encounter, and they are even less likely to identify problems in women and elderly people.⁸⁻¹³

DIAGNOSTIC STANDARDS FOR ALCOHOL ABUSE AND DEPENDENCY

Alcohol-related problems provide many diagnostic problems for clinicians. In our society, drinking is a common and socially complex behavior. At one end of the drinking spec-

Table 4-1 Diagnostic and Statistical Manual of Mental Disorders, Revised Third Edition (DSM-III-R) and International Statistical Classification of Diseases, 10th Revision (ICD-10) Diagnostic Criteria for Substance Abuse, Harmful Use, and Substance Dependence

DSM-III-R Dependence (3 Items Required)
1. Substance often taken in larger amounts or during a longer period than the person intended
2. Persistent desire or 1 or more unsuccessful efforts to cut down or control substance use
3. A great deal of time spent in activities necessary to get substance, taking substance, or recovering from its effects
4. (a) Recurrent use when substance use is physically hazardous (eg, drives while intoxicated) or (b) frequent intoxication or withdrawal symptoms when expected to perform major role obligations at work, school, or home
5. Important social, occupational, or recreational activities given up or reduced because of substance use
6. Continual substance use despite knowledge of having persistent or recurrent social, psychological, or physical problem that is caused or exacerbated by the use of substance
7. Marked tolerance: need for markedly increased amounts of substance (at least a 50% increase) to achieve intoxication or desired effect or markedly diminished effect with continued use of the same amount
8. Characteristic withdrawal symptoms
9. Substance often taken to relieve or avoid withdrawal symptoms.
DSM-III-R Abuse
1. Continued use despite knowledge of having persistent or recurrent social, occupational, psychological, or physical problem that is caused or exacerbated by the use of substance
2. Recurrent use in situations in which use is physically hazardous
ICD-10 Dependence (3 Items Required)
1. A strong desire or sense of compulsion to use a substance
2. Evidence of impaired capacity to control the use of a substance. This may relate to difficulties in avoiding initial use, difficulties in terminating use, or problems controlling levels of use.
3. A withdrawal state or use of the substance to relieve or avoid withdrawal symptoms and subjective awareness of the effectiveness of such behavior
4. Evidence of tolerance of the effects of the substance
5. Progressive neglect of alternative pleasures, behaviors, or interests in favor of substance use
6. Persisting with substance use despite clear evidence of harmful consequences
ICD-10 Harmful Use
1. Clear evidence that the use of a substance was responsible for causing actual psychological or physical harm to the user

trum, alcohol is used in moderation without adverse consequences to the drinkers or those around them. At the other end of the spectrum are those drinkers who have adverse effects medically, economically, and psychosocially from repeated abuse of alcohol. Between those who occasionally use alcohol in moderation and those who are frankly alcohol dependent lies a continuum of drinkers with varying consumption patterns and risks of alcohol-related problems.

The rational use of diagnostic tests to identify problem drinking or alcohol dependence demands a clear understanding of the definitions of the disorder being diagnosed. It will also become clear that diagnostic test characteristics,

such as sensitivity, specificity, and likelihood ratios (LRs), vary considerably, depending on the definition of problem drinking or alcohol dependence.

The *International Statistical Classification of Diseases, 10th Revision (ICD-10)* and the *Diagnostic and Statistical Manual of Mental Disorders, Revised Third Edition (DSM-III-R)* of the American Psychiatric Association present guidelines for the diagnosis of substance abuse disorders.^{14,15} The *ICD-10* recognizes 2 categories: harmful use and alcohol dependence. The *DSM-III-R* recognizes 2 categories: alcohol abuse and alcohol dependence. The diagnostic criteria for *DSM-III-R* and *ICD-10* are found in Table 4-1. There is another edition of the *DSM*, the *DSM-IV*. It is not significantly different from *DSM-III-R* with regard to the diagnosis of alcohol abuse and alcohol dependence. The following discussion refers to *DSM-III-R* because it has been used as a diagnostic standard for comparison with other diagnostic questionnaires.

Alcohol dependence represents a syndrome as diagnosed by *DSM-III-R* and *ICD-10*. The syndrome criteria of the 2 systems overlap considerably, but there are differences between *DSM-III-R* and *ICD-10*. The *ICD-10* does not include items that address the social or legal consequences of dependence, nor does it have criteria that assess dangerous use (eg, driving or working while intoxicated). The *ICD-10* criteria are restricted to the medical and psychological consequences of abuse and dependence. Despite these differences, there is excellent concordance between *DSM-III-R* and *ICD-10* in the diagnosis of alcohol dependence.¹⁶ This high degree of concordance illustrates the fact that dependence most commonly affects medical, psychological, and social aspects of life. Rarely are the consequences restricted to one sphere of life.

The *ICD-10* and *DSM-III-R* have separate categories of harmful or abusive drinking that do not meet the criteria for dependence. However, there is poor concordance between the 2 systems for these diagnostic categories.¹⁶ Because it does not include criteria for social/legal consequences of drinking, *ICD-10* makes fewer diagnoses than *DSM-III-R* does. For example, an individual who repeatedly drives while intoxicated would not be assigned a diagnosis under *ICD-10* but would be assigned a diagnosis as an alcohol abuser under *DSM-III-R*.

The *DSM-III-R* is the most widely used diagnostic framework for alcohol-related disorders, and it has been used as the diagnostic standard for comparison of other diagnostic questionnaires.^{6,7,15} The *DSM-III-R* criteria for alcohol abuse or dependence are structured to detect alcohol problems at any time in the life of the patient. This lifetime prevalence of alcohol problems may not represent an individual's current drinking status.⁸ Most studies that use the *DSM-III-R* criteria as the diagnostic standard for the identification of alcohol abuse or dependence also use a published structured interview, such as the Structured Clinical Interview for *DSM-III-R* (SCID), that asks specific interview questions that relate to the *DSM-III-R* diagnostic criteria.¹⁷

Other studies have used alcohol consumption questionnaires and interviews to define a level of "problem drinking" and then examined the diagnostic accuracy of screening

questionnaires to separate problem drinkers from nonproblem drinkers.^{7,18-21} However, in the following section, it will be seen that the sensitivities of screening questionnaires decrease as the definition of problem drinking is changed to include a greater proportion of at-risk drinkers.

It is clear that excessive alcohol consumption may be detrimental to medical and social health. The dangers associated with alcohol consumption represent a continuum of risk that makes it difficult to define "safe levels" of alcohol consumption. Some authors contend that ingestion of 4 or more drinks per day in men and 2 or more drinks per day in women constitute a "hazardous" consumption level that increases the risk of alcohol dependence and medical problems.^{4,22,23} A "drink" is defined as equivalent volume amounts that have an ethanol content of 0.6 oz. Twelve ounces of beer, 5 oz of wine, and 1.5 oz of liquor all contain 0.6 oz of ethanol. However, safe levels of consumption vary considerably, depending on the clinical or social context of drinking. One and one-half drinks per day may constitute at-risk drinking for pregnant women and represent a health threat to the developing child.^{18,24}

The World Health Organization (WHO) has developed a questionnaire, the Alcohol Use Disorders Identification Test (AUDIT), to identify persons with "hazardous" and "harmful" alcohol consumption who may not be captured by *DSM-III-R* or *ICD-10* diagnostic criteria.²⁵ WHO recognizes the following disorders of alcohol use: "Hazardous drinking" is use that increases the risk of subsequent psychological or medical harm and is judged to be 4 or more drinks per day in men and 2 or more drinks per day in women. "Harmful drinking" occurs in the person who has psychological or medical complications as defined in *ICD-10*. The WHO classification system attempts to identify persons who drink quantities that will increase their risk of subsequent problems. This modification is driven by concerns about the cost and effectiveness of treating alcohol dependence.²⁵ A review of alcohol treatment programs and their effectiveness is beyond the scope of this article, but there is a substantial body of evidence that brief, ambulatory interventions targeted to persons with hazardous drinking can decrease levels of consumption and, it is hoped, decrease the likelihood of subsequent harm and dependence.²⁶ However, diagnosis must precede treatment. It is the diagnosis of alcohol disorders in the context of the medical history that is the subject of the remainder of this article.

DIAGNOSTIC TESTS OF ALCOHOL ABUSE AND DEPENDENCY

Several questionnaires have been developed for the detection of alcohol disorders, including the *cut down*, *annoyed by criticism*, *guilty about drinking*, *eye-opener* drinks (CAGE) questionnaire, the Michigan Alcoholism Screening Test (MAST), and the AUDIT. The most widely used are the CAGE questionnaire and the MAST. Of these, the MAST has been more thoroughly studied in terms of reliability and accuracy. However, the MAST and its shortened versions are

more complicated than the CAGE questionnaire. The CAGE questionnaire is short, easily memorized, and reasonably accurate, making it the screening test of choice for busy house officers and practitioners.

CAGE Questionnaire

In 1968, Ewing²⁷ developed the CAGE questionnaire for the detection of alcoholism. CAGE is mnemonic for these 4 questions: (1) Have you ever felt you ought to cut down on your drinking? (2) Have people annoyed you by criticizing your drinking? (3) Have you ever felt bad or guilty about your drinking? (4) Have you ever had a drink first thing in the morning to steady your nerves or get rid of a hangover (eye opener)?

Some investigators have reasoned that alcohol abusers are more likely to give accurate responses to the CAGE questions if they are part of a series of questions on lifestyle that include drinking, smoking, diet, and exercise habits.^{7,28} The rationale behind this approach is that it may be less likely to trigger defensiveness and denial in people who are alcohol dependent. Other studies do not attempt to disguise the CAGE questionnaire. No studies that examine differences between CAGE interviews and written CAGE questionnaires were identified. There are no comparative studies of reliability or accuracy for the different modes of administering the CAGE questions. It seems reasonable to ask these questions in a frank, nonjudgmental manner as part of the medical history or review of symptoms.

MAST

The MAST was originally reported on by Selzer²⁹ in 1971. The MAST consists of 24 yes/no questions, with the "alcohol dependent" responses being scored as 1, 2, or 5 points. The MAST questions are listed in Table 4-2. The most common scoring for the MAST has 0 to 3 points as "non-alcohol dependent," 4 or 5 as "probably alcohol dependent," and greater than 5 as "definitely alcohol dependent."

Two modified, shortened versions of the MAST have been developed to make it a less time-consuming screening instrument for alcohol dependence. A 10-question version, the Brief MAST (BMAST), and a 13-question version, the Short MAST (SMAST), are available.^{30,31}

AUDIT

WHO sponsored a collaborative project to develop a screening test that would be able to detect persons with hazardous levels of consumption and those with harmful use and dependence. The AUDIT questions are listed in Table 4-3. Answers are scored from 0 to 4, for a maximum score of 40 points, with scores of 8 or more considered diagnostic of an alcohol use disorder.^{25,32}

Biochemical and Hematologic Tests

Increases in liver enzyme concentrations (aspartate aminotransferase, alanine aminotransferase, and γ -glutamyltrans-

Table 4-2 Michigan Alcoholism Screening Test (MAST)²⁹

Points	Question
2	1. Do you feel you are a normal drinker? ^a
2	2. Have you ever awakened the morning after some drinking the night before and found that you could not remember a part of the evening before?
1	3. Does your spouse or parents ever worry or complain about your drinking? ^a
2	4. Can you stop drinking without a struggle after 1 or 2 drinks?
1	5. Do you ever feel bad about your drinking? ^a
2	6. Do friends or relatives think you are a normal drinker? ^a
2	7. Are you always able to stop drinking when you want to? ^a
5	8. Have you ever attended a meeting of Alcoholics Anonymous? ^a
1	9. Have you gotten into fights when drinking?
2	10. Has drinking ever created problems with you and your spouse? ^a
2	11. Has your spouse or other family member ever gone to anyone for help about your drinking?
2	12. Have you ever lost friends or girlfriends/boyfriends because of your drinking?
2	13. Have you ever gotten into trouble at work because of drinking? ^a
2	14. Have you ever lost a job because of drinking?
2	15. Have you ever neglected your obligations, your family, or your work for 2 or more days in a row because you were drinking? ^a
1	16. Do you ever drink before noon?
2	17. Have you ever been told you have liver trouble? Cirrhosis?
2	18. Have you ever had delirium tremens (DTs), severe shaking, heard voices, or seen things that weren't there after heavy drinking?
5	19. Have you ever gone to anyone for help about your drinking? ^a
5	20. Have you ever been in a hospital because of your drinking? ^a
2	21. Have you ever been a patient in a psychiatric hospital or on a psychiatric ward of a general hospital when drinking was part of the problem?
2	22. Have you ever been treated at a psychiatric or mental health clinic or gone to a doctor, social worker, or clergyman for help with an emotional problem in which drinking had played a part?
2	23. Have you ever been arrested, even for a few hours, because of drunk behavior? ^a
2	24. Have you ever been arrested for drunk driving or driving after drinking? ^a

^aIncluded in the short version of the MAST.

ferase) and mean corpuscular volume have been investigated as biological markers of alcohol abuse. All of these tests are insensitive in detecting alcohol abusers. None of these tests, alone or in combination, perform as well as the MAST or the CAGE questionnaire in detecting alcohol abuse.^{19,22,33-35}

RELIABILITY OF THE MAST, CAGE, AND AUDIT QUESTIONNAIRES

Gibbs³⁶ reviewed the internal consistency (α) reliability coefficient of the MAST reported in 6 studies and found it to vary from .83 to .93. The α values in 6 studies of the SMAST or BMAST ranged from .75 to .81. Skinner and Sheu³⁷ reported the test-retest reliability of the MAST at .84. Reliability coeffi-

cients of 1.0 represent perfect test precision (perfect interobserver or intraobserver precision), and values close to 1.0 are highly precise. No reports measuring the reliability of the CAGE and AUDIT questionnaires were identified.

ACCURACY OF THE MAST, CAGE, AND AUDIT QUESTIONNAIRES

Determining the test accuracy of all questionnaires for alcohol use disorders presents some methodologic problems. The questions in the CAGE, MAST, and AUDIT questionnaires are embodied within the commonly used reference standards, *DSM-III-R* and *ICD-10*, which may result in inflated estimates of test accuracy. The advantage of the CAGE and AUDIT questionnaires over the much longer questionnaires is their brevity, which would allow them to be used as a screening or case-finding tool by busy clinicians.

The diagnostic accuracy of the MAST and its shorter versions has been reported, with sensitivities of 71% to 100% and specificities of 81% to 96%.^{8,19,38} The MAST can be criticized as a screening tool because of its length; it requires about 20 minutes to administer, making it less likely to be used by a busy clinician.

In most studies of the diagnostic accuracy of the CAGE questionnaire, a positive test result has been defined as 2 or more affirmative answers to the questions. The CAGE questionnaire has been validated in several environments, including psychiatric inpatients, medical and orthopedic inpatients, and ambulatory medical patients in the United States and Great Britain.^{6,7,18-21,28} Table 4-4 lists studies in which the diagnostic accuracy of the CAGE questionnaire has been reported and in which the authors specify the "diagnostic standard" used to define the patient's alcohol status. In all these studies, changing the criterion of a positive CAGE test result from a score of 2 to 1 results in greater test sensitivity but lower specificity. In other words, the test will identify more problem drinkers, but it will also misclassify more nonproblem patients as problem drinkers. Note that as the definition of problem drinking is lowered, for example, from 16 to 8 drinks per day or from 2 drinks to 1 drink per day in pregnant women, the sensitivity of the test decreases and the specificity increases for the same CAGE threshold.

The CAGE questionnaire is reasonably accurate at identifying those individuals who are alcohol dependent or heavy drinkers (>8 drinks/d). However, it is not at all sensitive at detecting the lower levels of consumption that may be dangerous, especially in pregnant women. It has not been tested as a tool to detect hazardous or at-risk drinking on the order of 4 drinks per day. It will be less sensitive in that situation. There is no difference in the diagnostic accuracy of the CAGE questionnaire when used in men or women, and it is equally effective in elderly people.^{6,39} However, there is a marked difference in the prevalence of alcohol disease in men and women. The prevalence of alcohol dependence in women is about one-third that in men. The predictive values for CAGE responses reflect the lower prevalence figures for women, with lower positive predictive values and higher negative predictive values.^{6,39}

The AUDIT is a newly developed tool, and only 1 validation study was identified. When a positive test result is considered to be a score of 8 or more points, the sensitivity of the AUDIT in detecting hazardous or harmful use is 92% and the specific-

ity is 94%.³² However, as noted above, there are methodologic reasons to believe that these estimates are inflated and may not be reliably testable. The 10 AUDIT questions were culled from a 150-item assessment of alcohol use. The AUDIT has not been

Table 4-3 Alcohol Use Disorders Identification Test (AUDIT) Questions^{25,32}

1. How Often Do You Have a Drink Containing Alcohol?				
Never	Monthly or less	2 to 4 times a month	2 or 3 times a week	4 or more times a week
2. How Many Drinks Containing Alcohol Do You Have on a Typical Day When You Are Drinking?				
1 or 2	3 or 4	5 or 6	7 to 9	10 or more
3. How Often Do You Have 6 or More Drinks on 1 Occasion?				
Never	Less than monthly	Monthly	Weekly	Daily or almost daily
4. How Often During the Last Year Have You Found That You Were Not Able to Stop Drinking Once You Had Started?				
Never	Less than monthly	Monthly	Weekly	Daily or almost daily
5. How Often During the Last Year Have You Failed to Do What Was Expected From You Because of Drinking?				
Never	Less than monthly	Monthly	Weekly	Daily or almost daily
6. How Often During the Last Year Have You Needed a First Drink in the Morning to Get Yourself Going After a Heavy Drinking Session?				
Never	Less than monthly	Monthly	Weekly	Daily or almost daily
7. How Often in the Last Year Have You Had a Feeling of Guilt or Remorse After Drinking?				
Never	Less than monthly	Monthly	Weekly	Daily or almost daily
8. How Often During the Last Year Have Been Unable to Remember What Happened the Night Before Because You Had Been Drinking?				
Never	Less than monthly	Monthly	Weekly	Daily or almost daily
9. Have You or Someone Else Been Injured as a Result of Your Drinking?				
No	Yes, but not in the last year		Yes, during the last year	
10. Has a Relative or Friend or a Doctor or Other Health Worker Been Concerned About Your Drinking or Suggested You Cut Down?				
No	Yes, but not in the last year		Yes, during the last year	

Table 4-4 Diagnostic Standards and Diagnostic Accuracy for the CAGE Questionnaire

Source, y	Type	No.	Diagnostic Standard	Patients			Prevalence of Alcohol Disease, %
				Positive CAGE Result	Sensitivity, %	Specificity, %	
Bernadt et al, ¹⁹ 1982	Psychiatric inpatients	385	Ethyl alcohol intake interview plus ≥ 16 drinks/d or medical record review diagnosis of alcoholism	≥ 2	97	76	17
Buchsbaum et al, ⁶ 1991	Ambulatory medical patients	821	<i>DSM-III-R</i> with SCID	≥ 2	73	91	36
				≥ 1	89	81	
Bush et al, ⁷ 1987	Medical and orthopedic inpatients	521	<i>DSM-III</i> with MAST, NIAAA intake questionnaire	≥ 2	75	96	2
				≥ 1	85	89	
King, ²⁰ 1986	Ambulatory general patients	407	Ethyl alcohol intake interview plus ≥ 8 drinks/d	≥ 2	82	95	4
				≥ 1	0	84	
Mayfield et al, ²⁸ 1974	Veterans Affairs hospital inpatients	366	Multidisciplinary team diagnosis	≥ 2	81	89	39
				≥ 1	90	72	
Sokol et al, ¹⁸ 1989	Prenatal clinic	971	Periconceptual ethyl alcohol intake interview plus ≥ 2 drinks/d	≥ 2	38	92	4
				≥ 1	59	82	
Waterson and Murray-Lyon, ²¹ 1989	Prenatal clinic	893	Periconceptual ethyl alcohol intake interview plus ≥ 2 (top row) vs ≥ 1 (bottom row) drink/d	≥ 2	33	95	2
				≥ 2	20	96	20

Abbreviations: CAGE, cut down, annoyed, guilty, eye opener; *DSM-III*, *Diagnostic and Statistical Manual of Mental Disorders, Third Edition*; *DSM-III-R*, *Diagnostic and Statistical Manual of Mental Disorders, Revised Third Edition*; MAST, Michigan Alcoholism Screening Test; NIAAA, National Institute on Alcohol Abuse and Alcoholism; SCID, Structured Clinical Interview for *DSM-III-R*.

tested as a discrete group of questions against an accepted diagnostic standard. As with the other questionnaires for alcohol disorders, items in the AUDIT are represented in the commonly used reference standards, *DSM-III-R* and *ICD-10*. This likely inflates the estimates of reliability coefficients and test accuracy. The AUDIT attempts to identify drinkers whose consumption places them at risk of harmful or dependent alcohol use before dependence has occurred. Three AUDIT questions relate to amounts and frequency of consumption. There is no reliable way to test the accuracy of patient responses concerning consumption. If heavy drinkers are defensive about their drinking and tend to underreport consumption, the AUDIT estimate of hazardous drinking may be conservative.

PREDICTIVE ACCURACY OF THE CAGE QUESTIONNAIRE

There are 2 ways for clinicians to calculate predictive value or posterior probability of disease.^{40,41} The first approach uses test sensitivity, specificity, and estimates of disease prevalence in Bayes theorem. The second approach multiplies the LR by the pretest odds of disease to obtain the posttest odds of disease. The 2 methods are equivalent when the diagnostic test

Table 4-5 Likelihood Ratios of CAGE Questions for the Diagnosis of Alcohol Abuse and Alcohol Dependence

	Buchsbaum et al, ⁶ 1991	Bush et al, ⁷ 1987	Mayfield et al, ²⁸ 1974
Prevalence of alcohol disease	0.36	0.20	0.39
LR by CAGE score			
0	0.14	0.18	0.13
1	1.5	1.4	0.90
2	4.5	6.8	1.6
3	13	158	15
4	101	∞	∞

Abbreviations: CAGE, cut down, annoyed, guilty, eye opener; LR, likelihood ratio.

Table 4-6 Posterior Probability of Alcohol Abuse or Alcohol Dependence Calculated With Likelihood Ratios^a

CAGE Score	LR	Posterior Probability	
		Prevalence of Alcohol Disease of 10%	Prevalence of Alcohol Disease of 36%
0	0.14	.02	.07
1	1.5	.14	.46
2	4.5	.33	.72
3	13	.59	.88
4	101	.92	.98

Abbreviations: CAGE, cut down, annoyed, guilty, eye opener; LR, likelihood ratio.

^aLRs are based on data from Buchsbaum et al.⁶

used gives dichotomous results. However, if the test results are not dichotomous, and most are not, these 2 methods may give surprisingly different results. The insistence that a given cut point be assigned to continuous data or multiple categorical levels can result in a loss of diagnostic power and even erroneous diagnostic conclusions.

In the introductory article to this series, Sackett⁴² introduced the concept of LRs for diagnostic tests with multiple levels of response. If you are not familiar with LRs, I encourage you to review that article. If one wishes to avoid some of the pitfalls that may occur when interpreting the results of questionnaires, it is important to be able to interpret the results with LRs. Table 4-5 lists 3 studies of the CAGE questionnaire in which LRs can be calculated.^{6,7,28} These studies have low LRs for CAGE scores of 0 (0.13–0.18), high LRs for CAGE scores of 3 (13–158), and very high LRs for CAGE scores of 4 (101 to infinity).

Table 4-6 shows the posterior probability of alcohol abuse or dependence for each CAGE score according to the Buchsbaum et al⁶ data and prevalences of 10% and 36%. Alcohol abuse or dependence is unlikely in persons with a score of 0. With a score of 3, the diagnosis is likely, and a score of 4 is virtually diagnostic of alcohol abuse or dependence in the higher-prevalence group. However, more caution needs to be exercised when interpreting CAGE scores of 1 or 2. The likelihood of alcohol abuse or dependence is increased in persons with scores of 2, but one might want to administer other confirmatory tests before the patient is given a diagnosis. A score of 1 has an LR of 1.5, and the posttest probability of disease is only marginally higher than the pretest probability of disease.

PROBLEMS IN THE IDENTIFICATION OF AT-RISK DRINKING IN PREGNANT WOMEN

Pregnant women who drink 2 or more drinks per day may expose the fetus to an increased risk of developmental delay, growth retardation, cardiac defects, and craniofacial abnormalities.^{18,24} Women drinking enough to expose the fetus to a teratogenic risk may underreport their consumption. This is most pronounced among those women with high MAST scores who are drinking heavily.^{43,44} It has also been shown that the BMAST and CAGE questionnaires are insensitive instruments for identifying pregnant drinkers who consume 2 or more drinks per day.^{18,21} Sokol et al¹⁸ modified the CAGE questionnaire by substituting for the question on “guilt” to one on alcohol tolerance: “How many drinks does it take to make you high?” The patient was considered tolerant if it took more than 2 drinks to make her feel high. The authors claim that this question is not likely to generate defensiveness and denial. This modified questionnaire, T-ACE (tolerance, annoyed, cut down, eye opener), was administered to 1065 women attending an inner-city obstetric clinic. The prevalence of at-risk drinking in this study was judged to be 4.3%. The T-ACE questionnaire was found to be more sensitive than the CAGE questionnaire (76% vs 59%) and equivalent to the MAST in identifying pregnant women drinking more

than 2 drinks per day when the cut point for a positive test result was a score of 1 or higher. Unfortunately, 40% of the women judged to be at-risk drinkers scored 0 on the CAGE questionnaire. Although the T-ACE questionnaire was more sensitive, 25% of at-risk drinkers had a score of 0. In this setting, the specificities of the T-ACE, CAGE, and MAST questionnaires were similar (76%-82%) and the positive predictive values were 13% to 14%.

Given the low sensitivity of these tests, a significant portion of pregnant drinkers will go undetected. The low prevalence of at-risk drinking in this population and the moderate specificity of these tests result in low positive predictive values. Consequently, these questionnaires cannot be expected to reliably identify problem pregnant drinkers.

THE BOTTOM LINE

In summary, the CAGE questionnaire can be a useful tool in the diagnosis of *DSM-III-R*-defined alcohol abuse and dependence and very heavy drinking (>8 drinks/d). A CAGE score of 0 has a good negative predictive value at a lower prevalence of disease. Scores of 3 or 4 strongly support the diagnosis of alcohol abuse. However, scores of 1 or 2 must be interpreted with caution, and one should use the LR approach to accurately interpret these intermediate scores. The CAGE questionnaire has not been tested as a tool for identifying persons who may be engaged in hazardous drinking of lesser amounts of alcohol; for example, 4 drinks per day. It is likely that the test will be insensitive in detecting these individuals. The AUDIT was recently developed to identify these hazardous drinkers. It has not been thoroughly tested, but the initial report suggests that it is reasonably accurate. Because 7 of the AUDIT questions are almost identical to questions in the MAST or CAGE, it should be good at identifying alcohol abuse and alcohol dependence. The other 3 AUDIT questions relate to consumption and constitute an attempt to identify hazardous drinkers. It may not be possible to determine the accuracy of these questions in the absence of a reliable, socially acceptable diagnostic standard for consumption. However, if heavy drinkers are defensive about their levels of consumption, the AUDIT may underestimate levels of consumption. The CAGE questionnaire is short and can be easily memorized. It has been field tested and shown to be a useful tool. The busy clinician could use the CAGE questionnaire to find unrecognized patients who are abusing or dependent on alcohol. The first 3 questions of the AUDIT are also easily memorized and can provide an estimate of the patient's typical alcohol consumption. The busy clinician could use these questions as a form of targeted preventive medicine. Men drinking more than 4 drinks per day and women drinking more than 2 drinks per day should be counseled about the risks of drinking.

Identifying pregnant women engaged in at-risk drinking is problematic. The prevalence of at-risk drinking among pregnant women is low, and the screening questionnaires to identify problem drinkers have relatively low sensitivities. Because none of these instruments is sufficiently reliable to use for case finding

in pregnant women, all pregnant women should be counseled about the risks of drinking while pregnant. Abstinence from alcohol would be the safest option, but women who choose to drink while pregnant should be strongly advised to avoid binge drinking and to drink fewer than 2 drinks per day.

Author Affiliation at the Time of the Original Publication

From the Department of Internal Medicine, University of Toronto, and the Division of General Internal Medicine, St Michael's Hospital, Toronto, Ontario, Canada.

Acknowledgments

The author thanks James Rankin, MD, for his critical review of the manuscript.

REFERENCES

- Eckardt MJ, Harford TC, Kaelber CT, et al. Health hazards associated with alcohol consumption. *JAMA*. 1981;246(6):648-666.
- Klatsky AL, Armstrong MA, Friedman GD. Alcohol and mortality. *Ann Intern Med*. 1992;117(8):646-654.
- West LJ, Cohen S. Provisions for dependency disorders. In: Holland WW, Detels R, Knox G, eds. *Oxford Textbook of Public Health*. New York, NY: Oxford University Press; 1985;vol 2; chap 9:106-114.
- Rankin JG, Ashley MJ. Alcohol-related health problems. In: Last JM, Wallace RB, eds. *Maxcy-Rosenau-Last Public Health and Preventive Medicine*. 13th ed. East Norwalk, CT: Appleton & Lange; 1992:chap 43:741-767.
- Schuckit MA. Genetics and the risk for alcoholism. *JAMA*. 1985;254(18):2614-2617.
- Buchsbaum DG, Buchanan RG, Centor RM, Schnoll SH, Lawton MJ. Screening for alcohol abuse using CAGE scores and likelihood ratios. *Ann Intern Med*. 1991;115(10):774-777.
- Bush B, Shaw S, O'Leary P, Delbanco T, Aronson MD. Screening for alcohol abuse using the CAGE questionnaire. *Am J Med*. 1987;82(2):231-235.
- Cleary PD, Miller M, Bush BT, Warburg MM, Delbanco TL, Aronson MD. Prevalence and recognition of alcohol abuse in a primary care population. *Am J Med*. 1988;85(4):4664-4671.
- Buchsbaum DG, Buchanan RG, Poses RM, Schnoll SH, Lawton MJ. Physician detection of drinking problems in patients attending a general medicine practice. *J Gen Intern Med*. 1992;7(5):517-521.
- Curtis JR, Geller G, Stokes EJ, Levine DM, Moore RD. Characteristics, diagnosis, and treatment of alcoholism in elderly patients. *J Am Geriatr Soc*. 1989;37(4):310-316.
- Marsh CM, Vaughn DA. Do MAST and CAGE scores help detect alcoholism? Presented at: 142nd Annual Meeting of the American Psychiatric Association; May 8, 1989; San Francisco, CA.
- Moore RD, Bone LR, Geller G, Mamon JA, Stokes EJ, Levine DM. Prevalence, detection, and treatment of alcoholism in hospitalized patients. *JAMA*. 1989;261(3):403-407.
- Dawson NV, Dadheech G, Speroff T, Smith RL, Schubert DSP. The effect of patient gender on the prevalence and recognition of alcoholism on a general medical service. *J Gen Intern Med*. 1992;7(1):38-45.
- World Health Organization. Mental and behavioral disorders: diagnostic criteria for research. In: *International Statistical Classification of Diseases, 10th Revision (ICD-10)*. Geneva, Switzerland: World Health Organization; 1992.
- American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders, Revised Third Edition*. Washington, DC: American Psychiatric Association; 1987.
- Rounsaville BJ, Bryant K, Babor T, Kranzler H, Kadden R. Cross system agreement for substance use disorders: *DSM-III-R*, *DSM-IV* and *ICD-10*. *Addiction*. 1993;88(3):337-348.
- Spitzer RL, Williams JBW, Gibbon M, First MB. *Structured Clinical Interview for DSM-III-R: Non-patient Edition (SCID-NP, Version 1.0)*. Washington, DC: American Psychiatric Press; 1990.
- Sokol RJ, Martier SS, Ager JW. The T-ACE questions: practical prenatal detection of risk drinking. *Am J Obstet Gynecol*. 1989;160(4):863-868.

19. Bernadt MW, Mumford J, Taylor C, Smith B, Murray RM. Comparison of questionnaire and laboratory tests in the detection of excessive drinking and alcoholism. *Lancet*. 1982;1(8267):325-328.
20. King M. At risk drinking among general practice attenders: validation of the CAGE questionnaire. *Psychol Med*. 1986;16(1):213-217.
21. Waterson NJ, Murray-Lyon IM. Screening for alcohol related problems in the antenatal clinic: an assessment of different methods. *Alcohol Alcohol*. 1989;24(1):21-30.
22. Skinner HA, Schuller R, Roy J, Israel Y. Identification of alcohol abuse using laboratory tests and a history of trauma. *Ann Intern Med*. 1984;101(6):847-851.
23. Pequingnot G, Tuyns AJ, Berta JL. Ascitic cirrhosis in relation to alcohol consumption. *Int J Epidemiol*. 1978;7(2):113-120.
24. Ernhart CB, Sokol RJ, Martier S, et al. Alcohol teratogenicity in the human: a detailed assessment of specificity, critical period, and threshold. *Am J Obstet Gynecol*. 1987;156(1):33-39.
25. Saunders JB, Aasland OG, Amundsen A, Grant M. Alcohol consumption and related problems among primary health care patients: WHO collaborative project on early detection of persons with harmful alcohol consumption, I. *Addiction*. 1993;88(3):349-362.
26. Bien TH, Miller WR, Tonigan JS. Brief interventions for alcohol problems: a review. *Addiction*. 1993;88(3):315-335.
27. Ewing JA. Detecting alcoholism: the CAGE questionnaire. *JAMA*. 1984; 252(14):1905-1907.
28. Mayfield D, McLeod G, Hall P. The CAGE questionnaire: validation of a new alcohol screening instrument. *Am J Psychiatry*. 1974;131:1121-1123.
29. Selzer ML. The Michigan Alcoholism Screening Test: the quest for a new diagnostic instrument. *Am J Psychiatry*. 1971;127(12):1653-1658.
30. Pokorny AD, Miller BA, Kaplan HB. The brief MAST: a shortened version of the Michigan Alcoholism Screening Test. *Am J Psychiatry*. 1972;129(3): 342-345.
31. Selzer ML, Vinokur A, van Rooijen MA. A self-administered Short Michigan Alcoholism Screening Test (SMAST). *J Stud Alcohol*. 1975;36(1):117-126.
32. Saunders JB, Aasland OG, Babor TF, De La Fuente JR, Grant M. Development of the Alcohol Use Disorders Identification Test (AUDIT): WHO collaborative project on early detection of persons with harmful alcohol consumption, II. *Addiction*. 1993;88(6):791-804.
33. Beresford T, Adduci R, Low D, Goggans F, Hall RCW. A computerized biochemical profile for detection of alcoholism. *Psychosomatics*. 1982;23:713-720.
34. Clark PMS, Kricka LJ. Biochemical tests for alcohol abuse. *Br J Alcohol Alcohol*. 1986;16:11-26.
35. Chick J, Kreitman N, Plant M. Mean cell volume and gamma-glutamyl-transpeptidase as markers of drinking in working men. *Lancet*. 1981;1 (8232):1249-1251.
36. Gibbs E. Validity and reliability of the Michigan Alcoholism Screening Test. *Drug Alcohol Depend*. 1983;12:279-285.
37. Skinner HA, Sheu WJ. Reliability of alcohol use indices. *J Stud Alcohol*. 1982;43(11):1157-1170.
38. Ross HE, Gavin DR, Skinner HA. Diagnostic validity of the MAST and the Alcohol Dependence Scale in the assessment of DSM-III alcohol disorders. *J Stud Alcohol*. 1990;51(6):506-513.
39. Buchsbaum DG, Buchanan RG, Welsh J, Centor RM, Schnoll SH. Screening for drinking disorders in the elderly using the CAGE questionnaire. *J Am Geriatr Soc*. 1992;40(7):662-665.
40. Sackett DL, Haynes RB, Guyatt G, Tugwell P. *Clinical Epidemiology: A Basic Science for Clinical Medicine*. 2nd ed. Boston, MA: Little Brown & Co Inc; 1991.
41. Sox HC, Blatt MA, Higgins MC, Marton KI. *Medical Decision Making*. Stoneham, MA: Butterworths-Heinemann; 1988.
42. Sackett DL. A primer on the precision and accuracy of the clinical examination. *JAMA*. 1992;267(19):2638-2644.
43. Ernhart CB, Morrow-Tlucak M, Sokol RJ, Martier S. Underreporting of alcohol use in pregnancy. *Alcoholism*. 1988;12(4):506-511.
44. Morrow-Tlucak M, Ernhart CB, Sokol RJ, Martier S, Ager J. Underreporting of alcohol use in pregnancy: relationship to alcohol problem history. *Alcoholism*. 1989;13(3):399-401.

UPDATE : Problem Alcohol Drinking

4

Prepared by David L. Simel, MD, MHS

Reviewed by Will Yancy, MD

CLINICAL SCENARIO

A 35-year-old woman requests an appointment for a gynecologic examination. Your nursing staff gives her the usual paperwork and self-administered questionnaires while she waits in the examination room. She fills them all out and gives a response of no to each question. What questions did your patient answer? Do you know how to evaluate her questionnaire? Could she be a problem drinker?

UPDATED SUMMARY ON SCREENING FOR ALCOHOL PROBLEMS

Original Review

Kitchens JM. Does this patient have an alcohol problem? *JAMA*. 1994;272(22):1782-1787.

UPDATED LITERATURE SEARCH

The perceived shortcomings of questionnaires for alcohol use disorders, coupled with the high prevalence of problems, prompted a worldwide effort to improve detection of alcohol use disorders. The US Preventive Services Task Force updated their recommendations (2004) according to new evidence concerning the effectiveness of screening and brief treatment interventions. Our literature search, conducted between 1993 and July 25, 2004, combined the search terms "alcoholism/di" and "alcohol drinking/cl, pc, ep" and the textwords "problem drinking" with "screening." The search was limited to "systematic reviews," and we used the Ovid MEDLINE database, along with the evidence-based medicine databases, to yield 19 English-language articles. We retained articles that were systematic (as opposed to nonsystematic reviews) and that focused on primary care (eg, rather than population-based samples, emergency or psychiatric care). This resulted in 4 articles that we obtained for review. We kept 1 article that had emergency department data to better assess the issues of screening women as opposed to men. We concentrated on

the shorter-form questionnaires that would be more applicable for primary care (see Appendix Tables 4-12, 4-14, 4-15, and 4-16 for the forms AUDIT, CAGE, T-ACE, and TWEAK, respectively). We also retrieved a recent systematic review that was published by the Agency for Health Care Policy and Research as part of an update to the *Guide to Clinical Preventive Services, Third Edition, Periodic Updates* (see <http://www.ahrq.gov/clinic/uspstf/uspsdrin.htm> [accessed May 17, 2008] for the article that first appeared in Whitlock EP, Polen MR, Green CA, Orleans CT, Klein J. Behavioral counseling interventions in primary care to reduce risky/harmful alcohol use by adults: a summary of the evidence for the US Preventive Services Task Force. *Ann Intern Med*. 2004;140(7):558-569). When necessary, we retrieved references from the systematic reviews to verify likelihood ratios (LRs) for reported instruments. After reviewing the retrieved studies and their reference lists, we repeated a literature search using the textwords CAGE, AUDIT, TWEAK, and T-ACE to make sure that we missed no original primary care studies that would have met inclusion criteria.

NEW FINDINGS

It is now abundantly clear that choosing to screen for problem drinking by using any standard approach is overwhelmingly more important than deciding on the screening form! However, once clinicians commit to screening for alcohol problems, there are advantages and disadvantages to the current questionnaires that require understanding (1) what disorder you are screening for and (2) your patient population. Problem drinking is drinking behavior that has not reached the level of abuse or dependence. Studies use various descriptors for problem drinking, including the terms *hazardous, at risk, or harmful* drinking.

The past decade has seen the continued validation of the AUDIT questionnaire, the recognition that screening for alcohol abuse differs from screening for hazardous or problem drinking, and the need for different approaches to screening according to the patient population. Screening women and, possibly, older patients requires different approaches than screening adult men.

DETAILS OF THE UPDATE

IMPROVEMENTS IN THE DATA PRESENTED IN THE ORIGINAL PUBLICATION

The results have not changed, but newer information allows revised estimates of the sensitivity, specificity, and LRs of screening tests for alcohol problems (Table 4-7).

CHANGES IN THE REFERENCE STANDARD

The reference standard for alcohol abuse and dependence remains the guidelines in the *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition*.¹ It is now important for clinicians to understand what constitutes a “drink” and the newer categories of patients’ drinking problems that have not reached the level of abuse or dependence. The definition of a drink changes across cultures, restaurants, and homes. A stan-

dard drink in Great Britain contains about 8 g of alcohol, as opposed to the standard of 19.75 g in Japan.² The US Department of Health and Human Services and the US Department of Agriculture define a standard drink in alcohol and volume content that approximates 12 fl oz of regular beer, 5 fl oz of wine, or 1.5 fl oz of 80-proof distilled spirits.^{2(p7)}

The National Institute on Alcohol Abuse and Alcoholism defines moderate drinking according to the frequency of drinking. Moderate male drinkers ingest 14 or fewer drinks/wk; moderate women drinkers, 7 or fewer drinks/wk; and adults older than 65 years, 7 or fewer drinks/wk.³ Men younger than 65 years would be considered “at risk” drinkers when they drink more than 14 drinks/wk or more than 4 drinks per occasion. Women have drinking problems at lower thresholds: more than 7 drinks/wk or more than 3 drinks per occasion defines “at risk” drinking among women. The World Health Organization uses slightly different descriptors that rely on the consequences of drinking rather than the amount and frequency: “hazardous” drinkers are those who are at risk

Table 4-7 Alcohol Problem Screening Results by Test and Population Profile

Screening ^a Test (n = Number of Studies)	Sensitivity (95% CI)	Specificity (95% CI)	LR+ (95% CI)	LR- (95% CI)
At Risk, Harmful, or Hazardous Drinking				
Adults				
AUDIT-C ≥ 8 (n = 1)	0.40	0.97	12 (5.0-30)	0.62 (0.52-0.74)
AUDIT ≥ 8 (n = 2)	0.57-0.59	0.91-0.96	6.8 (4.7-10)	0.46 (0.38-0.55)
CAGE ≥ 2 (n = 1; all patients > 60 y)	0.14	0.97	4.7 (3.7-6.0)	0.89 (0.86-0.91)
CAGE ≥ 2 (n = 2)	0.49-0.69	0.75-0.95	3.4 (1.2-10)	0.66 (0.54-0.81)
Pregnant Women^b				
TWEAK ≥ 3 (n = 2)	0.67 (0.61-0.73)	0.92 (0.91-0.93)	8.4	0.36
TWEAK ≥ 2 (n = 2)	0.91 (0.87-0.94)	0.77 (0.76-0.78)	4.0	0.12
T-ACE ≥ 1 (n = 3)	0.89 (0.81-0.94)	0.75 (0.70-0.79)	3.6	0.15
CAGE ≥ 2 (n = 3)	0.48 (0.44-0.53)	0.93 (0.92-0.93)	6.9	0.56
CAGE ≥ 1 (n = 3)	0.66 (0.62-0.70)	0.81 (0.81-0.82)	3.5	0.42
Alcohol Abuse or Dependence				
Adults				
CAGE ≥ 2 (n = 10)			6.9 (4.2-11)	0.33 (0.25-0.43)
CAGE ≥ 1 (n = 10)			3.4 (2.3 to 5.1)	0.33 (0.25-0.43)
AUDIT ≥ 8 (n = 2)	0.66-0.71	0.85-0.86	4.6 (3.5-6.1)	0.37 (0.28-0.49)
Women				
CAGE ≥ 2 (n = 2)	0.58 (0.32-0.80)	0.93 (0.90-0.95)	8.3	0.45
CAGE ≥ 1 (n = 1)	0.89 (0.82-0.93)	0.83 (0.79-0.86)	5.2	0.13
≥ 60 y				
CAGE ≥ 2 (n = 3)	0.13-0.82	0.82-0.99	5.2 (3.0-9.0)	0.37 (0.29-0.47)
CAGE ≥ 1 (n = 2)	0.79-0.98	0.56-0.88	2.6 (1.5-4.5)	0.24 (0.15-0.40)
AUDIT ≥ 8 (n = 1)	0.33	0.91	3.6 (1.6-8.0)	0.75 (0.58-0.90)

Abbreviations: AUDIT, Alcohol Use Disorders Identification Test; AUDIT-C, AUDIT Consumption Questions; CAGE, cut down, annoyed, guilty, eye opener; CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio; T-ACE, tolerance, annoyed, cut down, eye opener; TWEAK, tolerance, worry, eye opener, amnesia, cut (kut) down.

^aThe screening questionnaire should be assessed based on the patient population, the threshold that describes positivity, and whether you are screening for “at risk” drinking or dependence.

^bLikelihood ratio estimated from summary sensitivity and specificity measures.

of the adverse consequences of alcohol, whereas “harmful drinking” causes physical or psychological harm that does not yet meet the criteria for abuse.^{3(p1979)}

About 4.6% of US adults abuse alcohol, with men (6.9%) having about 3 times the rate compared to women (2.6%).⁴ An additional 3.8% display alcohol dependence (5.4% of men vs 2.3% of women).

RESULTS OF LITERATURE REVIEW

EVIDENCE FROM GUIDELINES

Canadian Task Force on Preventive Health Care

The Canadian Task Force has not updated their recommendations since 1994,⁵ at a time when the CAGE and the MAST had the best available data. Screening was recommended, although the limitations of these instruments in detecting hazardous drinking were recognized.

Web Resources for Alcohol Screening

A patient-administered screen: <http://www.alcoholscreening.org/> (accessed May 17, 2008). For clinicians: http://pubs.niaaa.nih.gov/publications/Practitioner/pocketguide/pocket_guide.htm (accessed May 17, 2008).

CLINICAL SCENARIO—RESOLUTION

Fortunately, your clinical practice is routinely screening for alcohol problems. However, it is important to know exactly how your patients are being screened. If your clinic is using the CAGE questionnaire, you may detect most patients with alcohol dependence, but you will likely fail to recognize patients who are problem drinkers. This is especially true for women because the sensitivity for all questionnaires is less compared with that for men. In addition to knowing which questionnaire your clinic nurses are using, you need to know how to score the results. Accepting a lower score as “positive” will improve the sensitivity so that you will not miss as many patients with alcohol problems. Because the prevalence of alcohol problems is so high, it is important not to miss these patients.

Assuming your patient drinks some alcohol, the negative LR for alcohol abuse or dependence is 0.18 for adults with at least 1 question positive in the CAGE. The sensitivity is better for the AUDIT, but primary care clinics

might not use the AUDIT because it contains more questions. If you want to detect potentially harmful or hazardous drinking, it would be good to ask the “Tolerance” question from the TWEAK (eg, “How many drinks does it take before you begin to feel the first effects of the alcohol?”). If the patient answers “at least 3,” then you need to assess more fully for problem drinking.

From a practice management standpoint, you and your clinic nurses should review your patient population (**Table 4-8**). If your clinic patients are mostly women, the best current screening forms are the TWEAK or the T-ACE. No data support the existence of 1 ideal questionnaire applicable to all patients, although making no choice of a screening instrument guarantees missed opportunities for intervention. If you are using the CAGE questions, you may choose to switch to the AUDIT (which will detect problem drinking, abuse, and dependence). If the AUDIT is too long for your patients, then you could select the CAGE, TWEAK, or T-ACE and use a low threshold for pursuing follow-up questions. Two alternate approaches combine the best features of the AUDIT (which detects hazardous drinking but is long) with the CAGE (which detects abuse and dependence and is short but does not detect problem drinking). The resulting AUDIT-C is a shorter questionnaire than the AUDIT (see Appendix **Table 4-13**) and, in one study, appears to have the same measurement characteristics as the full AUDIT.

Table 4-8 US Preventive Health Services Task Force Recommendations for Tests in Different Populations

Population	AUDIT	CAGE	TWEAK or T-ACE
Risky or Harmful Drinking			
Adults	Yes	No	No
≥ 65 y	Uncertain	No	No
Pregnant women	No	No	Yes
Alcohol Abuse or Dependence			
Adults	Yes	Yes	Yes
≥ 65 y	Uncertain	Uncertain	Uncertain
Pregnant women	No	No	Yes

Abbreviations: AUDIT, Alcohol Use Disorders Identification Test; CAGE, cut down, annoyed, guilty, eye opener; T-ACE, tolerance, annoyed, cut down, eye opener; TWEAK, tolerance, worry, eye opener, amnesia, cut (kut) down.

SCREENING FOR ALCOHOL PROBLEMS—MAKE THE DIAGNOSIS

PRIOR PROBABILITY

Data from the National Institute on Alcohol Abuse and Alcoholism suggest that 3 of 10 adults engage in risky drinking behaviors. In primary care clinics, the prevalence will be around 11% to 18%.

POPULATIONS FOR WHOM PROBLEM DRINKING SHOULD BE ASSESSED

- All adults (see Tables 4-9 and 4-10)
- Targeted populations/conditions requiring assessment include pregnant women (see Table 4-11), adolescents, and emergency patients

Table 4-9 Detecting the Likelihood of At-risk, Harmful, or Hazardous Drinking in Adults

	LR Range
AUDIT or AUDIT-C \geq 8	6.8-12
AUDIT or AUDIT-C \leq 8	0.46-0.62

Abbreviations: AUDIT, Alcohol Use Disorders Identification Test; AUDIT-C, AUDIT Consumption Questions; LR, likelihood ratio.

Table 4-10 Detecting the Likelihood of Alcohol Abuse or Dependence in Adults^a

	LR (95% CI)
CAGE \geq 1	3.4 (2.3-5.1)
CAGE = 0	0.18 (0.11-0.29)

Abbreviations: CAGE, cut down, annoyed, guilty, eye opener; CI, confidence interval; LR, likelihood ratio.

^aWomen have a lower sensitivity than men do but have a higher specificity. A cut point of \geq 1 optimizes the sensitivity and, therefore, the negative LR.

Table 4-11 Detecting the Likelihood of 2 or More Drinks/Day During Pregnancy^a

	LR Range
TWEAK \geq 2 or T-ACE \geq 1	3.6-4.0
TWEAK \leq 1 or T-ACE = 0	0.12-0.15

Abbreviations: LR, likelihood ratio; T-ACE, tolerance, annoyed, cut down, eye opener; TWEAK, tolerance, worry, eye opener, amnesia, cut (kut) down.

^aLRs are estimated from studies that have incorporation bias where the interviewer knew the results of the screening questionnaires.

REFERENCE STANDARD TESTS

Diagnostic interview schedule for *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition*,² interview performed by an experienced provider in an alcohol-related interview.

REFERENCES FOR THE UPDATE

1. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders (DSM-IV)*. 4th ed. Washington, DC: American Psychiatric Association; 2000.
2. Dufour MC. What is moderate drinking? *Alcohol Res Health*. 1999;23(1):5-14.
3. Fiellen DA, Reid MC, O'Connor PG. Screening for alcohol problems in primary care. *Arch Intern Med*. 2000;160(13):1977-1989.^a
4. Grant BF, Dawson DA, Stinson FS, Chou SP, Dufour MC, Pickering RP. The 12-month prevalence and trends in DSM-IV alcohol abuse and dependence: United States, 1991-1992 and 2001-2002. *Drug Alcohol Depend*. 2004;74(3):223-234.
5. Haggerty JL. Early detection and counseling of problem drinking. In: Canadian Task Force on the Periodic Health Examination. *Canadian Guide to Clinical Preventive Health Care*. Ottawa, Ontario, Canada: Health Canada; 1994:488-498. <http://www.ctfphc.org/sections/section06ch042.htm>. Accessed May 17, 2008.
6. Bradley KA, Boyd-Wickizer J, Powell SH, Burman ML. Alcohol screening questionnaires in women. *JAMA*. 1998;280(2):166-171.^a
7. Whitlock EP, Polen MR, Green CA, Orleans CT, Lein JT. *Behavioral Counseling Interventions in Primary Care to Reduce Risky/Harmful Alcohol Use*. Rockville, MD: Agency for Healthcare Research and Quality; 2004. Systematic Evidence Review No. 30. Electronic copies available at <http://www.ahrq.gov/clinic/3rduspstf/alcohol/alcomissum.pdf> (accessed May 17, 2008).

^aFor the Evidence to Support the Update for this topic, see <http://www.JAMAEvidence.com>.

APPENDIX—ALCOHOL SCREENING INSTRUMENTS^{6,7}

Adapted from Whitlock EP, Polen MR, Green CA, Orleans CT, Lein JT. *Behavioral Counseling Interventions in Primary Care to Reduce Risky/Harmful Alcohol Use*. Rockville, MD:

Agency for Healthcare Research and Quality; 2004. Systematic Evidence Review No. 30. Electronic copies available at <http://www.ahrq.gov/clinic/3rduspstf/alcohol/alcomissum.pdf> (accessed, May 17, 2008).

Table 4-12 AUDIT

Circle the number that comes closest to your alcohol use in the PAST YEAR.

1. How often do you have a drink containing alcohol?				
(0) Never	(1) Monthly or less	(2) 2 to 4 times a month	(3) 2 or 3 times a week	(4) 4 or more times a week
2. How many drinks containing alcohol do you have on a typical day when you are drinking?				
(0) 1 or 2	(1) 3 or 4	(2) 5 or 6	(3) 7 to 9	(4) 10 or more
3. How often do you have 6 or more drinks on 1 occasion?				
(0) Never	(1) Less than monthly	(2) Monthly	(3) Weekly	(4) Daily or almost daily
4. How often during the last year have you found that you were not able to stop drinking once you had started?				
(0) Never	(1) Less than monthly	(2) Monthly	(3) Weekly	(4) Daily or almost daily
5. How often during the last year have you failed to do what was expected from you because of drinking?				
(0) Never	(1) Less than monthly	(2) Monthly	(3) Weekly	(4) Daily or almost daily
6. How often during the last year have you needed a first drink in the morning to get yourself going after a heavy drinking session?				
(0) Never	(1) Less than monthly	(2) Monthly	(3) Weekly	(4) Daily or almost daily
7. How often in the last year have you had a feeling of guilt or remorse after drinking?				
(0) Never	(1) Less than monthly	(2) Monthly	(3) Weekly	(4) Daily or almost daily
8. How often during the last year have been unable to remember what happened the night before because you had been drinking?				
(0) Never	(1) Less than monthly	(2) Monthly	(3) Weekly	(4) Daily or almost daily
9. Have you or someone else been injured as a result of your drinking?				
(0) No	(1) Yes, but not in the last year		(2) Yes, during the last year	
10. Has a relative or friend or a doctor or other health worker been concerned about your drinking or suggested you cut down?				
(0) No	(1) Yes, but not in the last year		(2) Yes, during the last year	

Abbreviation: AUDIT, Alcohol Use Disorders Identification Test.

Scoring: A score of 8 or more is considered a positive screen for hazardous or harmful drinking.

Table 4-13 AUDIT-C

Circle the number that comes closest to your alcohol use in the PAST YEAR.

1. How often do you have a drink containing alcohol? Consider a “drink” to be 1 can or bottle of beer, 1 glass of wine, 1 wine cooler, 1 cocktail, or 1 shot of hard liquor (like scotch, gin, or vodka).				
(0) Never	(1) Monthly or less	(2) 2 to 4 times a month	(3) 2 to 3 times a week	(4) 4 or more times a week
2. How many drinks containing alcohol do you have on a typical day when you are drinking?				
(0) 1 or 2	(1) 3 or 4	(2) 5 or 6	(3) 7 to 9	(4) 10 or more
3. How often do you have 6 or more drinks on 1 occasion?				
(0) Never	(1) Less than monthly	(2) Monthly	(3) Weekly	(4) Daily or almost daily

Abbreviation: AUDIT-C, Alcohol Use Disorders Identification Test Consumption Questions.

Scoring: A score of 8 or more is considered a positive screen for hazardous or harmful drinking.

Table 4-14 CAGE

1. Have you ever felt you should cut down on your drinking?
2. Have people annoyed you by criticizing your drinking?
3. Have you ever felt bad or guilty about your drinking?
4. Have you ever had a drink first think in the morning to steady your nerves or to get rid of a hangover (*eye opener*)?

Abbreviation: CAGE, *cut down*, *annoyed*, *guilty*, *eye opener*.

Scoring: Two or more positive responses are considered a positive screen for problem drinking in most studies. Alternatively, you may select a cut point of just 1 positive response to improve the sensitivity.

Table 4-15 T-ACE

1. How many drinks does it take to make you feel high (*tolerance*)?
2. Have people annoyed you by criticizing your drinking?
3. Have you ever felt you should cut down on your drinking?
4. Have you ever had a drink first think in the morning to steady your nerves or to get rid of a hangover (*eye opener*)?

Abbreviation: T-ACE, *tolerance*, *annoyed*, *cut down*, *eye opener*.

Scoring: Positive response to the tolerance item (positive is considered more than 2 drinks) is scored 2 points; to other items, 1 point each. Total of 2 or more indicates risky drinking.

Table 4-16 TWEAK

1. How many drinks can you hold? ("Hold" version; ≥ 6 drinks indicates tolerance) *or* how many drinks does it take before you begin to feel the first effects of the alcohol? ("High" version; ≥ 3 indicates tolerance)?
2. Does your spouse (or do your parents) ever worry or complain about your drinking?
3. Have you ever had a drink first think in the morning to steady your nerves or to get rid of a hangover (*eye opener*)?
4. Have you ever awakened the morning after some drinking the night before and found that you could not remember a part of the evening before? (*amnesia*)
5. Have you ever felt you ought to cut (*kut*) down on your drinking?

Abbreviation: TWEAK, *tolerance*, *worry*, *eye opener*, *amnesia*, *cut (kut) down*.

Scoring: Positive responses to the tolerance or worry items score 2 points each; to other items, score 1 point each. A total score of 3 or more is considered positive for heavy/problem drinking. During pregnancy, it may be more appropriate to consider a score of 2 or more as positive.

EVIDENCE TO SUPPORT THE UPDATE:

Problem Alcohol Drinking

TITLE The Value of the CAGE in Screening for Alcohol Abuse and Alcohol Dependence in General Clinical Populations: A Diagnostic Meta-analysis.

AUTHORS Aertgeerts B, Buntinx F, Kester A.

CITATION *J Clin Epidemiol.* 2004;57(1):30-39.

QUESTION How well does the CAGE questionnaire (cut down, annoyed, guilty, eye opener) perform?

DESIGN A formal systematic review with meta-analytic techniques.

DATA SOURCES MEDLINE database and MEDION database for diagnostic reviews.

STUDY SELECTION AND ASSESSMENT A search for articles published from January 1974 to December 2001 was conducted, along with a manual search of Dutch-language articles. All languages (except Japanese) were included in the search. Studies had to be in a general clinical population and to report the data required for sensitivity and specificity. Studies with verification bias were eliminated, although studies that adjusted for verification bias were retained. Studies outside of general medical practices (eg, psychiatric settings or the emergency department) were excluded.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

CAGE questionnaire as compared with the diagnosis established by the *Diagnostic and Statistical Manual of Mental Disorders* criteria.

OUTCOME MEASURES

Sensitivity, specificity, and likelihood ratios (LRs) of the CAGE for diagnosis of alcohol abuse or dependence.

MAIN RESULTS

Thirty-five articles were identified, but only 10 were in compliance with all the inclusion and exclusion criteria (Table 4-17).

Table 4-17 Serial LR for the CAGE Questionnaire at Each Cut Point for Patients From Either Outpatient or Inpatient Settings

CAGE threshold	LR+ (95% CI)
CAGE = 4	25 (15-43)
CAGE ≥ 3	15 (8.2-29)
CAGE ≥ 2	6.9 (4.2-11)
CAGE ≥ 1	3.4 (2.3-5.1)
CAGE = 0	0.18 (0.11-0.29)

Abbreviations: CAGE, cut down, annoyed, guilty, eye opener; CI, confidence interval; LR, likelihood ratio.

When comparing primary care patients to ambulatory medical patients (excluding inpatients), the results for the LRs among these groups are clinically similar. While inpatients have positive LRs (confidence intervals [CIs]) that overlap at each threshold, the results for the negative LRs differ. The CAGE has much better sensitivity for inpatients, especially at lower thresholds: When patients have no more than 1 positive response on the CAGE, the LR is 0.17 (CI, 0.11-0.28), and when they answer all the questions negatively, the LR is 0.02 (CI, 0-0.11).

The authors conclude that the CAGE at a cut point of 2 or greater is of limited value.

CONCLUSION

LEVEL OF EVIDENCE Systematic review.

STRENGTHS High-quality systematic review with appropriate meta-analytic techniques. The study formulates the research question, includes a comprehensive search and selection of studies, critically appraises the studies and provides the results, and incorporates the results into their interpretation.

LIMITATIONS Users of the CAGE should be careful not to extrapolate these data to the diagnosis of hazardous or problem drinking because the studies evaluated alcohol abuse or dependence.

We see these data as suggesting that the CAGE is more useful than do the authors. However, it is very important to recognize that the CAGE, with its recommended cut point of CAGE of 2 or greater, is intended to diagnose alcohol abuse or dependence

and not lower levels of problem drinking. The CAGE is useful for this because getting an affirmative answer greatly increases the probability that the person has a problem. On the other hand, we agree with the authors that questionnaires with 0 to 1 positive responses do not sufficiently rule out abuse or dependence, especially in populations with higher prevalence of abuse or dependence.

What about accepting a threshold of only 1 positive response? Further studies are needed, but this would be a reasonable approach for screening. It should be noted that many patients who answer with only 1 positive question will not have an abuse or dependence problem, but it is likely that the sensitivity for such a question would be much higher for problem drinking and you would “miss” fewer patients. For many clinic populations, the LR of 0.18 when the patient answers in the negative for all CAGE questions may not be adequate. This has led many clinics to use a combination of the Alcohol Use Disorders Identification Test (AUDIT; for diagnosing problem drinking) and CAGE (for diagnosing abuse or dependence).

Reviewed by David L. Simel, MD, MHS

TITLE Screening for Alcohol Abuse and Dependence in Older People by Using *DSM* Criteria: A Review.

AUTHORS Beullens J, Aertgeerts B.

CITATION *Aging Ment Health.* 2004;8(1):76-82.

QUESTION Which alcohol screening questionnaires perform best in older patients?

DESIGN Formal systematic review without meta-analytic techniques.

DATA SOURCES MEDLINE and PsycINFO databases.

STUDY SELECTION AND ASSESSMENT Studies published from 1996 to 2002. Studies could be inpatient, outpatient, or nursing home settings for patients 60 years or older. One study of nursing home patients that included those as young as 50 years was included.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

CAGE (cut down, annoyed, guilty, eye opener), AUDIT (Alcohol Use Disorders Identification Test), MAST (Michigan Alcoholism Screening Test), and variations compared with the diagnosis established by the *Diagnostic and Statistical Manual of Mental Disorders (DSM)* criteria. We assessed the data for the CAGE and the AUDIT because these are shorter questionnaires than the longer MAST (see Appendix in the Update for the actual questionnaires).

Table 4-18 Performance of the CAGE Questionnaire Among Older Patients

Test (No. of Studies)	Sensitivity	Specificity	LR+ (95% CI)	LR- (95% CI)
CAGE ≥ 2 (n = 2)	0.63-0.70	0.82-0.91	5.3 (3.0-9.0)	0.37 (0.29-0.47)
CAGE ≥ 1 (n = 2)	0.79-0.86	0.56-0.78	2.6 (1.5-4.5)	0.24 (0.15-0.40)
AUDIT ≥ 8 (n = 1)	0.33	0.91	3.6 (1.6-8.0)	0.75 (0.58-0.90)

Abbreviations: AUDIT, Alcohol Use Disorders Identification Test; CAGE, cut down, annoyed, guilty, eye opener; CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

OUTCOME MEASURES

Sensitivity and specificity. The criterion standard assessed for alcohol abuse or dependence. We retrieved articles to calculate the LRs from the original data.

MAIN RESULTS

Seven articles were identified for inclusion; only 2 were done in the outpatient setting, and the results are displayed in Table 4-18.

CONCLUSION

LEVEL OF EVIDENCE Systematic review.

STRENGTHS The study formulates the research question, includes a comprehensive search and selection of studies, and provides the results.

LIMITATIONS There is no meta-analytic assessment. A formal quality assessment is not presented. Confidence intervals and sample sizes for the number of patients with alcohol abuse or dependence are not given.

The number of studies on drinking problems in older patients is disappointingly low. The authors provide a good rationale for why the existing questionnaires might not work as well in older patients. The authors’ impression is that the CAGE may be better for detecting alcohol abuse or dependence in older patients, which would be consistent with other studies about the use of the CAGE, but it is hard to be conclusive given the paucity of studies in ambulatory older patients. As in other studies, picking a threshold of just 1 or more positive answer to CAGE questions improves the sensitivity. The authors hypothesize that the T-ACE (tolerance, annoyed, cut down, eye opener) might be even more efficient than the CAGE because the “feeling guilty” question is replaced by a “tolerance” question that may be more appropriate for older patients. That hypothesis, along with assessing the proper threshold, needs assessment. The authors do not address the detection of harmful or hazardous drinking in older patients.

Reviewed by David L. Simel, MD, MHS

TITLE Alcohol Screening Questionnaires in Women.

AUTHORS Bradley KA, Boyd-Wickizer J, Powell SH, Burman ML.

CITATION *JAMA*. 1998;280(2):166-171.

QUESTION Which alcohol screening questionnaires perform best in women?

DESIGN Formal systematic review without meta-analytic techniques.

DATA SOURCES MEDLINE database and Social Science and Science Citations Index.

STUDY SELECTION AND ASSESSMENT Studies published from 1996 to July 1997 in English. Studies did not have to be performed in a general clinical population but did need to include a clinic population of women with the data reported separately for women. United States studies were the only studies included. All studies had to compare a brief screening questionnaire to a criterion standard.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

CAGE (cut down, annoyed, guilty, eye opener), AUDIT (Alcohol Use Disorders Identification Test), TWEAK (tolerance, worry, eye opener, amnesia, cut [kut] down), Brief Michigan Alcohol Screening Test (BMAST), T-ACE (tolerance, annoyed, cut down, eye opener), Trauma score, and NET* questionnaires¹ as compared with the diagnosis established by the *Diagnostic and Statistical Manual of Mental Disorders* criteria. In the obstetrics clinic studies, the criterion standard was the number of drinks per day, which is appropriate, given that any drinking may be harmful. In the primary care clinics studies, the criterion was alcohol abuse or dependence. We assessed the data only for the CAGE, AUDIT, TWEAK, and T-ACE for this review as these were the surveys studied in more than 1 location. In all studies, a person who was aware of the questionnaire results applied the criterion standard.

OUTCOME MEASURES

Sensitivity, specificity, and area under the receiver operating characteristic curve. Data were presented for women compared with men when the results were available.

MAIN RESULTS

Thirty-six articles were identified, but only 13 met all the inclusion criteria.

*NET stands for: N, Normal drinker: Do you feel you are a normal drinker?; E, "Eye opener" question from CAGE.; T, Tolerance: How many drinks does it take to make you feel high. These questions are found in the other questionnaires.

Table 4-19 Performance Characteristics of Screening Questionnaires in Women

Setting (No. of Studies, No. of Patients)	Test	Sensitivity (95% CI)	Specificity (95% CI)
Emergency Care (3 studies, 892 patients)			
Low cut point	TWEAK ≥ 2 (Hold version; only in 1 study)	0.87 (0.74-0.93)	0.87 (0.83-0.90)
Higher cut point	CAGE ≥ 2, AUDIT ≥ 8, TWEAK ≥ 3 (Hold version)	0.72 (0.66-0.77)	0.94 (0.92-0.96) ^a
Obstetrics Clinic (3 studies, 8431 patients)			
Low cut point	T-ACE ≥ 1	0.89 (0.81-0.94) ^a	0.75 (0.70-0.79) ^a
	CAGE ≥ 1	0.66 (0.62-0.70)	0.81 (0.81-0.82)
	TWEAK ≥ 2 (Hold version; only 1 study)	0.91 (0.87-0.94)	0.77 (0.76-0.78)
Higher cut point	T-ACE ≥ 2	0.79 (0.64-0.90) ^a	0.82 (0.71-0.90) ^a
	CAGE ≥ 2	0.48 (0.44-0.53)	0.93 (0.92-0.93)
	TWEAK ≥ 3 (Hold version; only 1 study)	0.67 (0.61-0.73)	0.92 (0.91-0.93)
Primary Care (2 studies, 758 patients)			
Low cut point	CAGE ≥ 1 (only 1 study)	0.89 (0.82-0.93)	0.83 (0.79-0.86)
Higher cut point	CAGE ≥ 2	0.58 (0.32-0.80) ^a	0.93 (0.90-0.95)

Abbreviations: AUDIT, Alcohol Use Disorders Identification Test; CAGE, cut down, annoyed, guilty, eye opener; CI, confidence interval; T-ACE, tolerance, annoyed, cut down, eye opener; TWEAK, tolerance, worry, eye opener, amnesia, cut (kut) down.

^aHeterogeneous, $P < .05$.

We extracted the data for sensitivity and specificity to assess for summary values (Table 4-19). The results are the random effects summary measures when there is more than 1 study.

We combined data for the sensitivity and specificity estimates by extracting the raw results. Because of concerns about incorporation bias, we assessed for heterogeneity. We chose not to report summary likelihood measures for women because of our uncertainty about the effect of incorporation bias.

The summary specificity for the CAGE of 2 or greater, AUDIT, TWEAK of 3 or greater, and T-ACE of 2 or greater is 0.92 (95% CI, 0.90-0.94), has narrow CIs, and suggests that a positive questionnaire at these thresholds is clinically similar no matter what population of women is included.

There is greater variability for the sensitivity. The CAGE questionnaire performs poorly in an obstetrics clinic. The AUDIT and the TWEAK of 3 or greater (hold version) have similar sensitivities across all settings (0.69 [95% CI, 0.64-0.74]). For every questionnaire studied (CAGE, AUDIT, and TWEAK), the sensitivity is always worse in women compared with men, whereas the specificity is always higher for women.

CONCLUSION

LEVEL OF EVIDENCE Systematic review.

STRENGTHS High-quality systematic review. The study formulates the research question, includes a comprehensive search and selection of studies, critically appraises the studies and provides the results, and incorporates the results into their interpretation.

LIMITATIONS There is no meta-analytic assessment. The studies in the emergency department and in primary care assessed only for abuse or dependence rather than for harmful or hazardous drinking. Each study was potentially affected by incorporation bias in which the interviewer knew the results of the screening questionnaires. We are uncertain how this affected the interpretation of the criterion standard. However, because all studies were affected by this bias, we still may make inferences on the relative value of the sensitivity and specificity.

No matter what the setting, the specificity of these tests is similarly high for women. Although it is possible that this uniformly good measurement property is a function of incorporation bias, it is also plausible that women with any positive screen result for alcohol are highly likely to be problem drinkers.

The results from the individual studies cited by these authors suggest poorer overall performance for the CAGE among women. Compared with the overall data in the meta-analysis by Aertgeerts et al,² the estimated positive likelihood ratio (LR) for women with a CAGE of 2 or greater appears to be the same (an estimated positive LR of 8.2 in women vs the meta-analytic summary estimate of 6.9 by Aertgeerts et al²), but the estimated LR of 0.45 does appear worse (summary positive LR 0.33 [95% CI, 0.25-0.43]). A study published just after this systematic review also suggested CAGE differences between men and women, along with differences based on race or country of origin.³ In that study, the sensitivity of the CAGE for white women and black women fell within the CI of that in the systematic review by Aertgeerts et al² but was less for Hispanic women. The AUDIT had a better sensitivity among all 3 groups of women studied.

The TWEAK and T-ACE were developed to detect alcohol problems during pregnancy, so they ought to work better than the CAGE for pregnant women. However, the TWEAK and T-ACE have not been as widely studied in primary care clinics.

The authors conclude that the TWEAK and AUDIT may be the best screening tests for women in any setting. They recommend a cut point of 2 or greater for the TWEAK, which does improve the sensitivity but was reported in only 1 study. Although the specificity is worse for the TWEAK of 2 or greater, this is not as an important an issue as failing to diagnose alcohol misuse during pregnancy. Dropping the cut point for the CAGE to 1 or greater improves the sensitivity, but it still does not perform as well as the TWEAK.

Our assessment is that the TWEAK does have statistically similar sensitivity to the AUDIT, with a narrow CI, and these appear to perform better than the CAGE. The TWEAK has the obvious advantage over the AUDIT in that it requires fewer questions. The “hold” version of the TWEAK has been studied more extensively than the “high” version (see Appendix in the Update for the actual questionnaires), but in the single study that compared them, the results were similar. Because many

women may never have passed out from alcohol, the authors recommend using the high version of the TWEAK with the question, “How many drinks does it take before you begin to feel the first effects of the alcohol?” (≥ 3 drinks indicates tolerance). They also recommend a cut point of 2 or greater as indicating positivity. They suggest this lower threshold because the improved sensitivity, especially for pregnant women, would be more important than a higher specificity.

The T-ACE should be studied further because it has fewer questions. It may be easier for primary care clinics to implement it because it is similar to the CAGE except that the “Feeling guilty” question is replaced by the “Tolerance” question.

Reviewed by David L. Simel, MD, MHS

REFERENCES FOR THE EVIDENCE

- Russell M, Martier SS, Sokol RJ, et al. Screening for pregnancy risk-drinking. *Alcohol Clin Exp Res.* 1994;18(5):1156-1161.
- Aertgeerts B, Buntinx F, Kester A. The value of the CAGE in screening for alcohol abuse and alcohol dependence in general clinical populations: a diagnostic meta-analysis. *J Clin Epidemiol.* 2004;57(1):30-39.
- Steinbauer JR, Cantor SB, Holzer CE, Volk RJ. Ethnic and sex bias in primary care screening tests for alcohol use disorders. *Ann Intern Med.* 1998;129(5):353-362.

TITLE Screening for Alcohol Problems in Primary Care.

AUTHORS Fiellin DA, Reid MC, O'Connor PG.

CITATION *Arch Intern Med.* 2000;160(13):1977-1989.

QUESTION Which alcohol screening questionnaires perform best in primary care patients?

DESIGN Formal systematic review.

DATA SOURCES MEDLINE database.

STUDY SELECTION AND ASSESSMENT Studies published in 1996-1998, English language, primary care setting, comparing a screening questionnaire to a criterion standard and including the sensitivity, specificity, or likelihood ratios (LRs). An assessment for evaluation bias or incorporation bias whereby the results of the screening test were used in the criterion standard and an analysis of clinical subgroups was done for each article.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

AUDIT (Alcohol Use Disorders Identification Test), CAGE (cut down, annoyed, guilty, eye opener), and SMAST (Short Michigan Alcoholism Screening Test) instruments for screening for alcohol problems compared with the *Diagnostic and Statistical Manual of Mental Disorders* as the criterion standard.

OUTCOME MEASURES

Adherence to quality standards of reporting the demographics, comorbidities, eligibility criteria and participation rate, criterion standard, blinding, and analysis of subgroups was presented for 38 studies. Sensitivity and specificity were presented without their confidence intervals (CIs). Meta-analytic techniques were not used.

MAIN RESULTS

Eleven articles assessed at-risk, hazardous, or harmful drinking whereas 27 articles studied alcohol dependence or abuse. The result for the SMAST was found in only 1 retrieved study. Table 4-20 includes the data only from studies that met standards for avoiding evaluation and incorporation bias. The sensitivity and specificity are the point estimates (single study) or ranges reported in the review. We retrieved the original articles to obtain the data for combining the results to get a summary LR for the AUDIT. We calculated the summary LR CIs for the AUDIT and AUDIT-C (AUDIT Consumption Questions) from the original data. (For alcohol abuse or dependence, a separate systematic review with a meta-analysis was used to combine the results.¹ The sensitivity and specificity values of the studies without verification bias cited in the publication are shown for comparison purposes to the AUDIT.)

CONCLUSION

LEVEL OF EVIDENCE Systematic review.

STRENGTHS This is an excellent systematic review that formulates the research question, includes a comprehensive

Table 4-20 Performance Characteristics of Screening Questionnaires in Primary Care

	Sensitivity	Specificity	LR+ (95% CI)	LR- (95% CI)
At-Risk, Harmful, or Hazardous Drinking				
AUDIT $\geq 8^{2,3}$	0.57-0.59	0.91-0.96	6.8 (4.7-10)	0.46 (0.38-0.55)
AUDIT-C $\geq 8^{2(p1974)}$	0.40	0.97	12 (5.0-30)	0.62 (0.52-0.74)
CAGE $\geq 2^{4,3(p385)}$	0.49-0.69	0.75-0.95	3.4 (1.2-10)	0.66 (0.54-0.81)
CAGE $\geq 2^4$ (patients all >60 y)	0.14	0.97	4.7 (3.7-6.0)	0.89 (0.86-0.91)
Current Abuse/Dependence				
AUDIT $\geq 8^{2(p1974),3(p385)}$	0.66-0.71	0.85-0.86	4.6 (3.5-6.1)	0.37 (0.28-0.49)
AUDIT-C $\geq 8^{2(p1974)}$	0.46	0.92	5.9 (3.3-10)	0.58 (0.44-0.73)
CAGE $\geq 2^{3(p385)}$	0.77	0.79		
Lifetime Abuse Dependence				
AUDIT $\geq 8^{3(p385)}$	0.39	0.89	7.0	0.46 (0.36-0.58)
CAGE $\geq 2^{3(p385),6}$	0.43-0.53	0.86		

Abbreviations: AUDIT, Alcohol Use Disorders Identification Test; AUDIT-C, AUDIT Consumption Questions; CAGE, cut down, annoyed, guilty, eye opener; CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

^aSource for sensitivity: Aithal et al.⁵

search and selection of studies, critically appraises the studies and provides the results, and incorporates the results into their interpretation.

LIMITATIONS There is no meta-analytic assessment of the AUDIT and CAGE. This makes the results a bit harder for the clinician to detect differences in the performance characteristics of these questionnaires.

The authors evaluated the sensitivity and specificity ranges to conclude that the AUDIT is best at identifying at-risk, hazardous, or harmful drinking. We retrieved the original reports to calculate the LRs. The CAGE appears inferior to the AUDIT for detecting at-risk, harmful, or hazardous drinking. However, a pragmatic problem occurs with the AUDIT in that it is much longer than the CAGE (10 questions vs 4). We retrieved the data from the AUDIT-C, which is a shorter version of the AUDIT, and it compares favorably to the AUDIT for diagnosing hazardous drinking, although it may not be as good for ruling out the problem. Because a subsequent systematic review performed a meta-analysis of the CAGE, we did not use this study to combine those data.

Reviewed by David L. Simel, MD, MHS

REFERENCES FOR THE EVIDENCE

1. Aertgeerts B, Buntinx F, Kester A. The value of the CAGE in screening for alcohol abuse and alcohol dependence in general clinical populations: a diagnostic meta-analysis. *J Clin Epidemiol*. 2004;57(1):30-39.
2. Bush K, Kivlahan DR, McDonell MB, Fish SD, Bradley KA. The AUDIT alcohol consumption questionnaires (AUDIT-C): an effective brief screening test for problem drinking. *Arch Intern Med*. 1998;158(16):1789-1795.
3. Bradley KA, Bush KR, McDonnell MB, Malone T, Fihn SD. Screening for problem drinking, comparison of CAGE and AUDIT. *J Gen Intern Med*. 1998;13(6):379-388.
4. Adams WL, Barry KL, Fleming MF. Screening for problem drinking in older primary care patients. *JAMA*. 1996;276(24):1964-1967.
5. Aithal GP, Thorne H, Dwarkanath AD, Tanner AR. Measurement of carbohydrate deficient transferrin (CDT) in a general medical clinic: is this test useful in assessing alcohol consumption? *Alcohol Alcohol*. 1998;33(3):304-309.
6. Fleming MF, Barry KL. A three-sample test of a masked alcohol screening questionnaire. *Alcohol Alcohol*. 1991;26(1):81-91.

TITLE Behavioral Counseling Interventions in Primary Care to Reduce Risky/Harmful Alcohol Use.

AUTHORS Whitlock EP, Green CA, Polen MR.

CITATION Contract No. 290-92-0018, Task No. 2, Technical support of the US Preventive Services Task Force, March 2004. <http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=hstat3.chapter.45217>. Accessed May 17, 2008.

QUESTION Which screening questionnaires for risky alcohol use among primary care patients identify those who might benefit from brief interventions?

DESIGN Formal systematic review without meta-analytic techniques.

DATA SOURCES MEDLINE, Cochrane, PsychInfo, HealthSTAR, and CINAHL databases.

STUDY SELECTION AND ASSESSMENT The goal was to identify new literature since the last US Preventive Services Task Force recommendations¹; thus, articles were sought from 1994 through April 2002. An extensive search was conducted to identify all relevant articles. Studies had to have been conducted in primary care settings (emergency care and inpatient studies were excluded). The study quality for all included and excluded articles is included. In addition to reviewing primary data, the authors reviewed other systematic reviews.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

The focus of this review was on brief treatment interventions for problem drinkers. The shorter questionnaires were used in the studies that were included: CAGE (*cut down, annoyed, guilty, eye opener*), AUDIT (Alcohol Use Disorders Identification Test), TWEAK (*tolerance, worry, eye opener, amnesia, cut [kut] down*), and T-ACE (*tolerance, annoyed, cut down, eye opener*).

OUTCOME MEASURES

Screening yield, sensitivity, and specificity.

MAIN RESULTS

Twelve studies were included in the review for assessing screening of primary care patients who might be enrolled in brief treatment intervention (Table 4-21).

The initial yield of screening primary care patients for all levels of drinking who are waiting for appointments is 11% to 18%. After further questioning, about 7% of primary care patients are candidates for brief treatment interventions. In trying to identify all patients with drinking disorders, the higher value of 11% to 18% would be the appropriate prevalence for adult US patients.

Table 4-21 Screening Questionnaires for Risky Alcohol Use Should Be Selected According to the Patient Population

Population	AUDIT	CAGE	TWEAK or T-ACE
Risky or Harmful Drinking			
Adults	Yes	No	No
≥65 y	Uncertain	No	No
Pregnant women	No	No	Yes
Alcohol Abuse or Dependence			
Adults	Yes	Yes	Yes
≥65 y	Uncertain	Uncertain	Uncertain
Pregnant women	No	No	Yes

Abbreviations: AUDIT, Alcohol Use Disorders Identification Test; CAGE, cut down, annoyed, guilty, eye opener; T-ACE, tolerance, annoyed, cut down, eye opener; TWEAK, tolerance, worry, eye opener, amnesia, cut (kut) down.

CONCLUSION

LEVEL OF EVIDENCE Systematic review.

STRENGTHS This is an excellent systematic review that formulates the research question, includes a comprehensive search and selection of studies, critically appraises the studies and provides the results, and incorporates the results into their interpretation.

LIMITATIONS There is no meta-analytic assessment. Confidence intervals and LRs are not presented. The studies included in this review were selected because they included randomized trials of patients suitable for brief interventions for problem drinking. Thus, these were not specifically studies of the diagnostic tests themselves. To determine the performance characteristics of screening tests, the authors also used published systematic reviews of the questionnaires.

According to data from systematic reviews of diagnostic tests, these authors conclude that the AUDIT is the best test for detecting risky harmful drinking in adults, although the TWEAK or T-ACE ought to be used for pregnant patients. For detecting alcohol abuse or dependence, they conclude that any of the 4 questionnaires is suitable other than during pregnancy.

The CAGE questionnaire is in widespread use, so the authors suggest that it might be improved by adding quantity/frequency questions. This has shown greater sensitivity and specificity in the emergency department but has not been studied in primary care.² It is available online as part of the National Institute on Alcoholism and Alcohol Abuse guide to physicians (http://pubs.niaaa.nih.gov/publications/Practitioner/pocketguide/pocket_guide.htm, accessed May 17, 2008) and also as a self-graded patient form (<http://www.alcoholscreening.org/>, accessed May 17, 2008).

Reviewed by David L. Simel, MD, MHS

REFERENCES FOR THE EVIDENCE

- US Preventive Services Task Force. *Guide to Clinical Preventive Services*. 2nd ed. Baltimore, MD: Lippincott Williams & Wilkins; 1996.
- Friedmann PD, Saitz R, Gogineni A, Zhang JX, Stein MD. Validation of the screening strategy in the NIAAA "Physicians guide to helping patients with alcohol problems." *J Studies Alcohol*. 2001;62(2):234-238.

CHAPTER

5

Does This Adult Patient Have Appendicitis?

James M. Wagner, MD

W. Paul McKinney, MD

John L. Carpenter, MD

CLINICAL SCENARIO

A 29-year-old patient presents to your office with abdominal pain and a fever. The patient was well until 1 day ago and had never experienced abdominal pain. A vague periumbilical pain awoke him from sleep 12 hours previously, and he soon developed anorexia, nausea, and vomiting. His wife consulted their family medical reference guide and then brought him to the office, concerned that his symptoms matched a description of appendicitis. The pain then migrated to the right lower quadrant (RLQ) and was much worse while he was riding in the car to the physician's office.

The patient's oral temperature is 37.8°C; the pulse rate and blood pressure are normal. He has RLQ tenderness, guarding but not rigidity, and rebound tenderness in the RLQ. A rectal examination reveals no tenderness, and he does not exhibit the psoas or obturator signs. Rovsing sign is positive.

WHY IS THIS AN IMPORTANT QUESTION TO ANSWER WITH A CLINICAL EXAMINATION?

In western countries, appendicitis represents a common cause of acute abdominal pain. According to National Center for Health Statistics data, approximately 500 000 patients underwent appendectomies from 1979 to 1984. Individuals carry a 7% lifetime risk of developing appendicitis.¹ The incidence of appendicitis causing abdominal pain depends on the clinical setting. In series from emergency departments or surgical services, 25% of patients younger than 60 years and evaluated for acute abdominal pain have acute appendicitis, whereas the incidence in those older than 60 years is approximately 4%.¹⁻⁵ Only 0.7% to 1.6% of all ambulatory patients with abdominal pain have appendicitis.^{6,7} Among children treated in the ambulatory care setting, appendicitis causes 2.3% of all abdominal pain episodes.⁸ In children admitted for acute abdominal pain, appendicitis is the etiology for approximately 32%.⁹⁻¹¹

The morbidity and mortality of appendicitis remain significant, even with the advent of antibiotics and effective surgical management. Although the overall mortality rate with appropriate treatment is less than 1%, in the elderly it remains approximately 5% to 15%.^{2,4} There is a significant amount of morbidity caused by appendiceal rupture.¹²⁻¹⁵ The incidence of perforation in patients with appendicitis ranges from 17% to 40%, with a median of 20%.^{16,17} The perforation rate is significantly higher in the elderly, with rates as high as 60% to 70%. Several factors contribute to the increased incidence of perforation in the elderly, including significant delay in seeking care, nonspecificity of the presenting symptoms and signs, diminished febrile response, and fewer abnormalities in important laboratory characteristics such as the white blood cell count (WBC).^{2,3,5,14,18,19} Children also have

an increased incidence of perforation because of delays in consulting a physician for abdominal pain.⁸ The negative laparotomy result rate in most series ranges from 15% to 35% and creates morbidity.^{16,17,20-22} In younger women, the negative laparotomy result rate is significantly higher (up to 45%) because of the prevalence of pelvic inflammatory disease and other common obstetric and gynecologic disorders.^{16,17,23,24}

THE ACCURACY OF OTHER DIAGNOSTIC MODALITIES

Routine medical history and physical examination remain the most effective and practical diagnostic modalities.^{25,26} Several other clinical methods for diagnosing appendicitis have been studied. Computer or algorithm-driven analyses of patients with abdominal pain have been evaluated,²⁷⁻³⁵ although most studies have incomplete controls and yield inconsistent results. Thus, the utility of computer-guided diagnosis compared with unassisted clinical diagnosis needs further evaluation. The authors of most of these studies believe that the improved utility they demonstrated was primarily because clinicians were forced to focus on specific clinical data that were readily available to be entered into the analysis tree. Finally, these authors observed that all of these modalities completely depend on the accuracy of the data gathered and interpreted by clinicians before the data are entered into the computer or algorithm analysis. The concept of an extended period of observation of patients with questionable appendicitis has been shown by some authors to be helpful.^{8,27,28} Its utility, like that of computer and algorithm analyses, depends on routine medical history and physical examination skills of clinicians.

The utility of radiographic techniques has also been evaluated. Plain abdominal radiographs and barium enemas are neither specific nor sensitive for appendicitis.³⁶ Ultrasonography is more effective in detecting a distended appendix than appendiceal perforation.^{10,15,36-44} No study has demonstrated ultrasonography to be clearly superior to the clinical examination, and many authors believe that its primary utility is to supplement the medical history and physical examination in patients with equivocal findings. The accuracy of computed tomography in diagnosing appendicitis has also been inconsistent.^{36,42,43}

Laparoscopy has been shown by some authors to be useful, particularly in young women in whom it can be difficult to differentiate between pelvic inflammatory disease, ectopic pregnancy, and appendicitis.²⁷ However, other series have not been as supportive, with negative appendectomy result rates from 20% to 30%.^{44,45} Studies of outcomes comparing laparoscopy with laparotomy have yielded conflicting results.^{46,47} Even though ultrasonography, computed tomography, and laparoscopy can be helpful, none are ideal techniques, and the clinician must depend on patient medical history and physical examination results.

APPENDICEAL ANATOMY AND PATHOPHYSIOLOGY OF APPENDICITIS

The adult's appendix averages 10 cm in length, arising from the posteromedial wall of the cecum, about 3 cm below the

ileocecal valve.⁴⁸ Its position in the abdominal cavity is variable, being described as retrocecal, retroileal, preileal, subcecal, or pelvic, and this variability in location may influence the clinical signs and symptoms associated with appendicitis. Although the physiologic role of the appendix is unproved, an immunologic function is suggested by its content of lymphoid tissue.⁴⁹

Appendiceal obstruction, followed by secondary bacterial invasion, causes the majority of appendicitis. Continued fluid secretion by the mucosa of the obstructed appendix distends the lumen, eventually exceeding venous pressure and leading to tissue ischemia and, ultimately, necrosis. Causes of obstruction include fecoliths, calculi, tumors, parasites, foreign bodies, or, rarely, barium. In the one-third of patients without apparent obstruction, infection by viruses, parasites, or bacteria, or either trauma or postoperative fecal stasis may be involved.⁵⁰⁻⁵⁵

Normally, appendicitis presents with a highly characteristic sequence of symptoms and signs.⁵⁶ Initially, appendicitis causes visceral pain poorly localized to the epigastrium or periumbilical region, presumably because of distention of the appendix. Anorexia, nausea, and vomiting soon follow as this pathophysiology worsens. More advanced inflammation causes irritation of adjacent structures or the peritoneum, low-grade fever, and peritoneal pain localized to the RLQ. The pathophysiology explains the classic migration of pain caused by appendicitis. The point of maximal tenderness may be distinct from McBurney point, 5 cm from the anterior superior iliac spine on a line running from the umbilicus.

Atypical locations of the appendix may lead to unusual clinical findings. In the case of retrocecal or retroiliac appendices,^{57,58} the pain may be poorly localized and may not undergo the transition from epigastric to RLQ locations. Pelvic appendicitis frequently causes pain in the left lower quadrant, with an absence of tenderness, and is reflected by increased pain during a rectal examination. Unusual symptoms of urinary and defecation urgency, caused by irritation of the ureter and rectum, respectively, plus dysuria and diarrhea may also occur.

Although often a diagnostic dilemma in the first trimester of pregnancy because of confusion with other diagnoses, appendicitis in later stages of gestation may present a challenge for the clinician because of displacement of the appendix by the enlarging uterus. In such cases, periumbilical or right subcostal tenderness may be found.

HOW TO ELICIT THE RELEVANT SYMPTOMS AND SIGNS

Pain is commonly the first symptom of appendicitis.^{9,59} Classically, the vague, midepigastric or periumbilical pain awakens the patient from sleep but is not initially severe. After reaching its peak in around 4 hours, it diminishes and then migrates to the RLQ. Most patients will seek medical attention within 12 to 48 hours. Pain usually occurs before vomiting, and the patient has usually not experienced similar symptoms before the present episode.

According to Cope's *Early Diagnosis of the Acute Abdomen*,⁶⁰ many patients feel constipated and anticipate that defecation will relieve discomfort, leading them to use cathartic agents. However, pain persists after a bowel movement.

Many signs have been associated with appendicitis or peritonitis. Some of obvious value, such as the pelvic examination, have not been adequately evaluated to merit mention in this systematic review or they lack an adequate description or standardization of the elicitation of the sign to ensure accurate reproduction. A common reference for definitions in the best studies is a text by De Dombal.⁶¹ What follows is the most consistent and useful description of the signs:

- Guarding: Guarding is a state of voluntary contraction of the abdominal muscles. The muscles are held tense by the patient because he or she knows (or fears) that further examination is likely to be painful. Fear can be partially, or fully, overcome by tact and persuasion.⁶¹
- Rigidity: Rigidity is also known as involuntary guarding. The best studies of abdominal pain have described rigidity as an involuntary reflex spasm of the muscles of the abdominal wall. It can never be overcome by tact and reassurance.⁶¹
- Rebound tenderness: (1) Press on the area of question with the flat of your hand, sufficient to depress the peritoneum. The patient should be experiencing pain. (2) Keep pressing with a constant intensity. As the patient adjusts to this pressure during 30 to 60 seconds, the pain diminishes. It may go away completely, although usually it does not. (3) Without warning, and preferably while the patient's attention is distracted, remove the hand suddenly to just above skin level. Watching the patient grimace is more indicative than a complaint of pain.⁶¹
- Rovsing sign: A sign related to the rebound tenderness test. Press deeply and evenly in the left lower quadrant and then release pressure suddenly. The presence of tenderness in the RLQ during palpation or referred rebound tenderness in the RLQ during release is considered a positive Rovsing sign.
- Psoas sign: With the patient in the supine position, ask the patient to lift the thigh against your hand, placed just above the knee. Alternatively, with the patient in the left lateral decubitus position (Figure 5-1), extend the patient's right leg at the hip. Increased pain with either maneuver is a positive sign and indicates irritation of the psoas muscle by an inflamed appendix.
- Obturator sign: This sign is similar mechanically to the psoas sign. It is elicited by passively flexing the right hip and knee and internally rotating the leg at the hip, stretching the obturator muscle (Figure 5-2). Resultant right-sided abdominal pain is a positive sign, indicating irritation of the obturator muscle. The obturator sign has not been studied independent of the psoas sign, but most clinicians would attribute the same significance.
- Rectal examination: Classically, tenderness and fullness perceived on the right but not the left side on rectal examination are indicative of a pelvic appendicitis.⁶⁰ This sign is subjective and poorly described in most major physical examination texts. No studies that assess rectal tenderness describe the examination technique.

PRECISION OF THESE SYMPTOMS AND SIGNS

There have been no studies published evaluating the precision of the clinical examination for appendicitis. A standardized clinical examination might produce strong interrater reliability.

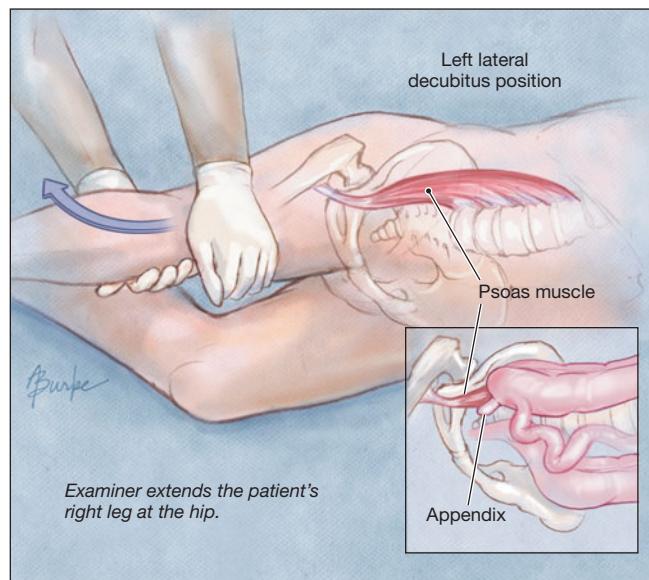


Figure 5-1 The Psoas Sign in Examination for Appendicitis

The sign can be elicited with 2 different patient positions. First, with the patient in the supine position, ask the patient to lift the right thigh against your hand placed just above the knee. With the patient in the left lateral decubitus position (as shown), extend the right leg at the hip. Increased pain with either maneuver is a positive sign and indicates irritation of the psoas muscle by an inflamed appendix.

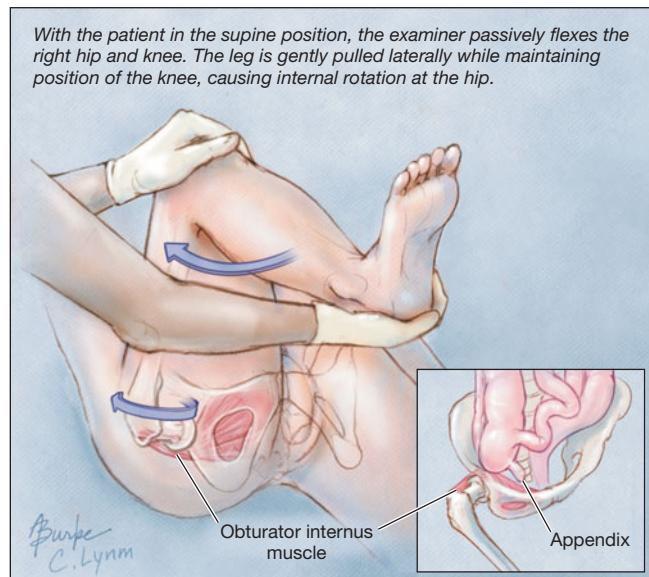


Figure 5-2 The Obturator Sign in Examination for Appendicitis

Elicit this sign by passively flexing the patient's right hip and knee and internally rotating the leg at the hip, stretching the obturator muscle. Resultant right-sided abdominal pain is a positive sign, indicating irritation of the obturator muscle.

ACCURACY OF THESE SYMPTOMS AND SIGNS

A handful of studies published during the past few decades have evaluated the accuracy of the clinical presentation of appendicitis. The studies are of various quality and design. Most are best described as cross-sectional in design because a clinical judgment is made, with outcomes measured in terms of pathologic confirmation of appendicitis vs a negative laparotomy result or no requirement for surgery. Eleven of the highest-quality studies, based on number of patients studied, the study design, and completeness of reported data, are summarized in Table 5-1.^{9,24,33,35,62-67} The search strategy for identifying these articles is available from the authors on request. This strategy yielded about 300 articles since 1966. Further limiting sets to adult age groups yielded 200 studies. The titles and abstracts were reviewed and chosen if adequate detail of the outcomes and aspects of the clinical examination allowed construction of 2 × 2 tables and subsequent calculation of likelihood ratios [LRs].

The 11 studies were divided into 2 groups by the patients on whom they focused. Approximately half of the studies focused on patients in whom appendicitis was suspected, and half, on those who were examined for acute abdomen. In the studies of suspected appendicitis, the inclusion criteria were not further defined. In the studies of acute abdomen, inclusion criteria usually involved pain for less than 1 week. Taken together, the studies report on the findings of more than 4000 patients and provide the best available evidence supporting the most valuable aspects of the clinical examination for appendicitis (Table 5-2).

Each study reports on a varying constellation of clinical findings. Many aspects of the clinical examination are not evaluated in all of these studies. Unfortunately, some of the aspects evaluated are poorly defined in the text of the studies,

so specific recommendations for these aspects are difficult to derive for medical education or the everyday practice of medicine.

Nonetheless, several points can be drawn from a systematic literature review. In evaluation of patients presenting with emergency and acute abdominal pain, usually defined as less than 1 week in duration before presenting to an emergency department or surgical ward, the prevalence (pretest probability) of acute appendicitis ranges from 12% to 26%.^{12,30,32,69} The clinical examination will influence this probability further. If various aspects of the clinical examination are viewed as diagnostic tests, LRs^{70,71} and posttest probability can be calculated.

From the medical history, 6 aspects have been evaluated. Seven physical examination items have also been studied well. These aspects are examined further in Table 5-3.⁷² The large number of patients studied and the similarities across studies make the data suitable for being combined into summary measures.

Three findings show a high positive LR (LR+) across all studies and, when present, are most useful for identifying patients at increased likelihood for appendicitis: RLQ pain (LR+, 8.0), rigidity (LR+, 4.0), and migration of initial periumbilical pain to the RLQ (LR+, 3.2). Rebound tenderness was studied in most patients, but its positive likelihood varied too much to allow a statistical point estimate of its effect (LR+, 1.1-6.3). Although the obturator sign has not been studied independently, the authors suspect that this sign has operating characteristics similar to those of the psoas sign.

Clinicians also collect evidence to help prove normality. Unfortunately, no single component consistently provided a low negative LR (LR-) that would rule out appendicitis. There were, however, many signs that proved to be helpful in ruling out appendicitis. The absence of RLQ pain and

Table 5-1 Studies of the Operating Characteristic of the Clinical Examination for Appendicitis

Authors	Year	Inclusion Criteria	Design	No. of Patients Studied (% Women)	Country	Age Range, y
Staniland et al ⁶²	1972	Admitted for acute abdomen	Retrospective	600 (49)	United Kingdom	<9 to >70
Brewer et al ⁶³	1976	ED evaluation for acute abdomen	Retrospective	1000 (0)	United States	15 to >65
Berry and Malt ²⁴	1984	Operation for suspected appendicitis	Retrospective	300 (40)	United States	10 to >50
Nauta and Magnant ⁶⁴	1986	Operation for suspected appendicitis	Prospective	97 (40)	United States	2 to 91
Alvarado ³³	1986	Admitted for suspected appendicitis	Retrospective	305 (42)	United States	4 to 80
Fenyö ³⁵	1987	Admitted for suspected appendicitis	Prospective	830 (57)	Sweden	15 to 86
Liddington and Thomson ⁶⁵	1991	Admitted for abdominal pain	Prospective	150 (58)	United Kingdom	7 to 84
Dixon et al ⁹	1991	Admitted for suspected appendicitis	Prospective	1204 (39)	Scotland	7 to 87
Izbicki et al ⁶⁶	1992	ED evaluation for suspected appendicitis	Prospective	150 (56)	Germany	11 to 88
Eskelinen et al ⁶⁷	1994	Admitted for abdominal pain	Prospective	222 (58)	Finland	65 to 90
Eskelinen et al ⁶⁸	1995	Admitted for abdominal pain	Prospective	417 (54)	Finland	>50
Total				5275 (41)		

Abbreviation: ED, emergency department.

the presence of similar previous pain demonstrated powerful LR- (0.28 and 0.25, respectively). The absence of the classic migration of pain also diminished the likelihood of appendicitis significantly (LR-, 0.5). The absence of RLQ guarding or rebound pain has excellent properties for ruling out appendicitis in some studies, but not others. The

presence of pain before vomiting needs further study to identify its diagnostic efficiency because, in its only evaluation, it was highly efficient in ruling out appendicitis. Astute clinicians will recognize that the absence of anorexia, nausea, or vomiting has little effect on the likelihood of appendicitis.

Table 5-2 Aspects of the Clinical Examination Studied^a

Author	Pain Migr	Anorexia	Nausea	Vomiting	Pain	Similar	Rectal	Psoas	RLQ Pain	Rebound	Rigid	Guard	Fever
Staniland et al ⁶²		×	×	×		×			×	×	×	×	
Brewer et al ⁶³	×	×		×	×	×				×			×
Berry and Malt ²⁴		×	×				×	×	×	×		×	
Nauta and Magnant ⁶⁴	×	×	×	×			×	×	×	×			
Alvarado ³³	×	×	×				×		×	×			×
Fenyö ³⁵	×									×	×		
Liddington and Thomson ⁶⁵										×			
Dixon et al ¹⁹							×		×	×	×	×	
Izbicki et al ⁶⁶	×					×		×	×	×	×	×	
Eskelinen et al ⁶⁷						×	×		×	×	×	×	×
Eskelinen et al ⁶⁸	×	×	×	×	×		×		×	×	×	×	×
No. of cases studied	1354	2161	1691	1684	651	1542	2349	450	3979	4688	3555	2267	1264

Abbreviations: Migr, migration of the initial periumbilical pain to the right lower quadrant; pain, pain before vomiting; psoas, positive psoas sign; rectal, pain on rectal examination; RLQ, right lower quadrant; similar, symptoms similar to those the patient previously experienced.

^aFor an explanation of rebound, rigid, and guard, see the "How to Elicit the Relevant Symptoms and Signs" section of the text.

Table 5-3 Summary of Clinical Examination Operating Characteristics for Appendicitis^a

Procedure	Sensitivity	Specificity	LR+ (95% CI)	LR- (95% CI)
Right lower quadrant pain	0.84	0.90	7.3-8.5 ^b	0-0.28 ^b
Rigidity	0.20	0.89	3.8 (3.0-4.8)	0.82 (0.79-0.85)
Migration of pain	0.64	0.82	3.2 (2.4-4.2)	0.50 (0.42-0.59)
Pain before vomiting ^c	1.0	0.64	2.8 (1.9-3.9)	NA
Psoas sign	0.16	0.95	2.4 (1.2-4.7)	0.90 (0.83-0.98)
Fever	0.67	0.79	1.9 (1.6-2.3)	0.58 (0.51-0.67)
Rebound tenderness test	0.63	0.69	1.1-6.3 ^b	0-0.86 ^b
Guarding	0.73	0.52	1.7-1.8 ^b	0-0.54 ^b
No similar pain previously	0.86	0.40	1.50 (1.46-1.7)	0.32 (0.25-0.42)
Rectal tenderness	0.41	0.77	0.83-5.3 ^b	0.36-1.1 ^b
Anorexia	0.68	0.36	1.3 (1.2-1.4)	0.64 (0.54-0.75)
Nausea	0.58	0.37	0.69-1.2 ^b	0.70-0.84 ^b
Vomiting	0.51	0.45	0.92 (0.82-1.0)	1.1 (0.95-1.3)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio; NA, not available.

^aAll studies were used to create 2 × 2 tables and then tested for homogeneity of the odds ratio with the Breslow-Day statistic. If studies were not rejected as heterogeneous by this statistic, $P = .05$, CIs were manually reviewed to exclude type II errors. Studies satisfying both criteria were combined, and LRs were calculated with the Mantel-Haenszel method. The 95% CIs were calculated according to the method of Simel et al.⁷² Only 1 study evaluated pain before vomiting. For an explanation of procedure terms, see Table 5-2 or the "How to Elicit the Relevant Symptoms and Signs" section of the text.

^bIn heterogeneous studies, the LRs are reported as ranges.

^cOnly 1 study on this in the meta-analysis.

THE ROLE OF COMBINED FINDINGS

Clinicians rarely rely on a single sign or symptom for diagnosis but instead rely on a combination of findings. Unfortunately, the precision and accuracy of combinations of findings have not been reported in these studies. Several studies do assess, however, various decision rules that do combine these findings.^{6,33-35,66,73-77} Four of the most powerful rules were validated on an independent set of 1254 patients older than 50 years and presenting with abdominal pain. No single score was found to be superior; however, it was observed that the decision rules reported in the original work to be most powerful incorporated at least 2 of 5 common variables: site and duration of pain, site of tenderness, rebound tenderness, and leukocytosis.⁷⁸

THE BOTTOM LINE

Returning to the beginning clinical scenario, the historical components of the presentation are highly suggestive of appendicitis. Our patient demonstrates the classic sequence of abdominal pain before vomiting, culminating with the migration of the initial midepigastric pain to the RLQ. The combination of these LR+s alone makes appendicitis more likely.

The findings of guarding but not rigidity tend to neutralize each other's effect. The rectal examination results and the psoas and related signs are helpful if present but are not helpful when absent, as in this case. In sum, we suspect appendicitis in this man, so further evaluation is warranted.

A surgical doctrine suggests that a decrease in the perforation rate will be achieved only by an increase in the negative laparotomy result rate in suspected acute appendicitis. The truth of this doctrine has been called into question, given the results of large- and small-area variation studies.²⁹ Improved clinical evaluation is suggested as a remedy for a high rate of negative laparotomy results without increasing the perforation rate. Evidence suggests the essential nature of clinical details.^{79,80} Clinicians often do not collect enough clinical details for accurate and precise diagnosis.⁸¹⁻⁸³ Correction of this deficit, therefore, may well increase diagnostic accuracy without increasing the perforation rate.

In summary, there are several conclusions that can be made concerning the clinical presentation, pathophysiology, and diagnosis of appendicitis:

1. Appendicitis is a common clinical entity, with significant morbidity and mortality, particularly at the extremes of age.
2. The pathophysiology of appendicitis consists of initial dilatation of the appendix, followed by appendiceal ischemia, necrosis, and parietal peritoneal irritation. Clinical findings are predictable, predicated on knowledge of this pathophysiology.
3. The characteristic sequence of symptoms and signs includes the following: (1) vague pain initially located in the epigastric or periumbilical region; (2) anorexia, nausea, or unsustained vomiting; (3) migration of the initial pain to the RLQ; and (4) low-grade fever.

4. Migration of pain in the characteristic manner, RLQ pain, and the presence of pain before vomiting are historical findings that suggest appendicitis. The presence of rigidity, a positive psoas sign, fever, or rebound tenderness is a sign on physical examination indicating an increased likelihood of appendicitis.
5. Conversely, the absence of RLQ pain, the absence of the classic migration of pain, and the presence of similar pain previously are powerful symptoms in the medical history that make appendicitis less likely. In the physical examination, the lack of RLQ pain, rigidity, or guarding makes appendicitis less likely.
6. Because no finding on the clinical examination can effectively rule out appendicitis, prudence dictates close follow-up of patients with abdominal pain who do not receive further diagnostic testing.

Author Affiliations at the Time of the Original Publication

Department of Internal Medicine, University of Texas, Southwestern Medical Center, Dallas, Texas (Drs Wagner, McKinney, and Carpenter).

Acknowledgments

We appreciate the expert advice offered by gynecologist Joanne Piscitelli, MD, and general surgeon Ted Pappas, MD, both of Duke University, Durham, North Carolina, during the preparation of the manuscript.

REFERENCES

1. Irvin TT. Abdominal pain: a surgical audit of 1190 emergency admissions. *Br J Surg.* 1989;76(11):1121-1125.
2. Balsano N, Cayten CG. Surgical emergencies of the abdomen. *Emerg Med Clin North Am.* 1990;8(2):399-410.
3. Bugliosi TF, Meloy TD, Vukov LF. Acute abdominal pain in the elderly. *Ann Emerg Med.* 1990;19(12):1383-1386.
4. Fenyo G. Acute abdominal disease in the elderly. *Am J Surg.* 1982; 143(6):751-754.
5. Fenyo G. Diagnostic problems of acute abdominal diseases in the aged. *Acta Chir Scand.* 1974;140(5):396-405.
6. Wasson JH, Sox HC Jr, Sox CH. The diagnosis of abdominal pain in ambulatory male patients. *Med Decis Making.* 1981(3);1:215-224.
7. Britt H, Bridges-Webb C, Sayer GP, Neary S, Traynor V, Charles J. The diagnostic difficulties of abdominal pain. *Aust Fam Physician.* 1994;23 (3):375-381.
8. White JJ, Santillana M, Haller JA. Intensive in-hospital observation: a safe way to decrease unnecessary appendectomy. *Am Surg.* 1975;41(12):793-798.
9. Dixon JM, Elton RA, Rainey JB, Macleod DA. Rectal examination in patients with pain in the right lower quadrant of the abdomen. *BMJ.* 1991;302(6781):386-388.
10. Gamal R, Moore TC. Appendicitis in children aged 13 years and younger. *Am J Surg.* 1990;159(6):589-592.
11. Putnam TC, Gagliano N, Emmens RW. Appendicitis in children. *Surg Gynecol Obstet.* 1990;170(6):527-532.
12. Jess P, Bjerregaard B, Brynitz S, Holst-Christensen J, Kalaja E, Lund-Kristensen J. Acute appendicitis: prospective trial concerning diagnostic accuracy and complications. *Am J Surg.* 1981;141(2):232-234.
13. Mittelpunkt A, Nora PF. Current features in the treatment of acute appendicitis: an analysis of 100 consecutive cases. *Surgery.* 1966;60 (5):971-975.
14. Franz MG, Norman J, Fabri PJ. Increased morbidity of appendicitis with advancing age. *Am Surg.* 1995;61(1):40-44.
15. Wade DS, Morrow SE, Balsara ZN, Burkhardt TK, Goff WB. Accuracy of ultrasound in the diagnosis of acute appendicitis compared with the surgeon's clinical impression. *Arch Surg.* 1993;128(9):1039-1046.

16. Lewis FR, Holcroft JW, Boey J, Dunphy E. Appendicitis: a critical review of diagnosis and treatment in 1000 cases. *Arch Surg.* 1975;110(5):677-684.
17. Addiss DG, Shaffer N, Fowler BS, Tauxe RV. The epidemiology of appendicitis and appendectomy in the United States. *Am J Epidemiol.* 1990;132(5):910-925.
18. Owens BJ, Hamit HF. Appendicitis in the elderly. *Ann Surg.* 1978;187(4):392-396.
19. Peltokallio P, Jauhainen K. Acute appendicitis in the aged patient. *Arch Surg.* 1970;100(2):140-143.
20. Jerman RP. Removal of the normal appendix: the cause of serious complications. *Br J Clin Pract.* 1969;23(11):466-467.
21. Chang FC, Hogle HH, Welling DR. The fate of the negative appendix. *Am J Surg.* 1973;126(6):752-754.
22. Howie JGR. Death from appendicitis and appendicectomy. *Lancet.* 1966;2(7477):1334-1335.
23. Bongard F, Landers DV, Lewis F. Differential diagnosis of appendicitis and pelvic inflammatory disease. *Am J Surg.* 1985;150(1):90-96.
24. Berry J, Malt RA. Appendicitis near its centenary. *Ann Surg.* 1984;200(5):567-575.
25. Peterson MC, Holbrook JH, Hales DV, Smith NL, Staker LV. Contributions of the history, physical examination, and laboratory investigation in making medical diagnoses. *West J Med.* 1992;156(2):163-165.
26. Hampton JR, Harrison MJG, Mitchell JRA, Prichard JS, Seymour C. Relative contributions of history-taking, physical examination, and laboratory investigation to diagnosis and management of medical outpatients. *BMJ.* 1975;2(5969):486-489.
27. Jones PF. Practicalities in the management of the acute abdomen. *Br J Surg.* 1990;77(4):365-367.
28. Schwartz SI. Tempering the technological diagnosis of appendicitis. *N Engl J Med.* 1987;317(11):703-704.
29. Neutra RR. Appendicitis: decreasing normal removals without increasing perforations. *Med Care.* 1978;16(11):956-961.
30. Adams ID, Chan M, Clifford PC, et al. Computer aided diagnosis of acute abdominal pain: a multicentre study. *BMJ.* 1986;293:800-804.
31. De Dombal ET. Educational assessment of clinical diagnostic skills: studies across Europe on acute abdominal pain. *Postgrad Med J.* 1993;69(suppl 2):S94-S96.
32. De Dombal ET, Leaper DJ, Horrocks JC, Staniland JR, McCann AP. Human and computer-aided diagnosis of abdominal pain: further report with emphasis on performance of clinicians. *BMJ.* 1974;1(5904):376-380.
33. Alvarado A. A practical score for the early diagnosis of acute appendicitis. *Ann Emerg Med.* 1986;15(5):557-564.
34. Bond GR, Tully SB, Chan LS, Bradley RL. Use of the MANTRELS score in childhood appendicitis: a prospective study of 187 children with abdominal pain. *Ann Emerg Med.* 1990;19(9):1014-1018.
35. Fenyo G. Routine use of a scoring system for decision-making in suspected acute appendicitis in adults. *Acta Chir Scand.* 1987;153(9):545-551.
36. Brazaitis MP, Dachrran AH. The radiologic evaluation of acute abdominal pain of intestinal origin. *Med Clin North Am.* 1993;77(5):939-961.
37. Puylaert JBCM, Rutgers PH, Lalisaang RI, et al. A prospective study of ultrasonography in the diagnosis of appendicitis. *N Engl J Med.* 1987;317(11):666-669.
38. John H, Neff U, Kelemen M. Appendicitis diagnosis today: clinical and ultrasonic deductions. *World J Surg.* 1993;17(2):243-249.
39. Davies AH, Mastorakou I, Cobb R, Rogers C, Lindell D, Mortenson NJM. Ultrasonography in the acute abdomen. *Br J Surg.* 1991;78(10):1178-1180.
40. Anteby SO, Schenker JG, Polishuk WZ. The value of laparoscopy in acute pelvic pain. *Ann Surg.* 1974;118(4):484-486.
41. Kang WM, Lee CH, Chou YH, et al. A clinical evaluation of ultrasonography in the diagnosis of acute appendicitis. *Surgery.* 1989;105(2 pt 1):154-159.
42. Taourel P, Baron MP, Pradel J, Fabre JM, Seneterre E, Bruel JM. Acute abdomen of unknown origin: impact of CT on diagnosis and management. *Gastrointest Radiol.* 1992;17(4):287-291.
43. Lim HK, Bae SH, Seo GS. Diagnosis of acute appendicitis in pregnant women: value of sonography. *AJR Am J Roentgenol.* 1992;159(3):539-542.
44. Sarafati MR, Hunter GC, Witzke DB, et al. Impact of adjunctive testing on the diagnosis and clinical course of patients with acute appendicitis. *Am J Surg.* 1993;166(6):664-665.
45. Whitworth CM, Whitworth PW, Sanfillipo J, Polk HCJ. Value of diagnostic laparoscopy in young women with possible appendicitis. *Surg Gynecol Obstet.* 1988;167(3):187-190.
46. Schroder DM, Lathrop JC, Lloyd LR, Coccaccio JE, Hawasli A. Laparoscopic appendectomy for acute appendicitis: is there really any benefit? *Am Surg.* 1993;59(8):541-547; discussion 547-548.
47. Koliias J, Harries RH, Otto G, Hamilton DW, Cox JS, Gallery RM. Laparoscopic versus open appendectomy for suspected appendicitis: a prospective study. *Aust N Z J Surg.* 1994;64(12):830-835.
48. Buschard K, Kjaeldgaard A. Investigation and analysis of the position, fixation, length, and embryology of the vermiform appendix. *Acta Chir Scand.* 1973;139(3):293-298.
49. Bjerke K, Brantzaeg P, Rognum TO. Distribution of immunoglobulin producing cells is different in normal human appendix and colon mucosa [abstract]. *Gut.* 1986;27(6):667.
50. Lin J, Bleiweiss U, Mendelson MH. Cytomegalovirus-associated appendicitis in a patient with the acquired immunodeficiency syndrome. *Am J Med.* 1990;89(3):377-379.
51. Lopez-Navidad A, Domingo P, Cadafalch J. Acute appendicitis complicating infectious mononucleosis: case report and review. *Rev Infect Dis.* 1990;12(2):297-302.
52. Nadler S, Cappell MS, Bhatt B. Appendiceal infection by *Entamoeba histolytica* and *Strongyloides stercoralis* presenting like acute appendicitis. *Dig Dis Sci.* 1990;35(5):603-608.
53. Hennington MH, Tinsley EJ, Proctor HJ. Acute appendicitis following blunt abdominal trauma: incidence or coincidence? *Am Surg.* 1991;214(1):61-63.
54. Houghton A, Aston N. Appendicitis complicating colonoscopy. *Gastrointest Endosc.* 1988;34(6):489.
55. Barr D, van Heerden JA, Mucha PJ. The diagnostic challenge of postoperative acute appendicitis. *World J Surg.* 1991;15(4):526-528.
56. Schrock TR. Acute appendicitis. In: Sleisinger MH, Fortran JS, eds. *Gastrointestinal Disease: Pathophysiology, Diagnosis, Management.* Philadelphia, PA: WB Saunders Co; 1996:1339-1347.
57. Poole GV. Anatomic basis for delayed diagnosis of appendicitis [abstract]. *South Med J.* 1990;83(7):771-773.
58. Shen GK, Wong R, Daller J, et al. Does the retrocecal position of the vermiform appendix alter the clinical course of acute appendicitis? a prospective analysis [abstract]. *Arch Surg.* 1991;126(5):569-570.
59. Maxwell JM, Ragland JJ. Appendicitis: improvements in diagnosis and treatment. *Am Surg.* 1991;57(5):282-285.
60. Silen W. *Cope's Early Diagnosis of the Acute Abdomen.* New York, NY: Oxford University Press Inc; 1987:1-290.
61. De Dombal FT. *Diagnosis of Acute Abdominal Pain.* New York, NY: Churchill Livingstone Inc; 1991:1-259.
62. Staniland JR, Ditchburn J, De Dombal ET. Clinical presentation of acute abdomen: study of 600 patients. *BMJ.* 1972;3(5823):393-398.
63. Brewer RJ, Golden OT, Hitch DC, Rudolf LE, Wangenstein SL. Abdominal pain: an analysis of 1000 consecutive cases in a university hospital emergency room. *Am J Surg.* 1976;131:219-223.
64. Nauta RI, Magnant C. Observation versus operation for abdominal pain in the right lower quadrant: roles of the clinical examination and the leukocyte count. *Am J Surg.* 1986;151(6):746-748.
65. Liddington MI, Thomson WHF. Rebound tenderness test. *Br J Surg.* 1991;78(7):795-796.
66. Izicki JR, Wolfman TK, Dietmar KW, et al. Accurate diagnosis of acute appendicitis: a retrospective and prospective analysis of 686 patients. *Eur J Surg.* 1992;158(4):227-231.
67. Eskelinen M, Ikonen J, Lippinen P. Acute appendicitis in patients over the age of 65 years: comparison of clinical and computer based decision making. *Int J Biomed Comput.* 1994;36:239-249.
68. Eskelinen M, Ikonen J, Lippinen P. The value of history-taking, physical examination, and computer assistance in the diagnosis of acute appendicitis in patients more than 50 years old. *Scand J Gastroenterol.* 1995;30(4):349-355.
69. Wilson DH, Wilson PD, Walmsley RG, Horrocks JC, De Dombal FT. Diagnosis of acute abdominal pain in the accident and emergency department. *Br J Surg.* 1977;64(4):250-254.
70. Nardone DA, Lucas LM, Palac DM. Physical examination: a revered skill under scrutiny. *South Med J.* 1988;81(6):770-773.
71. Sackett KL, Haynes RB, Guyatt GH, Tugwell P. *Clinical Epidemiology: A Basic Science for Clinical Medicine.* Boston, MA: Little Brown & Co; 1991:173-186.
72. Simel DL, Samsa GP, Matchar DB. Likelihood ratios with confidence: sample size estimation for diagnostic test studies. *J Clin Epidemiol.* 1991;44(8):763-770.

73. Christian F, Christian GP. A simple scoring system to reduce the negative appendicectomy rate. *Ann R Coll Surg Engl.* 1992;74(4):281-285.
74. Orient JM, Kettel LJ, Lim J. A test of a linear discriminant for identifying low-risk abdominal pain. *Med Decis Making.* 1985;5(1):77-87.
75. Ramirez JM, Deus J. Practical score to aid decision making in doubtful cases of appendicitis. *Br J Surg.* 1994;81(5):680-683.
76. Teicher I, Landa B, Cohen M, Kabnick LS, Wise L. Scoring system to aid in the diagnosis of appendicitis. *Ann Surg.* 1983;198(6):753-759.
77. Morgan DL, Hayes JE, Zenarosan N. Predicting which abdominal pain patients can be appropriately discharged home [abstract]. *Ann Emerg Med.* 1995;25:140.
78. Ohmann C, Yang Q, Franke C. Diagnostic scores for acute appendicitis. *Eur J Surg.* 1995;161(4):273-281.
79. Lavelle SM, Kanagaratnam B. The information value of clinical data. *Int J Biomed Comput.* 1990;26(3):203-209.
80. Todd BS, Stamper R. Limits to diagnostic accuracy. *Med Inform.* 1993;18(3):255-270.
81. Avorn J, Everitt DE, Baker MW. The neglected medical history and therapeutic choices for abdominal pain: a nationwide study of 799 physicians and nurses. *Arch Intern Med.* 1991;151(4):694-698.
82. Johnson JE, Carpenter JL. Medical house staff performance in physical examination. *Arch Intern Med.* 1986;146(5):937-941.
83. Wiener S, Nathanson M. Physical examination: frequently observed errors. *JAMA.* 1976;236(7):852-855.

Prepared by Jim Wagner, MD
Reviewed by Kaveh Shojania, MD

CLINICAL SCENARIO

A 24-year-old woman presents with abdominal pain, nausea, and vomiting. She describes the pain as beginning in her midabdomen 3 days ago, and it has gotten progressively worse. Her last menstrual period was 3 weeks ago and was normal; she is not sexually active. The pain has stayed in the midabdomen and not moved to other locations. On examination, she has a fever and right lower quadrant (RLQ) and rebound tenderness; her pelvic and rectal examination results are unremarkable. Laboratory evaluation reveals a left shift without leukocytosis and ketonuria.

UPDATED SUMMARY ON ADULT APPENDICITIS

Original Review

Wagner JM, McKinney WP, Carpenter JL. Does this patient have appendicitis? *JAMA*. 1996;276(19):1589-1594.

UPDATED LITERATURE SEARCH

Our literature search used the parent search for the Rational Clinical Examination series, combined with the subject headings “exp appendicitis” published between 1994 and September 2004. This search yielded more than 400 titles, which were narrowed down to approximately 50 by excluding studies of laboratory and radiologic tests and case studies.

There have been few new studies that focused on the operating characteristics of individual components of the clinical examination for appendicitis. However, there have been several studies that have looked at combinations of findings. That is, instead of examining the likelihood ratio (LR) of rebound tenderness alone, studies have explored the combination of fever, migration of pain, and rebound tenderness.

The studies of clinical decision rules were selected if the components, derivation, and validation of the prediction rule were clearly defined in the article and the patients included were those from a general population with abdominal pain or were suspected of having appendicitis. Our previous literature search was reviewed, and studies conducted before 1994 were included if they fit these criteria.

NEW FINDINGS

- Combinations of findings from the clinical examination are more powerful than any single finding.
- Most of the decision rules formed by these combinations of findings include migration of pain from periumbilical to RLQ, rebound tenderness, RLQ tenderness, nausea-vomiting, male sex, fever, rigidity, and white blood cell (WBC) count.

Details of the Update

Eighteen studies that derived or validated clinical decision rules for appendicitis were identified. The most important studies were those by Alvarado,¹ Eskelinen et al,² and Fenyo et al.³ These studies were chosen because of their methodology, large sample sizes, simplicity of the decision rule, or familiarity with physicians. In addition, a study that compared several clinical decision rules on the same population provided a good perspective of the relative value of these rules.⁴

The Alvarado¹ study was one of the first of the clinical decision rules published, demonstrating the power of the rule beyond individual findings. Although the methods are rudimentary and the rule is not validated in the study, it represents the most widely accepted and the simplest of the clinical decision rules. By combining the results for 8 findings from the medical history or the examination (which conveniently spells out the mnemonic MANTRELS), the resulting score provides guidance on whether to operate in the setting of suspected appendicitis. Of 10 potential points, patients with a score of 7 or higher are recommended for surgical intervention. The various components are Migration, Anorexia-acetone, Nausea-vomiting, Tenderness in RLQ, Rebound pain, Elevation of temperature, Leukocytosis, and Shift to the left of normal WBC count.

The Eskelinen et al² study evaluated more than a thousand patients with a rule that includes 7 variables in men and 5 in women. The disadvantages of this study are that the rule is complex, computer based, and was validated with a small number of patients.

The Fenyo et al³ study assessed 10 variables used in a complex equation. The results for the individual findings showed that a WBC count of less than $8.9 \times 10^9/L$ (LR, 0.16) was the one finding that had reasonable measurement properties, leading to a lower likelihood of appendicitis.

The Ohmann et al⁴ study displayed a parallel analysis of 10 available studies, including the 3 mentioned above. A database of 45 variables prospectively collected from 1254 consecutive patients on a standardized form was used to evaluate these studies. A surprising outcome of the study was that none of the rules produced sufficiently low rates (<15%) for either unneeded appendectomy result (rule advised surgery but normal appendix found) or delayed appendectomy (rule advised delay but the patient proved to have appendicitis). However, the clinicians in these studies did not perform much better than the rules. Although the clinicians who chose not to use the rules performed similarly to the decision rule results, implementing decision models in actual clinical practice may identify a subset of clinicians who improve with the rules. The authors recommend the Alvarado¹ and Eskelinen et al² studies as those warranting further evaluation.

Table 5-4 Accuracy of Selected Decision Rules

Author	Year	n	LR+ (95% CI)	LR− (95% CI)	DOR (95% CI)
Alvarado ¹	1986	277	3.1 (1.9-5.0)	0.26 (0.19-0.35)	12 (6.0-25)
Christian and Christian ⁹	1992	58	4.5 (1.8-11)	0 (0-0.23)	311 (15-6426)
Fenyö et al ³	1997	1167	5.6 (4.6-6.8)	0.31 (0.26-0.37)	18 (13-24)
Eskelinen et al ²	1994	1333	10 (8.3-12)	0.06 (0.04-0.10)	164 (93-287)
Izbicki et al ¹⁰	1992	150	1.9 (1.5-2.5)	0.21 (0.09-0.45)	9.5 (3.7-24)
Kalan et al ¹¹	1994	49	1.3 (0.81-2.1)	0.38 (0.11-1.3)	3.5 (0.66-19)
Ramirez and Deus ¹²	1994	166	4.3 (2.1-8.8)	0.25 (0.17-0.36)	17 (6.4-46)
Saidi and Ghasemi ¹³	2000	128	6.0 (3.6-9.9)	0.08 (0.03-0.24)	75 (20-280)

Abbreviations: CI, confidence interval; DOR, diagnostic odds ratio; LR+, positive likelihood ratio; LR−, negative likelihood ratio.

Table 5-5 The Alvarado Clinical Decision Rule (MANTRELS Mnemonic)

Variable	Score
Migration	1
Anorexia-acetone	1
Nausea-vomiting	1
Tenderness in RLQ	2
Rebound pain	1
Elevation of temperature	1
Leukocytosis	2
Shift to the left	1
Maximum total score	10
Positive	≥7

Abbreviation: RLQ, right lower quadrant.

What lessons can be learned from these studies? The rules recommended by experts incorporate a description of the pain location (and change of location) from the medical history, rebound and RLQ tenderness on the physical examination, and leukocytosis. Other commonly included variables are nausea, vomiting, male sex, fever, and rigidity. Decision rules do not vary dramatically between women and men with abdominal pain.

IMPROVEMENTS IN THE DATA PRESENTED IN THE ORIGINAL PUBLICATION

A series of letters to the editor prompted by the original publication pointed out some ways the presentation of the data could be improved.^{5,6} The numbers reported for the sensitivities and specificities did not match the reported LRs; this error was explained⁷ and corrected in Table 5-3 of the original publication.

CHANGES IN THE REFERENCE STANDARD

There were no changes in the reference standard; appendicitis is still a histologically proven diagnosis, and the absence of appendicitis is still a clinical diagnosis (ie, no surgery after adequate follow-up). A recent systematic review suggests that computed tomography may be more accurate than ultrasonography for identifying patients with appendicitis, but neither test is sufficient to serve as the reference standard.⁸

RESULTS OF LITERATURE REVIEW

The LR and diagnostic odds ratio of the 8 best clinical decision rules are displayed in Table 5-4. The studies with the highest numbers of participants that evaluated rules with the highest diagnostic odds ratios (a measure of overall accuracy) were Alvarado,¹ Fenyö et al,³ and Eskelinen et al.²

Although the approaches of Fenyö et al³ and Eskelinen et al² have a higher diagnostic odds ratio than the Alvarado¹ rule, both have a large number of variables and require multivariate modeling that make them hard to use without a handheld calculator or coding sheet. According to these results, as well as expert opinion expressed by the parallel evaluation of Ohmann et al⁴ of 10 of the studies, the Alvarado¹ clinical prediction rule (Table 5-5) is used by most clinicians who prefer decision rules because it balances accuracy with simplicity of use and familiarity to clinicians.

CLINICAL SCENARIO—RESOLUTION

The patient's presentation is suggestive but not clearly diagnostic of appendicitis. The clinician resorts to Alvarado's¹ clinical decision rule and calculates that the patient has 7 of 10 possible points, a positive test result with an LR of 3.1. The patient was referred for surgery, which revealed an inflamed appendix.

APPENDICITIS—MAKE THE DIAGNOSIS

PRIOR PROBABILITY

The incidence of appendicitis among emergency patients with abdominal pain is up to 25% for patients younger than 60 years. For those older than 60 years, the incidence is up to 5%.

POPULATIONS FOR WHOM APPENDICITIS SHOULD BE CONSIDERED

- All patients with abdominal pain.

DETECTING THE LIKELIHOOD OF APPENDICITIS AMONG EMERGENCY PATIENTS WITH ABDOMINAL PAIN IN THE RLQ

The Alvarado¹ model is recommended as the most user-friendly while being among the most powerful. The details of the clinical decision rule are displayed in Table 5-6. Note that the MANTRELS mnemonic is helpful in that it is easy to remember and is organized according to medical history, physical examination, and laboratory data.

Table 5-6 Operating Characteristics of the Alvarado Model

Test	Sensitivity	Specificity	LR+ (95% CI)	LR- (95% CI)
Alvarado score (≥7 is positive)	0.81	0.74	3.1 (1.9–5.0)	0.26 (0.19–0.35)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

REFERENCE STANDARD

Histologically proven diagnosis or no surgery after adequate follow-up (which allows the inference that appendicitis was not present).

REFERENCES FOR THE UPDATE

1. Alvarado A. A practical score for the early diagnosis of acute appendicitis. *Ann Emerg Med.* 1986;15(5):557-564.^a
2. Eskelinen M, Ikonen J, Lipponen P. Sex-specific diagnostic scores for acute appendicitis. *Scand J Gastroenterol.* 1994;29(1):59-66.^a
3. Fenyo G, Lindberg G, Blind P, Enochsson L, Oberg A. Diagnostic decision support in suspected acute appendicitis: validation of a simplified scoring system. *Eur J Surg.* 1997;163(11):831-838.^a
4. Ohmann C, Yang O, Franke C, and the Abdominal Pain Study Group. Diagnostic scores for acute appendicitis. *Eur J Surg.* 1995;161(4):273-281.^a
5. Witt K, Mäkelä M, Olsen O. Likelihood ratios to determine "Does this patient have appendicitis?": comment and clarification. *JAMA.* 1997; 278(10):819.
6. Kloving A. Likelihood ratios to determine "Does this patient have appendicitis?": comment and clarification. *JAMA.* 1997;278(10):819.
7. Wagner JM. Likelihood ratios to determine "Does this patient have appendicitis?": comment and clarification. *JAMA.* 1997;278(10):819-820.
8. Terasawa T, Blackmore CC, Bent S, Kohlwes RJ. Systematic review: computed tomography and ultrasonography to detect acute appendicitis in adults and adolescents. *Ann Intern Med.* 2004;141(7):537-546.
9. Christian F, Christian GP. A simple scoring system to reduce the negative appendicectomy rate. *Ann R Coll Surg Engl.* 1992;74(4):281-285.
10. Izbicki JR, Wolfram TK, Dietmar KW, et al. Accurate diagnosis of acute appendicitis: a retrospective and prospective analysis of 686 patients. *Eur J Surg.* 1992;158(4):227-231.
11. Kalan M, Talbot D, Cunliffe WJ, Rich AJ. Evaluation of the modified Alvarado score in the diagnosis of acute appendicitis: a prospective study. *Ann R Coll Surg Engl.* 1994;76(6):418-419.
12. Ramirez JM, Deus J. Practical score to aid decision making in doubtful cases or appendicitis. *Br J Surg.* 1994;81(5):680-683.
13. Saidi RF, Ghasemi M. Role of Alvarado score in diagnosis and treatment of suspected acute appendicitis. *Am J Emerg Med.* 2000;18(2):230-231.

^aFor the Evidence to Support the Update for this topic, see <http://www.JAMAEvidence.com>.

This page intentionally left blank

EVIDENCE TO SUPPORT THE UPDATE:

Appendicitis, Adult

TITLE A Practical Score for the Early Diagnosis of Acute Appendicitis.

AUTHOR Alvarado A.

CITATION *Ann Emerg Med.* 1986;15(5):557-564.

QUESTION Can the negative appendectomy rate be reduced without increasing the risk of perforation by using a practical score?

DESIGN Retrospective chart review to derive a decision rule based on bayesian statistics.

SETTING One Philadelphia hospital.

PATIENTS Three hundred five patients hospitalized from January 1975 to December 1976 with abdominal pain suggestive of appendicitis.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

Two-by-two tables were constructed for each clinical characteristic found with the chart review. The 8 most accurate characteristics were used in the clinical decision rule ([Table 5-7](#)), making it one of the simplest rules available.

MAIN OUTCOME MEASURES

Appendicitis was diagnosed when pathologically proven. No appendicitis was defined as a normal appendix discovered at operation or resolution of pain without surgery. The length of follow-up of the nonsurgical patients was not defined.

MAIN RESULTS

See [Table 5-8](#).

CONCLUSIONS

LEVEL OF EVIDENCE Level 4.

STRENGTHS Although the level of evidence of this study is low, this study is noteworthy because of its early appearance in the literature and its wide acceptance.

Table 5-7 The Alvarado Scoring System (MANTRELS Mnemonic)

Variable	Score ^a
Migration	1
Anorexia-acetone	1
Nausea-vomiting	1
Tenderness in RLQ	2
Rebound pain	1
Elevation of temperature	1
Leukocytosis	2
Shift to the left	1
Total	10

Abbreviation: RLQ, right lower quadrant.

^aA score of 7 or higher requires operation.

Table 5-8 Likelihood Ratios for the Alvarado Score

Test	Sensitivity	Specificity	LR+ (95% CI)	LR- (95% CI)	DOR (95% CI)
Alvarado score (≥ 7)	0.81	0.74	3.1 (2.2-5.0)	0.25 (0.21-0.35)	12 (6.0-25)

Abbreviations: CI, confidence interval; DOR, diagnostic odds ratio; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

WEAKNESSES This study's retrospective design is a weakness, but it has much face validity. This study has been validated in several later studies.¹⁻³

REFERENCES FOR THE EVIDENCE

1. Saidi HS, Chavda SK. Use of a modified Alvarado score in the diagnosis of acute appendicitis. *East Afr Med J.* 2003;80(8):411-414.
2. Saidi RF, Ghasemi M. Role of Alvarado score in diagnosis and treatment of suspected acute appendicitis. *Am J Emerg Med.* 2000;18(2):230-231.
3. Kalan M, Talbot D, Cunliffe WJ, Rich AJ. Evaluation of the modified Alvarado score in the diagnosis of acute appendicitis: a prospective study. *Ann R Coll Surg Engl.* 1994;76(6):418-419.

Reviewed by James Wagner, MD

TITLE Sex-Specific Diagnostic Scores for Acute Appendicitis.

AUTHORS Eskelinen M, Ikonen J, Lipponen P.

CITATION *Scand J Gastroenterol*. 1994;29(1):59-66.

QUESTION Can the diagnosis of acute appendicitis in women and men with acute abdominal pain be improved by using computer-based diagnostic scores?

DESIGN This was prospective derivation of a clinical decision rule from a convenience sample of patients with a standardized data collection sheet. The rule was derived using logistic stepwise multivariate regression analysis.

SETTING Two Finnish hospitals.

PATIENTS A total of 1333 patients with acute abdominal pain of less than 7 days' duration who were admitted to one of the 2 study hospitals during a 6-year period.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

Separate clinical decision rules were derived for men and women. The scoring system for the clinical decision rule for the men involved 7 indicators (Table 5-9); the scoring system for women involved 5 (Table 5-10). Computers were used in this study to take the data entered from a standardized form and calculate the discriminate score (DS) and compare it with the diagnostic standard.

MAIN OUTCOME MEASURES

The DS was compared with a diagnostic standard: appendicitis was defined as that pathologically proven. No appendicitis was defined as a normal appendix discovered at operation or resolution of pain without surgery. The length of follow-up of the nonsurgical patients was not defined.

MAIN RESULTS

Several computer models and cutoffs were analyzed; the cutoff for the results reported in Table 5-11 was as follows. Patients with DS values below -2.00 should not have surgery,

patients with a DS above -0.48 should have surgery, and patients with DS values between -2.00 and -0.48 were considered nondefined. That is, they required follow-up before the decision to operate or not was made.

Table 5-9 The Scoring System for Men

Variable	Indicator	Score
Constant		-7.69
Previous abdominal surgery	Yes	0
	No	1.88
Pain at diagnosis	RLQ	1.3
	Other	0
Fever	$\geq 37.1^{\circ}\text{C}$	1.05
	$< 37.1^{\circ}\text{C}$	0
Tenderness	RLQ	1.97
	Other	0
Rebound tenderness	Yes	1.61
	No	0
Guarding	Yes	1.14
	No	0
Rigidity	Yes	1.43
	No	0

Abbreviation: RLQ, right lower quadrant.

Table 5-10 The Scoring System for Women

Variable	Indicator	Score
Constant		-7.22
Pain at diagnosis	RLQ	1.33
	Other	0
Tenderness	RLQ	2.98
	Other	0
Renal tenderness	Yes	0
	No	0.88
Guarding	Yes	2.08
	No	0
Rigidity	Yes	2.45
	No	0

Abbreviation: RLQ, right lower quadrant.

Table 5-11 Results of the Study

Test	Sensitivity	Specificity	LR+ (95% CI)	LR- (95% CI)	DOR (95% CI)
Men	0.95	0.89	8.6	0.05	163
Women	0.93	0.92	12	0.07	163
Total	0.94	0.91	10 (9.3-12)	0.06 (0.05-0.10)	164 (93-287)

Abbreviations: CI, confidence interval; DOR, diagnostic odds ratio; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

CONCLUSIONS

LEVEL OF EVIDENCE Levels 2 and 3.

STRENGTHS This study had a large sample size. A standardized form was used to record all clinical data.

LIMITATIONS This study used convenience sampling of patients that was described by the authors as “although not entirely consecutive, the series was collected by the same surgeon with regard to data collection and comprised a representative and unselected sample.”

The clinical decision rule was not validated in the original report. It has since been validated on an even larger sample of patients¹; the results were less impressive but still indicated significant power of this scoring system.

REFERENCE FOR THE EVIDENCE

1. Zielke A, Sitter H, Rampp T, Bohrer T, Rothmund M. Clinical decision-making, ultrasonography, and scores for evaluation of suspected acute appendicitis. *World J Surg.* 2001;25(5):578-584.

Reviewed by James M. Wagner, MD

TITLE Diagnostic Decision Support in Suspected Acute Appendicitis: Validation of a Simplified Scoring System.

AUTHORS Fenyo G, Lindberg G, Blind P, Enochsson L, Oberg A.

CITATION *Eur J Surg.* 1997;163(11):831-838.

QUESTION Can a scoring system for the diagnosis of appendicitis be validated?

DESIGN Prospective validation of previously derived decision rule.

SETTING One Swedish county district hospital and 1 university hospital. One center accounted for 86% of the patients. The authors state that “virtually all” patients in that center were enrolled. At the second center that enrolled a minority of patients, only 60% of the potentially eligible patients were enrolled.

PATIENTS A total of 1167 patients with suspected appendicitis, that is, patients who had not previously had an appendectomy and who presented with pain, tenderness, or both in the right lower quadrant (RLQ).

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

The scoring system validated in this study has been used routinely in the 2 hospitals involved in this study. A pocket chart with 10 variables and their associated scores was carried by

clinicians ([Table 5-12](#)); scores suggest “consider operation,” “observe with repeated examinations,” or “observation or discharge of the patient.”

MAIN OUTCOME MEASURES

The diagnostic standard was positive for histologically proven appendicitis and negative for histologically disproven appendicitis or the resolution of symptoms without operation. A positive result was defined as a score of -2 or more.

MAIN RESULTS

See [Tables 5-13](#), [5-14](#), [5-15](#), and [5-16](#).

Table 5-12 Scoring System

Variable	Indicator	Score
Constant (apply to all patients as the starting point)		-10
Sex	Male	+8
	Female	-8
White blood cell count (per μL)	≥ 14000	+15
	9000-13900	+2
	≤ 8900	-15
Duration of pain, h	≤ 24	+3
	24-48	0
	≥ 48	-12
Progression of pain	Yes	+3
	No	-4
Relocation of pain	Yes	+7
	No	-9
Vomiting	Yes	+7
	No	-5
Aggravation by coughing	Yes	+4
	No	-11
Rebound tenderness	Yes	+5
	No	-10
Rigidity	Yes	+15
	No	-4
Tenderness outside RLQ	Yes	-6
	No	+4

Abbreviation: RLQ, right lower quadrant.

Table 5-13 Probability of Appendicitis According to Score

	Probability of Appendicitis	Recommended Strategy
-2 or greater	≥ 0.45	Consider operation
-3 to -16	0.44-0.17	Observe with repeated examination
-17 or less	≤ 0.17	Observe or discharge to home

Table 5-14 Univariate Results

	LR+ (95% CI) ^a	LR- (95% CI)
Male (n = 531)	1.6 (1.4-1.7)	
Female (n = 636)	0.66 (0.58-0.75)	
White blood cell count (per μ L)		
\geq 14000	3.1 (2.5-3.8)	
9000-13900	1.3 (1.1-1.5)	
\leq 8900	0.16 (0.11-0.22)	
Duration of pain, h		
\leq 24	1.4 (1.2-1.5)	
24-48	1.1 (0.78-1.5)	
\geq 48	0.39 (0.30-0.51)	
Progression of pain	1.3 (1.2-1.5)	0.57 (0.47-0.68)
Relocation of pain	2.2 (2.0-2.6)	0.46 (0.40-0.54)
Vomiting	1.7 (1.4-1.9)	0.74 (0.66-0.82)
Aggravation by coughing	1.5 (1.4-1.6)	0.35 (0.28-0.45)
Rebound tenderness	1.8 (1.6-1.9)	0.38 (0.31-0.46)
Rigidity	2.8 (2.3-3.5)	0.70 (0.64-0.76)
Tenderness outside RLQ	0.67 (0.58-0.77)	1.4 (1.3-1.6)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio; RLQ, right lower quadrant.

^aSerial likelihood ratios are reported for results on an ordinal scale.

The overall accuracy from the receiver operating characteristic curve for the multivariate model was 0.89 for the center with consecutive enrollment vs 0.83 for the center that did not capture all eligible patients.

Using the model with a cut point of -2 or greater, as presented by the authors, produces a likelihood ratio (LR) of 5.6 (95% confidence interval [CI], 4.6-6.8) for a score of -2 or greater; when the score is less than -2, the LR is 0.31 (95% CI, 0.26-0.37).

CONCLUSIONS

LEVEL OF EVIDENCE

Level 3.

STRENGTHS This study had a large sample size, and it used a standardized form on which all clinical data were

Table 5-16 Serial LRs for the Recommended Cut Points

Score	LR (95% CI) for Appendicitis
-2 or greater	6.0 (4.9-7.2)
-3 to -16	0.59 (0.42-0.84)
-17 or less	0.26 (0.21-0.32)

Abbreviations: CI, confidence interval; LR, likelihood ratio.

recorded. The study reports a clinical decision rule that was derived and validated with good technique.

LIMITATIONS This study reported the results from 2 centers. Most patients enrolled in the study were reported as being “consecutive.” It is difficult to assess the effect of non-consecutive enrollment at the second hospital, which accounted for approximately 15% of the patients. However, because the overall accuracy of the score at the hospital with nonconsecutive patients was slightly worse (83% vs 89%), it is likely that the findings *underestimate* the true accuracy.

None of the individual clinical findings had values distinctly different from 1, allowing the clinician the opportunity to reliably rule in appendicitis. A variable with good measurement properties that decreased the likelihood of appendicitis was a normal white blood cell count ($<8900/\mu\text{L}$), with an LR of 0.16.

The investigators’ goal was to compare the predicted probability of a score by using the clinical variables with the actual outcomes. The authors recommend a cut point of -2 or greater as suggesting the need for surgery and a value -17 or less as appropriate for discharging a patient home without observation and a repeated examination. The data are presented in a fashion that allows clinicians to calculate the LR for the 3 levels of appendicitis scores. The serial LRs perform better than the dichotomous LR and match the clinical recommendations for the different levels of LRs.

Reviewed by James M. Wagner, MD

Table 5-15 Multivariate Results for a Score of -2 or Greater as a Function of Sex

Test	Sensitivity	Specificity	LR+ (95% CI)	LR- (95% CI)	DOR (95% CI)
Men	0.80	0.79	3.8 (3.1-4.8)	0.24 (0.18-0.31)	16 (10-24)
Women	0.61	0.93	8.8 (6.2-12)	0.42 (0.34-0.51)	21 (13-34)

Abbreviations: CI, confidence interval; DOR, diagnostic odds ratio; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

TITLE Diagnostic Scores for Acute Appendicitis.

AUTHORS Ohmann C, Yang O, Franke C, and the Abdominal Pain Study Group.

CITATION Eur J Surg. 1995;161(4):273-281.

QUESTION Which of 10 predictive scores used in the diagnosis of acute appendicitis are most valuable?

DESIGN Multicenter prospective collection of 25 variables from the medical history and 20 from the physical examination.

SETTING Six German hospitals.

PATIENTS A total of 1254 consecutive patients with acute abdominal pain of less than 1 week's duration, excluding patients with trauma or postsurgical pain.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

The 10 tests evaluated were the Lindberg et al,¹ Eskelinen et al,² Alvarado,³ Fenyo,⁴ Izicki et al,⁵ Christian and Christian,⁶ van Way et al,⁷ Teicher et al,⁸ Arnbjornsson,⁹ and De Dombal¹⁰ scores. These scores were grouped according to the population in which the score was intended for use (Table 5-17). Group A (Lindberg and Eskelinen scores) contained scores intended for use with a population with acute abdominal pain. Group B (Alvarado, Fenyo, Izicki, and Christian scores) scores were intended for use on patients suspected of having appendicitis. Group C (van Way, Teicher, and Arnbjornsson scores) scores were derived from patients who had appendicitis. Group D (De Dombal) were scores intended for use with any patient with abdominal pain, but the diagnosis of interest was "nonspecific abdominal pain"; that is, instead of diagnosing appendicitis, it "diagnoses" pain in which surgical intervention is unnecessary.

Appendicitis was diagnosed when confirmed by pathology specimens. "No appendicitis" was defined as a normal appendix discovered at operation or resolution of pain without surgery. Patients not receiving operation were followed by telephone interview for a length of time that was undefined.

MAIN OUTCOME MEASURES

The collected data were used to calculate the 10 predictive scores. Patients were retrospectively assigned to outcomes that would have resulted from management that followed the score's suggestion. A 15% negative appendectomy rate is accepted as standard of care, so a score that resulted in assignments of patients leading to more than 15% was deemed unacceptable. This was done to define a minimally acceptable performance of a score.

There were 4 such criteria used for comparing outcomes from the 10 scores: (1) "[i]nitial negative appendectomy [sic] rate" (defined as proportion of patients who did not have acute appendicitis who were assigned to the operation group), (2) "[p]otential perforation rate" (defined as proportion of patients with acute appendicitis not assigned to the operation group), (3) "[i]nitial missed perforation rate" (defined as proportion of patients with perforated appendicitis not assigned to the operation group), and (4) "[m]issed appendicitis rate" (defined as the proportion of patients with acute appendicitis who were assigned to the exclusion group).

MAIN RESULTS

The prevalence of appendicitis in this study was 17%.

None of the tested scores fulfilled the criterion for an acceptable score since all had high missed appendicitis rates (Table 5-17). By calculation of sensitivity and specificity from the data provided in the study, it appears that there was a

Table 5-17 Clinical Outcomes That Would Have Accrued From Management Guided by the Scores

Group	Score	Initial Negative Appendectomy Rate, %	Potential Perforation Rate, %	Initial Missed Perforation Rate, %	Missed Appendicitis Rate, %
A	Lindberg ¹	53	53	56	30
	Eskelinen ²	30	30	63	61
B	Alvarado ³	29	29	76	65
	Fenyo ⁴	25	25	76	57
	Izicki ⁵	47	49	38	17
	Christian ⁶	42	42	69	57
C	van Way ⁷	12	12	27	15
	Teicher ⁸	40	11	21	... ^a
	Arnbjornsson ⁹	13	13	10	15
D	De Dombal ¹⁰	39	39	82	76
Standard		15	35	15	5
Clinicians	Initial	52	11	16	11
	Actual	8	33	44	0

^aEllipsis indicates data not available.

deflation of the sensitivity and inflation of the specificity in this study compared with the other attempts at validating these data, which suggests a referral bias in this or the analyzed studies. The initial diagnostic accuracy of the clinicians also did not perform at a minimally acceptable level.

Despite the disappointing performance of the scores, the investigators reported the performance of each score compared with one another. After applying each score to the entire database of patients presenting with abdominal pain (not just the populations for which the scores were intended or derived), the investigators recommended further testing of 2 scores in patients with abdominal pain or suspected of having appendicitis: the Alvarado and Eskelinen scores. The investigators also report the variables used most frequently: site and duration of pain, site of tenderness, rebound tenderness, and white blood cell count.

CONCLUSIONS

LEVEL OF EVIDENCE

Level 1.

STRENGTHS This was an unprecedented, prospective, multicenter study that compared 10 clinical prediction rules for appendicitis on a single, large population at several German hospitals. The methods were fairly well described, and the criteria against which all rules were compared seemed thoughtful.

WEAKNESSES The clinical database used in this study contained most, but not all, of the clinical criteria used in each clinical prediction rule. This was a complex study, and its description and tables were somewhat confusing. The division of studies into groups A and B was based on subjective data and seemed arbitrary. The included studies typically did not explain the difference between “acute abdominal pain” and “suspected appendicitis.” The data reported most thoroughly were those analyzed by groups; perhaps the most useful data presented were in the text, where groups A and B were compared on the same population.

The performance of each of the rules was surprising. The investigators provide several suggestions to explain the poor performance, mainly positive bias of the original studies and geographic differences in patient characteristics. Beyond what was explored in the discussion, the difference between the initial and actual treatment plan may explain the poor performance of the scores. Given time, the patient may lose symptoms or signs and therefore exhibit a lower score than initially recorded.

Nonetheless, it appears that the Alvarado and Eskelinen scores are the best clinical decision rules for appendicitis in patients with abdominal pain. This judgment is based on the practicality of the score and the use of the most powerful individual findings. In addition, the Alvarado rule is the oldest rule most familiar to clinicians and is the simplest to implement.

REFERENCES FOR THE EVIDENCE

1. Fenyo G, Lindberg G, Blind P, Enochsson L, Oberg A. Diagnostic decision support in suspected acute appendicitis: validation of a simplified scoring system. *Eur J Surg*. 1997;163(11):831-838.
2. Eskelinen M, Ikonen J, Lippinen P. A computer-based diagnostic score to aid in diagnosis of acute appendicitis: a prospective study of 1333 patients with acute abdominal pain. *Theor Surg*. 1992;7:86-90.
3. Alvarado A. A practical score for the early diagnosis of acute appendicitis. *Ann Emerg Med*. 1986;15(5):557-564.
4. Fenyo G. Routine use of a scoring system for decision-making in suspected acute appendicitis in adults. *Acta Chir Scand*. 1987;153(9):545-551.
5. Izbicki JR, Wolfram TK, Dietmar KW, et al. Accurate diagnosis of acute appendicitis: a retrospective and prospective analysis of 686 patients. *Eur J Surg*. 1992;158(4):227-231.
6. Christian F, Christian GP. A simple scoring system to reduce the negative appendicectomy rate. *Ann R Coll Surg Engl*. 1992;74(4):281-285.
7. van Way C, Murphy J, Dunn E, Elerdeing S. A feasibility study of computer aided diagnosis in appendicitis. *Surg Gynecol Obstet*. 1982;155(5):685-688.
8. Teicher I, Landa B, Cohen M, Kabnick LS, Wise L. Scoring system to aid in the diagnosis of appendicitis. *Ann Surg*. 1983;198(6):753-759.
9. Arnbjornsson E. Scoring system for computer-aided diagnosis of acute appendicitis. *Ann Chir Gynaecol*. 1985;74(4):159-166.
10. De Dombal FT. *Diagnosis of Acute Abdominal Pain*. 2nd ed. New York, NY: Churchill Livingstone; 1991:1-259.

Reviewed by James Wagner, MD

CHAPTER 6

Does This Patient Have Ascites?

How to Divine Fluid in the Abdomen

John W. Williams, Jr, MD

David L. Simel, MD, MHS

CLINICAL SCENARIOS

In each of the following cases, the clinician will need to determine whether the patient has ascites.

CASE 1 A 44-year-old man with cirrhosis is admitted with fever but has no obvious source of infection.

CASE 2 A 57-year-old woman presents with an adnexal mass and recent weight gain but otherwise feels well.

CASE 3 A 65-year-old man with a history of myocardial infarction is admitted for decreased exercise tolerance, increased abdominal girth, and ankle edema.

WHY IS THIS AN IMPORTANT QUESTION TO ANSWER WITH A CLINICAL EXAMINATION?

Free fluid in the abdominal cavity is *ascites*. Ascites may have important diagnostic, prognostic, and therapeutic implications. When clinically detectable, ascites may indicate underlying heart failure, liver disease, nephrotic syndrome, or malignancy. In patients with liver disease, ascites has prognostic significance because operative mortality is increased and overall survival is decreased; ascites may also signal metastases in patients with malignancy.¹ Although patients with small amounts of ascites do not generally require specific therapy, patients with larger amounts of ascites may require intervention to relieve symptoms caused by their distended abdomen. Furthermore, the degree of ascites is useful in monitoring the efficacy of treatment for the underlying condition that caused it (eg, monitoring response to chemotherapy for malignancy).

The 3 clinical scenarios are specific examples of why ascites detection is clinically important. For example, ascites detection in the first patient may lead to the diagnosis of spontaneous bacterial peritonitis as the source of the patient's fever. If ascites is found by clinical examination, the physician may be able to proceed directly to abdominal paracentesis without pausing for imaging procedures. In the second patient, the presence of ascites would heighten the clinician's suspicion of ovarian carcinoma with peritoneal metastases, implying a more advanced stage and poorer prognosis. In the third patient, the finding of ascites may trigger the physician's consideration of diagnostic possibilities other than severe left-sided congestive heart failure, such as a pericardial effusion causing marked signs of right-sided heart failure. Clearly, clinical determination of the presence or absence of ascites has the advantages of speed, convenience, and cost savings on diagnostic imaging.

It is easy to identify large volumes of ascites clinically, but smaller amounts of ascites are not as obvious. When diagnostic confirmation is necessary, paracentesis is the definitive

test, although less invasive radiographic procedures are ordinarily used to corroborate the clinician's suspicion. Ultrasonography can detect as little as 100 mL of abdominal fluid and is considered the gold standard for diagnosing ascites.^{2,3} Abdominal computed tomography can also detect small amounts of fluid but is more expensive. Unfortunately, there are no general guidelines for correlating small amounts of ascites observed on ultrasonographic examination or computed tomography with pathophysiologic conditions.

The reference standard for ascites is fluid aspiration by paracentesis and fluid visualization by ultrasonography or computed tomography.

Pathophysiology of Ascites

An understanding of the pathophysiologic basis for ascites facilitates assessment of each patient's risk by alerting the examiner to conditions disrupting normal physiology ([Table 6-1](#)). Under physiologic conditions, intravascular and extravascular hydrostatic and colloid osmotic pressures are balanced, preventing accumulation of extravascular fluid.⁵ Any process disrupting this balance may precipitate ascites. For example, fibrotic constriction of the hepatic sinusoids secondary to alcoholic cirrhosis leads to increased venous hydrostatic pressure and, ultimately, to ascites by forcing lymphatic drainage into the abdomen through the hepatic capsule.¹ Cirrhotic patients with ascites show avid renal retention of sodium and water, which is an important mechanism for continued ascites formation.⁶ A second, less important mechanism for ascites formation is a loss of osmotic pressure because of inadequate protein synthesis (eg, malnutrition, liver disease) or protein wasting (eg, the nephrotic syndrome). Because of protein loss, transudative fluid moves from the

intravascular space into the abdominal extravascular space to balance hydrostatic and osmotic forces. Finally, infection or malignancy in the peritoneum may produce inflammatory exudates or malignant effusions in the abdominal extravascular space faster than it can be absorbed intravascularly.

How to Elicit the Symptoms and Signs of Ascites

A complete evaluation for ascites includes a focused medical history and physical examination. The examiner should ask about recent ankle edema, weight gain, or change in abdominal girth. Other potentially important items are a history of liver disease or congestive heart failure. A focused physical examination for ascites includes (1) inspection for bulging flanks, (2) percussion for flank dullness, (3) a test for shifting dullness, and (4) a test for a fluid wave.

Bulging flanks occur when the weight of abdominal free fluid is sufficient to push the flanks outward. However, it is sometimes difficult to distinguish bulging flanks caused by ascites from bulging flanks caused by obesity. One method for discriminating between the 2 is to test for flank dullness. With the patient recumbent, gas-filled loops of bowel will characteristically float on top of ascites, making the percussion note tympanitic at the umbilicus and dull beyond the fluid meniscus into the flanks ([Figure 6-1A](#)). The examiner can confirm this pattern by progressively percussing the abdomen, beginning at the umbilicus and moving toward the flanks, listening for the transition from tympany to dullness when the meniscus is reached.⁷ Having identified and marked the transition between tympany and dullness, further evidence for ascites can be obtained by testing for shifting dullness. This is done by rolling the patient away from the examiner and repeating the percussion. With ascites, the area of dullness shifts to the dependent side, and the area of tympany shifts toward the top ([Figure 6-1B](#)).

Another potentially useful method for detecting ascites is testing for a fluid wave. The test is performed by having the patient, or an assistant, place the medial edges of both hands firmly down the midline of the abdomen to block transmission of a wave through subcutaneous fat ([Figure 6-2](#)). The examiner taps one flank sharply while using the fingertips to feel for an impulse on the opposite flank. When ascites is present, an impulse may be felt in the receiving hand after a barely perceptible lag.

Two additional maneuvers, the puddle sign and auscultatory percussion, cannot currently be recommended. The puddle sign was initially advocated because of its purported high sensitivity.⁸ However, it is infrequently used now because it is difficult to perform properly and has low sensitivity (43%-55%).^{9,10} A method of auscultatory percussion was described by Guarino,¹¹ but its precision and accuracy have not yet been reported. After voiding, the patient sits or stands so that free fluid gravitates to the pelvis, and the examiner places a stethoscope in the midline, immediately above the pubic crest. Finger-flicking percus-

Table 6-1 Pathophysiologic Classification of Ascites^a

I. Elevated hydrostatic pressure
A. Cirrhosis
B. Congestive heart failure
C. Constrictive pericarditis
D. Inferior vena cava obstruction
E. Hepatic vein obstruction (Budd-Chiari syndrome)
II. Decreased osmotic pressure
A. Nephrotic syndrome
B. Protein-losing enteropathy
C. Malnutrition
D. Cirrhosis or hepatic insufficiency
III. Fluid production exceeding resorptive capacity
A. Infections
1. Bacterial
2. Tuberculosis
3. Parasitic
B. Neoplasms

^aAdapted from Bender.⁴

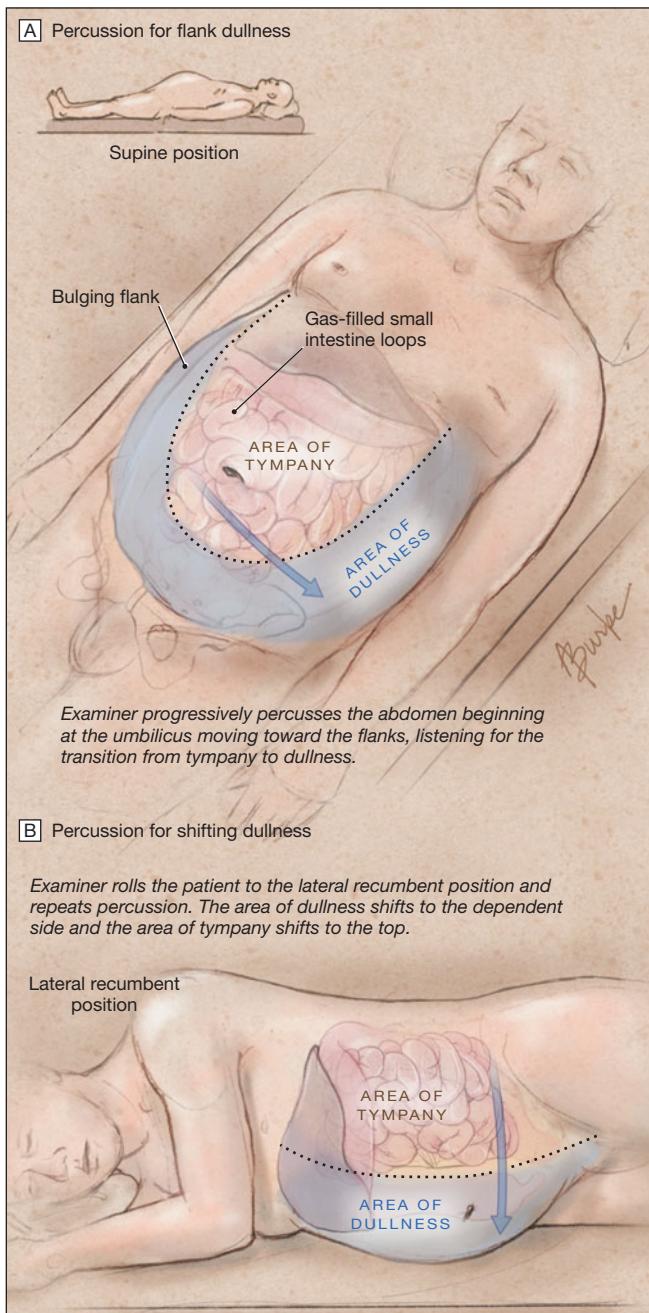


Figure 6-1 Percussion Techniques for Detecting Ascites

sion is performed along radial spokes from the subcostal margin downward toward the pelvis. The percussion note is initially dull but changes sharply to a loud note at the border of increased pelvic density. In the absence of ascites, the border is approximately 4.5 cm above the pelvic crest (the pelvic baseline). In patients with ascites, free fluid raises the demarcating border clearly above the pelvic baseline. When the patient is supine, this clear line of demarcation is obliterated because the free fluid gravitates to the flanks.

Although most of the physical examination for ascites should focus on the abdomen, extra-abdominal signs may

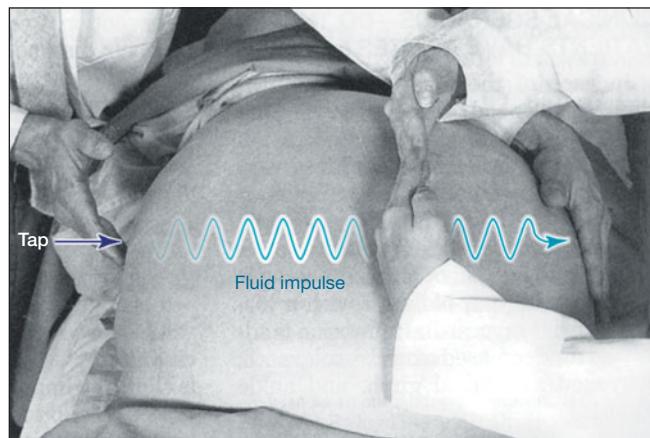


Figure 6-2 Testing for a Fluid Wave

provide evidence for conditions associated with ascites. Physical findings that may be useful by their presence or absence include evidence of liver disease (eg, jaundice, spider angiomas) or heart disease (eg, cardiac gallop).

ACCURACY OF HISTORY AND SYMPTOMS FOR ASCITES

We examined the effect of medical history items on the probability of ascites in male veteran inpatients (Table 6-2).⁹ Medical histories, obtained by internal medicine house staff, were compared with reference standard abdominal ultrasonographic findings. Positive histories of hepatitis or heart failure generated likelihood ratios (LRs) of 3.2 and 2.0, respectively. However, alcoholism (positive LR [LR+], 1.4) or a history of carcinoma (LR+, 0.91) had little effect on the odds of ascites.

Other questions about the patient's present illness may be even more useful. In this same study, the patient's symptoms of increased abdominal girth, weight gain, or ankle edema gave LR+ values of 4.2, 3.2, and 2.8, respectively. The absence of increased abdominal girth (negative LR [LR-], 0.17) or ankle swelling (LR-, 0.10) decreased appreciably the diagnostic likelihood of ascites. For example, in a patient with a low pretest probability of ascites (<20%), the absence of recent ankle edema decreases the probability of ascites to less than 2.5%. Clearly, the patient's medical history and current symptoms are valuable for at least 2 reasons. First, certain items may suggest the presence or absence of ascites. Second, in patients suspected of having ascites, a focused physical examination for ascites is needed. The clinical history distinguishes patients with high and low probabilities for ascites. Ascites is unlikely when patients report no increase in abdominal girth, and ascites is very unlikely in male patients who report no history of recent ankle swelling.

PRECISION OF THE SIGNS FOR ASCITES

Six gastroenterologists examined 50 hospitalized alcoholic patients for the presence or absence of ascites. Their overall

agreement was good (intraclass correlation, 0.75), and it was excellent among senior physicians (0.95).¹² In another study, 90 veteran inpatients with evidence of liver disease were examined by 3 internists for 4 signs of ascites. For each sign, there was good agreement: presence or absence of abdominal distention (86%), bulging flanks (79%), shifting dullness (78%), and detection of prominent fluid waves (76%).¹³ There is good agreement among physicians on the presence or absence of traditional signs for ascites.

ACCURACY OF SIGNS FOR ASCITES

Three investigations have compared physical examination findings for ascites with findings from reference standard abdominal ultrasonographic examinations.^{9,10,13} Despite the various levels of training (internal medicine interns to staff gastroenterologists), the results were similar in each study (Table 6-3). There was no single sign for ascites that was both sensitive and specific. However, flank dullness ($\geq 80\%$) and bulging flanks ($\geq 72\%$) were sensitive in all studies. Shifting dullness had a high sensitivity ($\geq 83\%$) in 2 investigations. The puddle sign, purported to be the most sensitive test for ascites, performed poorly, yielding at best a sensitivity of 55%. The absence of a fluid wave was the only sign with a high specificity (82%-92%) across all studies. Shifting dullness was highly specific in only 1 study⁹; results may be inconsistent because of

differences in the study populations (general medical vs patients with liver disease). To date, no investigator has studied how to best use these signs in combination.

The clinician must know the pretest probability or prevalence of a disease to apply sensitivity and specificity data to an individual patient. The LRs for the physical examination signs from the 3 studies are displayed in Table 6-4. We combined the study results according to the number of unique patients in each study to yield pooled sensitivity, specificity, and LRs (Table 6-5). The finding of a fluid wave, shifting dullness, or peripheral edema increased the likelihood of ascites the most. The absence of bulging flanks, flank dullness, shifting dullness, or peripheral edema decreased the likelihood of ascites the most.

Finally, is the whole greater than the sum of the parts? Is an examiner's overall clinical impression more accurate than individual signs or symptoms of ascites? Two studies evaluated the accuracy of the overall clinical assessment for ascites. In the study by Cattau et al¹⁰ of patients who were referred because their physicians were unsure about the presence of ascites, the examiners correctly determined the presence or absence of ascites in only 56% of patients in this most difficult clinical scenario. In the study by Simel et al,⁹ examiners categorized the probability of ascites as high, intermediate, or low. Examiners at all levels of training (intern through chief resident) were accurate when assigning a high probability of ascites (LR+, 38-83) but were less accurate at low probability of ascites (LR-, 0.77-0.87). Apparently, a high probability of ascites in hospitalized patients was sufficient to make the diagnosis, but a low probability was not enough to rule out ascites. This rule may not apply for outpatients.

The following suggestions should guide clinical teaching and performance of the clinical examination for detecting ascites:

1. The most useful findings for ruling out ascites are no history of ankle swelling or increased abdominal girth and the inability to demonstrate bulging flanks, flank dullness, or shifting dullness.
2. The most powerful findings for making the diagnosis of ascites are a positive fluid wave result, shifting dullness, or peripheral edema.

Table 6-2 Accuracy of the Clinical History^a

Historical Item or Symptom	Sensitivity	Specificity	LR+	LR-
Increased girth	0.87	0.77	4.2	0.17
Recent weight gain	0.67	0.79	3.2	0.42
Hepatitis	0.27	0.92	3.2	0.80
Ankle swelling	0.93	0.66	2.8	0.10
Heart failure	0.47	0.73	2.0	0.73
Alcoholism	0.60	0.58	1.4	0.69
History of carcinoma	0.13	0.85	0.91	1.0

Abbreviations: LR+, positive likelihood ratio; LR-, negative likelihood ratio.

^aAdapted from Simel et al.⁹

Table 6-3 Sensitivity and Specificity of the Physical Examination for Ascites

Sign	Sensitivity			Specificity		
	Cummings et al ¹³	Simel et al ⁹	Cattau et al ¹⁰	Cummings et al ¹³	Simel et al ⁹	Cattau et al ¹⁰
Flank dullness	NA	0.80	0.94	NA	0.69 ^a	0.29
Bulging flanks	0.72	0.93	0.78	0.70	0.54	0.44
Shifting dullness	0.88	0.60	0.83	0.56	0.90 ^a	0.56
Fluid wave	0.53	0.80	0.50	0.90	0.92	0.82
Puddle sign	NA	0.43	0.55	NA	0.83	0.51

Abbreviation: NA, not available.

^aTest for heterogeneity suggests these values are significantly better across studies ($P < .01$).

Table 6-4 Likelihood Ratios for the Physical Examination for Ascites^a

Sign	LR+			LR-		
	Cummings et al ¹³	Simel et al ⁹	Cattau et al ¹⁰	Cummings et al ¹³	Simel et al ⁹	Cattau et al ¹⁰
Bulging flanks	2.4	2.0	1.4	0.4	0.1	0.5
Flank dullness	NA	2.6	1.3	NA	0.3	0.2
Shifting dullness	2.0	5.8	1.9	0.2	0.5	0.4
Fluid wave	5.3	9.6	2.8	0.5	0.2	0.6
Puddle sign	NA	2.6	1.1	NA	0.7	0.9
Peripheral edema	NA	3.8	NA	NA	0.2	NA

Abbreviations: LR+, positive likelihood ratio; LR-, negative likelihood ratio; NA, not available.

^aExaminers were board-certified general internists in the study by Cummings et al,¹³ internal medicine house staff in that by Simel et al,⁹ and staff gastroenterologists in that by Cattau et al.¹⁰

Table 6-5 Pooled Results of Physical Examination Studies

Physical Sign	LR+ (95% CI)	LR- (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
Bulging flanks	2.0 (1.5-2.6)	0.3 (0.2-0.6)	0.81 (0.69-0.93)	0.59 (0.50-0.68)
Flank dullness	2.0 (1.5-2.9)	0.3 (0.1-0.7)	0.84 (0.68-1.00)	0.59 (0.47-0.71)
Shifting dullness	2.7 (1.9-3.9)	0.3 (0.2-0.6)	0.77 (0.64-0.90)	0.72 (0.63-0.81)
Fluid wave	6.0 (3.3-11)	0.4 (0.3-0.6)	0.62 (0.47-0.77)	0.90 (0.84-0.96)
Puddle sign	1.6 (0.8-3.4)	0.8 (0.5-1.2)	0.45 (0.20-0.70)	0.73 (0.61-0.85)

Abbreviations: CI, confidence interval; LR+ positive likelihood ratio; LR-, negative likelihood ratio.

- The puddle sign is difficult to perform, uncomfortable for patients, and not sensitive to small amounts of ascites. It should not be performed.

Author Affiliations at the Time of the Original Publication

Ambulatory Care Service and The Center for Health Services Research in Primary Care, Durham Veterans Administration Medical Center, and the Division of General Internal Medicine, Duke University Medical Center, Durham, North Carolina.

Acknowledgments

This work was supported, in part, by a grant from the A. W. Mellon Foundation, New York, New York.

We wish to thank Jan Irvine, MD, and Valerie Lawrence, MD, for reviewing an earlier draft of this overview.

REFERENCES

- Bender MD, Ockner RK. Ascites. In: Sleisenger MH, Fordtran JS, eds. *Gastrointestinal Diseases and Pathophysiology: Diagnosis and Management*. 4th ed. Philadelphia, PA: WB Saunders Co; 1989:1943-1945.
- Goldberg BB, Goodman GA, Clearfield HR. Evaluation of ascites by ultrasound. *Radiology*. 1970;96(1):15-22.
- Goldberg BB, Clearfield HR, Goodman GA, Morales JO. Ultrasonic determination of ascites. *Arch Intern Med*. 1973;131(2):217-220.
- Bender MD. Diseases of the peritoneum. In: Wyngaarden JB, Smith LH, eds. *Cecil Textbook of Medicine*. 18th ed. Philadelphia, PA: WB Saunders Co; 1988:790-793.
- Guyton AC. *Textbook of Medical Physiology*. 7th ed. Philadelphia, PA: WB Saunders Co; 1983:836.
- Schrier RW. Pathogenesis of sodium and water retention in high-output and low-output cardiac failure, nephrotic syndrome, cirrhosis, and pregnancy (2 parts). *N Engl J Med*. 1988;319(16-17):1065-1072, 1127-1134.
- LeBlond RF, DeGowin RL, Brown DD. DeGowin's *Diagnostic Examination*. 8th ed. New York, NY: McGraw-Hill; 2004:543.
- Lawson JD, Weissbein CA. The puddle sign—an aid in the diagnosis of minimal ascites. *N Engl J Med*. 1959;260(13):652-654.
- Simel DL, Halvorsen RA, Feussner JR. Quantitating bedside diagnosis: clinical evaluation of ascites. *J Gen Intern Med*. 1988;3(5):423-428.
- Cattau EL Jr, Benjamin SB, Knuff TE, Castell DO. The accuracy of the physical examination in the diagnosis of suspected ascites. *JAMA*. 1982;247(8):1164-1166.
- Guarino JR. Auscultatory percussion to detect ascites. *N Engl J Med*. 1986;315(24):1555-1556.
- Espinosa P, Ducot B, Pelletier G, et al. Interobserver agreement in the physical diagnosis of alcoholic liver disease. *Dig Dis Sci*. 1987;32(3):244-247.
- Cummings S, Papadakis M, Melnick J, Gooding GAW, Tierney LM. The predictive value of physical examinations for ascites. *West J Med*. 1985;142(5):633-636.

This page intentionally left blank

Prepared by David L. Simel, MD, MHS
Reviewed by Rose Hatala, MD, and David Edelman, MD

CLINICAL SCENARIO

A 48-year-old man became intoxicated and fell down several steps. He presents to the emergency department with a normal blood pressure despite some abdominal pain. He has been a moderate to heavy drinker since his teenage years. Your examination reveals mild, diffuse abdominal discomfort and a bruise on the flank where he fell. There is bilateral ankle edema. You cannot appreciate a fluid wave, although the flanks seem to bulge.

UPDATED SUMMARY ON ASCITES

Original Review

Williams JW Jr, Simel DL. Does this patient have ascites? how to divine fluid in the abdomen. *JAMA*. 1992;267(19):2645-2648.

UPDATED LITERATURE SEARCH

Our literature search used the parent search strategy for the Rational Clinical Examination series, combined with the subject “exp ascites” published in English from 1991 to 2004. The results yielded 118 titles, for which we reviewed the titles and abstracts. Only 1 article evaluated the clinical signs for ascites in a general clinical population.

NEW FINDINGS

- The accepted reference standard (ultrasonography) detects peritoneal fluid in smaller amounts than could ever be detected by clinical examination.
- The presence of a fluid wave or shifting dullness is confirmed as the most useful finding. Because the reference standard detects such small amounts of ascites, the absence of any physical examination finding does not reliably exclude the presence of peritoneal fluid.

IMPROVEMENTS IN THE DATA PRESENTED IN THE ORIGINAL PUBLICATION

A reappraisal of the original publication showed that confidence intervals (CIs) around the symptoms we reported

would help display their potential importance. We added CIs to this update.

CHANGES IN THE REFERENCE STANDARD

Ascites refers to abnormally large collections of peritoneal fluid. Studies now confirm that very small amounts of fluid can be detected by transabdominal ultrasonography¹ or endoscopic ultrasonography.² However, there are no defined cut points at which the presence of small amounts of peritoneal fluid detected by imaging procedures meets a standard of ascites. All the studies in the original review and subsequent studies consider any amount of peritoneal fluid as “ascites.”

RESULTS OF LITERATURE REVIEW

Since the original review was published, no additional studies have evaluated a patient’s symptoms for ascites or combinations of symptoms and signs. The information about symptoms suggesting ascites comes from 1 study ([Table 6-6](#)). The finding of auscultatory percussion was evaluated, but the CIs around both the positive likelihood ratio and negative likelihood ratio include 1, suggesting that it is not a useful maneuver.

An additional study³ in a selected population of thin patients validated the *presence of the fluid wave* as the most useful finding from the clinical examination ([Table 6-7](#)). All published studies counted the presence of *any* fluid on ultrasonography as “positive”; this rigorous reference standard would, not surprisingly, demonstrate that the physical findings fail frequently in proving the absence of small

Table 6-6 Results for Symptoms of Ascites

Symptoms (1 Study, 64 Patients)	LR+ (95% CI)	LR- (95% CI)
Increased girth	4.1 (2.3-7.4)	0.17 (0.05-0.62)
Recent weight gain	3.2 (1.7-6.2)	0.42 (0.20-0.87)
Ankle swelling	2.8 (1.8-4.3)	0.10 (0.01-0.67)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

Table 6-7 Pooled Results for the Physical Signs for Ascites

Physical Sign	LR+ (95% CI)	LR- (95% CI)
Fluid wave (4 studies, 372 patients)	5.3 (2.9-9.5)	0.57 (0.38-0.85)
Peripheral edema (1 study, 63 patients)	3.8 (2.2-6.8)	0.17 (0.05-0.50)
Shifting dullness (4 studies, 372 patients)	2.1 (1.6-2.9)	0.40 (0.21-0.78)
Bulging flanks (4 studies, 372 patients)	1.8 (1.4-2.5)	0.48 (0.28-0.83)
Flank dullness (3 studies, 192 patients)	1.7 (1.0-2.7)	0.44 (0.20-1.0)
Puddle sign (3 studies, 172 patients)	1.3 (0.93-2.00)	0.79 (0.59-1.1)
Auscultatory percussion (1 study, 66 patients)	1.3 (0.85-2.00)	0.71 (0.39-1.3)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

amounts of fluid. This standard may also be why the presence of peripheral edema (evaluated in only 1 study), which is easier to detect than ascites and is a marker for extracellular fluid, may be both sensitive and specific for the presence of peritoneal fluid detected by ultrasonography. On the other hand, when the signs for ascites are absent, the lower bounds of the CIs suggest that physicians may be able to rule out large amounts of ascites.

We feel confident that the puddle sign and auscultatory percussion are not useful.

Given the low pretest probability of ascites in the general population, patients should *not* be evaluated for ascites

during a routine physical examination. When it is important to detect smaller amounts of peritoneal fluid, radiologic images will be necessary because the clinical examination will not be useful, which is especially important when evaluating for ovarian carcinoma (or other abdominal malignancies) and for patients with blunt abdominal trauma when the clinical significance of missing a small amount of peritoneal fluid is high.

EVIDENCE FROM GUIDELINES

No guidelines advocate for the routine assessment of ascites.

CLINICAL SCENARIO—RESOLUTION

Alcoholism alone does not appreciably change the likelihood of ascites (likelihood ratio, 1.4). If the baseline prevalence of ascites in general medical patients is 5%, a diagnosis of alcoholism increases the probability to only 7%. The patient in the scenario could have preexisting ascites from cirrhosis, but he could also have hemoperitoneum from the fall. Unfortunately, none of the symptoms or signs of ascites have been evaluated well for their utility during blunt trauma. The presence of peripheral edema is a useful finding when present (suggesting ascites) or when absent (suggesting no ascites). You decide you need to know for certain whether the patient has a hemoperitoneum, so you must proceed to additional testing such as ultrasonography, diagnostic peritoneal lavage, or computed tomography.^{4,5}

ASCITES—MAKE THE DIAGNOSIS

During the general physical examination, patients should *not* be evaluated for ascites. When it is important to detect smaller amounts of peritoneal fluid, radiologic images will be necessary because the clinical examination will not be useful, which is especially important when evaluating for abdominal malignancies or for patients with blunt abdominal trauma.

PRIOR PROBABILITY

The prevalence of ascites in an unselected population is low, likely on the order of less than 1% (expert opinion).

The prevalence of ascites among general medical patients will be slightly higher, but still less than 5% (expert opinion).

POPULATION FOR WHOM THE SYMPTOMS AND SIGNS SHOULD BE EVALUATED

- Cirrhosis
- Congestive heart failure
- Constrictive pericarditis
- Nephrotic syndrome
- Malnutrition, chronic diarrhea
- Neoplastic disorders (any peritoneal fluid might be important)
- Systemic infectious diseases
- Blunt abdominal trauma (any peritoneal fluid might be important)

Table 6-8 Symptoms of Ascites

	LR (95% CI)
Make Ascites More Likely	
Increased abdominal girth	4.1 (2.3-4.7)
Recent weight gain	3.2 (1.7-6.2)
Ankle swelling	2.8 (1.8-4.3)
Make Ascites Less Likely	
No ankle swelling	0.10 (0.01-0.67)
No increase in abdominal girth	0.17 (0.05-0.62)
No recent weight gain	0.42 (0.20-0.87)

Abbreviations: CI, confidence interval; LR, likelihood ratio.

Table 6-9 Signs for Ascites

	LR (95% CI)
Fluid wave	5.3 (2.9-9.5)
Shifting dullness	2.1 (1.6-2.9)

Abbreviations: CI, confidence interval; LR, likelihood ratio.

The absence of findings does not rule out the presence of smaller amounts of peritoneal fluid. See Tables 6-8 and 6-9.

REFERENCE STANDARD TESTS

- Ultrasonography
- Computed tomography
- Diagnostic paracentesis

REFERENCES FOR THE UPDATE

1. Paajanen H, Lahti P, Nordback I. Sensitivity of transabdominal ultrasonography in detection of intraperitoneal fluid in humans. *Eur Radiol*. 1999;9(7):1423-1425.
2. Nguyen PT, Chang KJ. EUS in the detection of ascites and EUS-guided paracentesis. *Gastrointest Endosc*. 2001;54(3):336-339.
3. Chongtham DS, Singh MM, Kalantri SP, Pathak S, Jain AP. Accuracy of clinical manoeuvres in detection of minimal ascites. *Indian J Med Sci*. 1998;52(11):514-520.^a

4. Branney SW, Wolfe RE, Moore EE, et al. Quantitative sensitivity of ultrasound in detecting free intraperitoneal fluid. *J Trauma*. 1995;39(2):375-380.

5. Branney SW, Moore EE, Cantrill SV, Burch JM, Terry SJ. Ultrasound based key clinical pathway reduces the use of hospital resources for the evaluation of blunt abdominal trauma. *J Trauma*. 1997;42(6):1086-1090.

^aFor the Evidence to Support the Update for this topic, see <http://www.JAMAevidence.com>.

This page intentionally left blank

EVIDENCE TO SUPPORT THE UPDATE:

Ascites

TITLE Accuracy of Clinical Maneuvers in Detection of Minimal Ascites.

AUTHORS Chongtham DS, Singh MM, Kalantri SP, Pathak S, Jain AP.

CITATION *Indian J Med Sci.* 1998;52(11):514-520.

QUESTION How well do commonly used maneuvers for detecting ascites work on a general medical ward?

DESIGN One examiner identified patients for study, whereas a second examiner performed the maneuvers on all enrolled patients. An ultrasonographer, blinded to the findings, identified all patients with any degree of ascites.

SETTING Medical ward in India.

PATIENTS A total of 66 patients admitted to a ward for cardiac, hepatic, renal, nutritional, infectious, or neoplastic disorders. Those with a history of ascites, paracentesis, or "evidence of ascites from history" were excluded. These were thin patients by western standards, with a mean weight of about 49 kg (108 lb) for men and 46 kg (101 lb) for women (there was no difference in the weight of those with vs those without ascites).

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

An examiner, blinded to the entrance criteria, evaluated each patient. The ultrasonographer was blinded to the entrance criteria and clinical findings.

MAIN OUTCOME MEASURES

Sensitivity, specificity, and likelihood ratios.

MAIN RESULTS

See [Table 6-10](#).

Table 6-10 Likelihood Ratios for Signs of Ascites^a

Test	Sensitivity	Specificity	LR+ (95% CI)	LR- (95% CI)
Bulging flanks	0.51	0.64	1.4 (0.79-2.5)	0.75 (0.49-1.1)
Flank dullness	0.57	0.61	1.5 (0.88-2.5)	0.70 (0.44-1.1)
Shifting dullness	0.46	0.74	1.8 (0.9-3.6)	0.73 (0.51-1.0)
Fluid wave	0.20	100	13 (0.79-224)	0.80 (0.67-0.95)
Puddle sign	0.46	0.68	1.4 (0.75-2.6)	0.80 (0.54-1.2)
Auscultatory percussion	0.66	0.48	2.0 (0.86-2.0)	0.71 (0.40-1.3)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

^aThe authors observed that most of the patients had "minimal" ascites.

CONCLUSION**LEVEL OF EVIDENCE** Level 2.

This study was performed with high quality, although it used only 1 examiner. The results confirm that the presence of a fluid wave is the best finding in favor of ascites. In addition, the puddle sign (as in previous studies) and auscultatory percussion have poor discriminative ability.

This study population is unique in that it consisted of patients different from those in previous studies—these patients have a small body habitus, creating an expectation that the physical examination might have yielded better results. On the other hand, the patients in this study were selected because it was not obvious whether they had ascites. Furthermore, the definition of ascites was any peritoneal fluid detected by ultrasonography (as in previous work), and the authors observed that most of the patients had minimal ascites. This study confirms that the physical examination cannot detect small amounts of peritoneal fluid.

Reviewed by David L. Simel, MD, MHS

This page intentionally left blank

CHAPTER

7

What Can the Medical History and Physical Examination Tell Us About Low Back Pain?

Richard A. Deyo, MD, MPH

James Rainville, MD

Daniel L. Kent, MD

Back pain ranks second only to upper respiratory illness as a symptomatic reason for office visits to physicians.¹ Approximately 70% of adults have low back pain at some time, but only 14% have an episode that lasts more than 2 weeks. About 1.5% have such episodes with features of sciatica.^{2,3} Most causes of back pain respond to symptomatic and physical measures, but some are surgically remediable and some are systemic diseases (cancer or disseminated infection) requiring specific therapy, so careful diagnostic evaluation is important. Features of the clinical history and physical examination influence not only therapeutic choices but also decisions about diagnostic imaging, laboratory testing, and specialist referral.

ANATOMIC/PHYSIOLOGIC ORIGINS OF FINDINGS IN THE LOW BACK

Low back pain may arise from several structures in the lumbar spine, including the ligaments that interconnect vertebrae, outer fibers of the annulus fibrosus, facet joints, vertebral periosteum, paravertebral musculature and fascia, blood vessels, and spinal nerve roots. The causes of low back pain generated through these structures include (1) musculoligamentous injuries; (2) degenerative changes in the intervertebral disks and facet joints; (3) herniation of the nucleus pulposus of an intervertebral disk, with irritation of adjacent nerve roots; (4) spinal stenosis (narrowing of the central spinal canal or the lateral recesses of the canal in which the nerve roots travel caudally; this usually results from hypertrophic degenerative changes in the disks, ligamentum flavum, and facet joints); (5) anatomic anomalies of the spine, such as scoliosis and spondylolisthesis, which are often asymptomatic but may cause pain when they are severe; (6) underlying systemic diseases, such as primary or metastatic cancer, spinal infections, and ankylosing spondylitis; and (7) visceral diseases unrelated to the spine, including diseases of the pelvic organs, kidneys, gastrointestinal tract, and aorta (diagnosis of which will not be discussed in the present report).

PREVALENCE OF DISEASES THAT PRODUCE LOW BACK PAIN

Up to 85% of patients cannot be given a definitive diagnosis because of weak associations among symptoms, pathologic changes, and imaging results.^{4,5} We assume that many of these cases are related to musculoligamentous injury or degenerative changes.

Anatomic evidence of a herniated disk is found in 20% to 30% of imaging tests (myelography, computed tomography [CT], and magnetic resonance imaging [MRI]) among normal persons.^{6,7} These herniations are asymptomatic and result in no clinical disease. The proportion of all persons with low back pain who undergo surgery for a disk herniation is only about 2%.²

In primary care, about 4% of patients with back pain will prove to have compression fractures, 3% have spondylolisthesis,

and only 0.7% have spinal malignant neoplasms (primary or metastatic).⁸⁻¹³ Even fewer have ankylosing spondylitis (about 0.3%) or spinal infections (0.01%).^{8,14,15} Widespread recognition of spinal stenosis has occurred only in the last 15 years. It is most common in older adults, but its prevalence is unknown.

Because a specific cause frequently cannot be identified, diagnostic efforts are often disappointing. Instead of seeking a precise cause in every case of back pain, it may be most useful to answer 3 basic questions: (1) Is there a serious systemic disease causing the pain? (2) Is there neurologic compromise that might require surgical evaluation? (3) Is there social or psychological distress that may amplify or prolong pain? These questions can generally be answered according to medical history and physical examination alone, and a minority of patients requires further diagnostic testing.

IS THERE EVIDENCE OF SYSTEMIC DISEASE?

Cancer

Malignant neoplasm (primary or metastatic) is the most common systemic disease affecting the spine, although it

accounts for less than 1% of episodes of low back pain. Approximately 80% of patients with this diagnosis are older than 50 years (**Table 7-1**). A history of cancer has such high specificity (0.98) that such patients should be considered to have cancer until proven otherwise. However, only one-third of patients with an underlying malignant neoplasm causing their back pain have a prior cancer diagnosis (sensitivity, 0.31). Unexplained weight loss, pain duration greater than 1 month, and failure to improve with conservative therapy are moderately specific findings. Most patients with back pain caused by cancer report that pain is unrelieved by bed rest (sensitivity > 0.90), but the finding is nonspecific.¹⁰ In a study of nearly 2000 patients with back pain, no cancer was identified in any patient younger than 50 years and without a history of cancer, unexplained weight loss, or a failure of conservative therapy (combined sensitivity, 100%).¹⁰

The physical examination is less useful than the medical history for detecting underlying cancer,¹⁰ except in late stages. Because the breast, lung, and prostate are the most common sources of spinal metastases, these organs should be examined when cancer is suspected.

Table 7-1 Estimated Accuracy of the Medical History in the Diagnosis of Spine Diseases Causing Low Back Pain

Diseases to Be Detected	Source, Year	Medical History	Sensitivity	Specificity
Cancer	Deyo and Diehl, ¹⁰ 1988	Age ≥ 50 y History of cancer Unexplained weight loss Failure to improve with a month of therapy No relief with bed rest Duration of pain > 1 mo Age ≥ 50 y or history of cancer or unexplained weight loss or failure of conservative therapy	0.77 0.31 0.15 0.31 >0.90 0.50 1.0	0.71 0.98 0.94 0.90 0.46 0.81 0.60
Spinal osteomyelitis	Waldvogel and Vasey, ¹⁶ 1980	Intravenous drug abuse, urinary tract infection, or skin infection	0.40	NA
Compression fracture	Unpublished data ^a	Age ≥ 50 y Age ≥ 70 y Trauma Corticosteroid use	0.84 0.22 0.30 0.06	0.61 0.96 0.85 0.995
Herniated disk	Deyo and Tsui-Wu, ² 1987; Spangfort, ¹⁷ 1972	Sciatica	0.95	0.88
Spinal stenosis	Turner et al, ¹⁸ 1992	Pseudoclaudication Age ≥ 50 y	0.60 0.90 ^b	NA 0.70
Ankylosing spondylitis	Gran, ¹⁹ 1985	4 of 5 positive responses ^c Age at onset ≤ 40 y Pain not relieved supine Morning back stiffness Pain duration ≥ 3 mo	0.23 1.0 0.80 0.64 0.71	0.82 0.07 0.49 0.59 0.54

Abbreviation: NA, not available.

^aFrom 833 patients with back pain at a walk-in clinic, all of whom received pain lumbar radiographs.

^bAuthors' estimate.

^cThe 5 screening questions were (1) Onset of back discomfort before age 40 years? (2) Did the problem begin slowly? (3) Persistence for at least 3 months? (4) Morning stiffness? and (5) Improved by exercise?

Spinal Infections

Spinal infections usually are blood-borne from other sites, including urinary tract infections, indwelling urinary catheters, skin infections, and injection sites for illicit intravenous drugs. One of these sites is identified in approximately 40% of patients with spinal infections (sensitivity, 0.40).¹⁶

In patients with spinal infections, the sensitivity of fever is disappointing, varying from 0.27 for tuberculous osteomyelitis to 0.50 for pyogenic osteomyelitis²⁰ and 0.83 for spinal epidural abscess.²¹ Because 2% of patients in primary care with mechanical low back pain have fever (perhaps because of viral syndromes), specificity for bacterial infection is approximately 0.98.¹⁰ Spine tenderness in response to percussion has a sensitivity of 0.86 for bacterial infection, but specificity is poor (0.60).^{10,22,23}

Compression Fractures

Although spinal compression fractures are not systemic diseases, they often occur in persons with generalized osteoporosis. Most patients with this problem do not have a history of identifiable trauma (sensitivity, 0.30). A person with back pain who is receiving long-term corticosteroid therapy is considered to have a compression fracture until proven otherwise (specificity, 0.99). Black and Hispanic women have only one-fourth as many compression fractures as white women.²⁴ As shown in Table 7-1, age greater than 70 years is a relatively specific finding (specificity, 0.96).

Ankylosing Spondylitis and Spine Range-of-Motion Measures

Ankylosing spondylitis shares several historical features with other inflammatory arthropathies, such as rheumatoid arthritis. Calin et al²⁵ described 5 screening questions for ankylosing spondylitis: (1) Is there morning stiffness? (2) Is there improvement in discomfort with exercise? (3) Was the onset of back pain before age 40 years? (4) Did the problem begin slowly? (5) Has the pain persisted for at least 3 months?

With at least 4 positive answers to define a positive "test" result, the sensitivity of these questions was 0.95 and specificity was 0.85,²⁵ although other authors report lower sensitivity.^{19,26} When screening for a rare disease such as ankylosing spondylitis, typically, the predictive value of a positive test is low. In an industrial screening program, only 16 of 367 persons with positive criteria proved to have ankylosing spondylitis (a predictive value of 0.04).²⁷ "Inflammatory" symptoms (morning stiffness, night pain, and relief with exercise) are moderately sensitive but nonspecific. All patients with ankylosing spondylitis in 1 population survey reported symptom onset before age 40 years, making this history highly sensitive but nonspecific (Table 7-1).¹⁹

Reduced spinal mobility results from fusion of adjacent vertebrae in this condition. The Schober test, which measures distraction between 2 marks on the skin during forward flexion, is a commonly described method for quantifying

reduced flexion. Although it is moderately reproducible,^{23,28} reduced spine flexion is not specific for inflammatory spondylopathies, being equally common in patients with chronic back pain or spine tumors.²⁹ Reduced chest expansion (using a strict criterion for abnormality, such as expansion ≤ 2.5 cm) is highly specific (0.99) but insensitive in early ankylosing spondylitis (0.09),^{19,30} so that predictive values are poor.

Tests for sacroiliac joint tenderness (to discriminate ankylosing spondylitis from mechanical spine conditions) include a hip extension test, anteroposterior pelvic pressure, lateral pelvic compression, and direct pressure on the sacroiliac joints. Unfortunately, these tests are poorly reproducible^{23,31} and inaccurate in distinguishing ankylosing spondylitis from mechanical spine complaints.^{32,33} Early ankylosing spondylitis is most often suspected from radiographs obtained because of persistent pain.

Although spine flexion is of limited diagnostic value, it may be useful in planning or monitoring physical therapy in patients with low back pain of any cause.³⁴ Range of motion in multiple directions can be assessed with 2 inclinometers (used in the construction industry) with good precision.^{28,34} The technique is detailed elsewhere.³⁴

IS THERE EVIDENCE OF NEUROLOGIC COMPROMISE?

The spinal cord, cauda equina, and nerve roots are vulnerable to several disorders that cause back pain and sciatica. The most common of these is a herniated intervertebral disk, but other causes include nerve root entrapment in the root canals by bony and ligamentous hypertrophy, spinal stenosis, spinal or paraspinal infections, and neoplasms. Irritation of neurologic structures is manifested as motor, reflex, or sensory dysfunction in the lower extremities and (rarely) as bowel or bladder dysfunction.

The first clue to nerve root irritation is usually sciatica, a sharp or burning pain radiating down the posterior or lateral aspect of the leg (usually to the foot or ankle), often associated with numbness or paresthesia. The pain is sometimes aggravated by coughing, sneezing, or the Valsalva maneuver. Among patients with low back pain alone (no sciatica or neurologic symptoms), the prevalence of neurologic impairments is so low that extensive neurologic evaluation is usually unnecessary.

Lumbar Disk Herniations

Sciatica has such a high sensitivity (0.95) that its absence makes a clinically important lumbar disk herniation unlikely.^{17,35} Using the accuracy of sciatica in Table 7-1 and a prevalence of surgically important disk herniations of 2%, we estimate the likelihood of disk herniation in a patient without sciatica to be 1 in 1000. Most patients have a long history of recurrent back pain before the onset of sciatica, but when a frank disk herniation occurs, leg pain usually overshadows the back pain. The peak incidence of herniated lumbar disks is in adults between the ages of 30 and 55 years.¹⁷

A symptomatic disk herniation tethers the affected nerve root, so pain results from stretching the nerve by straight-leg raising (SLR) from the supine position. This is performed by cupping the heel in 1 hand and keeping the knee fully extended with the other. The straight leg is slowly raised from the examining table until pain occurs. Tension is transmitted to the nerve roots once the leg is raised beyond 30 degrees, but after 70 degrees, further movement of the nerve is negligible.³⁶ A typical positive SLR sign is one that reproduces the patient's sciatica between 30 degrees and 60 degrees of leg elevation.^{17,37,38}

A related test is the crossed SLR (CSLR) sign. This occurs when SLR is performed on the patient's well leg and is found to elicit pain in the leg with sciatica. The precision of tests for SLR is shown in **Table 7-2.**^{23,39-41} Visual estimation is reasonably accurate, but a goniometer or inclinometer improves interobserver agreement.

Pain on ipsilateral SLR at 60 degrees is moderately sensitive for herniated lumbar disks but nonspecific, because limitation is often observed in the absence of disk herniations (**Table 7-3**).⁴³⁻⁴⁵ CSLR is less sensitive but highly specific.^{17,44,45,48} Thus, a positive CSLR test result substantially increases the likelihood of a disk herniation, whereas a negative result is of limited value. The lower the angle of a positive SLR test, the more specific the test becomes and the larger the disk protrusion found at surgery.^{46,49}

Straight-leg raising is most appropriate for testing the lower lumbar nerve roots (L5 and S1), where the majority of herniated disks occur. Irritation of higher lumbar roots is tested with the femoral nerve stretch test (flexing the knee with

patient prone), but the precision and accuracy of this test are unknown.

Assessment of Motor, Reflex, and Sensory Function

Ninety-eight percent of clinically important lumbar disk herniations occur at either the L4 to L5 or the L5 to S1 intervertebral level,^{17,44-46} causing neurologic impairments in the motor and sensory territories of the L5 and S1 nerve roots. Thus, the most common neurologic impairments are weakness of the ankle and great-toe dorsiflexors (L5), diminished ankle reflexes (S1), and sensory loss in the feet (L5 and S1).^{17,44-46} In a patient with sciatica, the neurologic examination can focus on these functions.

Ankle dorsiflexor strength is tested by having the supine patient dorsiflex the ankle against the examiner's resistance. Inability to maintain dorsiflexion against the examiner should be considered weakness, and the healthy side should be checked for comparison. This method shows excellent precision (Table 7-2) and is more reproducible than the patient's ability to heel stand.²³ Ankle dorsiflexor weakness rarely occurs in isolation and is nearly always associated with weak toe dorsiflexion, sensory deficits, or impaired reflexes.⁵⁰ For toe strength, the supine patient is instructed to maximally dorsiflex the great toe ("point your big toe at your nose" seems to work well) and resist the examiner's effort to flex the toe with 2 fingers.

Ankle reflexes are more difficult to reproduce, and patient positioning may be important. The side-lying, prone, and

Table 7-2 Reproducibility of Physical Examination Findings

Category	Test	Unit of Measurement	Interobserver Agreement (Statistic)	Source, Year
Tenderness	Bone tenderness	Yes/No	0.40 (κ)	McCombe et al, ²³ 1989
	Soft-tissue tenderness	Yes/No	0.24 (κ)	McCombe et al, ²³ 1989
	Muscle spasm	Yes/No	"Discarded; too unreliable"	Waddell et al, ³⁹ 1982
SLR	Ipsilateral SLR, inclinometer	Degrees	0.78-0.97 (r)	Hoehler and Tobis, ⁴⁰ 1982 Hsieh et al, ⁴¹ 1983
	Ipsilateral SLR goniometer	Degrees	0.69 (r)	McCombe et al, ²³ 1989
	SLR causes leg pain	Yes/No	0.66 (κ)	McCombe et al, ²³ 1989
	Ipsilateral SLR < 75° by visual estimation	Yes/No	0.56 (κ)	Waddell et al, ³⁹ 1982
	CSLR, causes pain	Yes/No	0.74 (κ)	McCombe et al, ²³ 1989
Neurologic examination	Ankle dorsiflexion weak	Yes/No	1.00 (κ)	McCombe et al, ²³ 1989
	Great toe extensors weak	Yes/No	0.65 (κ)	McCombe et al, ²³ 1989
	Ankle reflexes normal	Yes/No	0.39-0.50 (κ)	McCombe et al, ²³ 1989 Schwartz et al, ⁴² 1990
	Any sensory deficit	Yes/No	0.68 (κ)	McCombe et al, ²³ 1989
	Calf wasting	Yes/No	0.80 (κ)	McCombe et al, ²³ 1989
Inappropriate signs	Superficial tenderness	Yes/No	0.29 (κ)	McCombe et al, ²³ 1989
	Simulated rotation or axial loading causes pain	Yes/No	0.25 (κ)	McCombe et al, ²³ 1989
	SLR with distraction causes pain	Yes/No	0.40 (κ)	McCombe et al, ²³ 1989
	Inexplicable pattern, neurologic examination	Yes/No	0.03 (κ)	McCombe et al, ²³ 1989
	Overreaction	Yes/No	0.29 (κ)	McCombe et al, ²³ 1989

Abbreviations: CSLR, crossed straight-leg raising; SLR, straight-leg raising.

Table 7-3 Estimated Accuracy of Physical Examination for Lumbar Disk Herniation Among Patients With Sciatica

Test	Source, Year	Sensitivity ^a	Specificity ^a	Comments
Ipsilateral SLR	Kosteljanetz et al, ⁴³ 1984; Hakelius and Hindmarsh, ⁴⁴ 1972	0.80	0.40	Positive test result; leg pain at < 60°
CSLR	Spangfort, ¹⁷ 1972; Hakelius and Hindmarsh, ^{44,45} 1972	0.25	0.90	Positive test result: reproduction of contralateral pain
Ankle dorsiflexion weakness	Spangfort, ¹⁷ 1972; Hakelius and Hindmarsh, ⁴⁴ 1972	0.35	0.70	HNP usually at L4-5 (80%)
Great toe extensor weakness	Hakelius and Hindmarsh, ⁴⁴ 1972; Kortelainen et al, ⁴⁶ 1985	0.50	0.70	HNP usually at L5-S1 (60%) or L4-5 (30%)
Impaired ankle reflex	Spangfort, ¹⁷ 1972; Hakelius and Hindmarsh, ⁴⁴ 1972	0.50	0.60	HNP usually at L5-S1; absent reflex increases specificity
Sensory loss	Kosteljanetz et al, ⁴³ 1984; Kortelainen et al, ⁴⁶ 1985	0.50	0.50	Area of loss poor predictor of HNP level
Patella reflex	Aronson and Dunsmore, ⁴⁷ 1963	0.50	NA	For upper lumbar HNP only
Ankle plantar flexion weakness	Hakelius and Hindmarsh, ⁴⁴ 1972	0.06	0.95	
Quadriceps weakness	Hakelius and Hindmarsh, ⁴⁴ 1972	<0.01	0.99	

Abbreviations: CSLR, crossed straight-leg raising; HNP, herniated nucleus pulposus; SLR, straight-leg raising; NA, not available.

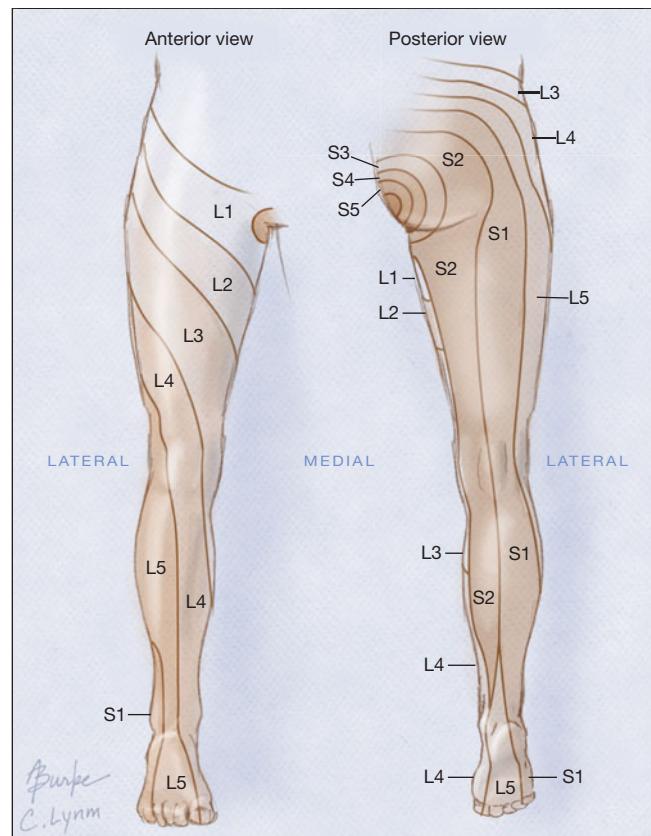
^aSensitivity and specificity were calculated by the authors of the present report. Values represent rounded averages where multiple references were available. All results are from surgical case series.

kneeling positions are probably best (rather than the sitting position), but we are unaware of comparative data. The foot is gently rocked until relaxation is obtained, and the calf muscles should be held under slight tension by dorsiflexing the foot. Estimated κ values for the precision of ankle reflexes range from 0.39 to 0.50.^{23,48} Schwartz et al⁴² found that a plantar tap is as good as an Achilles tendon tap (estimated $\kappa = 0.55$). In this technique, the patient lies supine and the ball of the foot is tapped with the reflex hammer. The plantar tap was preferred by patients and could be elicited in 91% of patients younger than 65 years but in only 71% of patients older than 65 years.

Ankle plantar flexion is an S1 function, but only severe impairments can be clinically detected, and sensitivity for disk herniation is low (Table 7-3). Toe walking appears to be an unreliable method of assessing plantar flexion strength ($\kappa = 0$).²³ Hamstring and hip extensor strength have been used to evaluate S1 root injuries, but their precision and accuracy are unknown. Muscle wasting indicates longstanding denervation or disease and may be detected visually. Good precision was noted for observations of anterior compartment and hamstring wasting in one study (Table 7-2).²³

Sensory examination of the lower extremities takes time. Patients distinguish differences in pain intensity by pinprick more accurately than differences in touch or temperature, and sensory impairment from nerve root compression is most frequent in the distal extremes of the dermatomes.⁵¹ Therefore, an efficient strategy is to check for symmetry of pain elicited by pinprick in the extremes of the L4, L5, and S1 dermatomes (the medial aspect, dorsum, and lateral aspect of the feet) (Figure 7-1).

Higher lumbar nerve roots account for only about 2% of lumbar disk herniations. They are suspected when numbness or pain involves the anterior thigh more prominently

**Figure 7-1** Lower-extremity Dermatomes

than the calf (Figure 7-1). Testing includes knee reflexes, quadriceps strength, and psoas strength.^{17,47,50} Quadriceps weakness is virtually always associated with impairment in the patella reflex.⁵⁰

The accuracy of neurologic findings for the diagnosis of a herniated disk is only moderate (Table 7-3). Considering combinations is helpful, however, because a finding of impaired ankle reflexes or weak foot dorsiflexion would have a sensitivity of almost 90% for patients with surgically proven disk herniations.¹⁷ Multiple findings related to SLR or neurologic examination increase the probability that a herniated disk will be found at surgery.⁵²

Spinal Stenosis

The mean age of patients at surgery for spinal stenosis is 55 years, with an average symptom duration of 4 years.¹⁸ The characteristic history is that of neurogenic claudication: pain in the legs and occasionally neurologic deficits that occur after walking. In contrast to arterial ischemic claudication, neurogenic claudication is more likely to occur on standing alone (without ambulation), may increase with cough or sneeze, and is associated with normal arterial pulses.⁵³ The sensitivity of neurogenic claudication is modest (about 0.60),¹⁸ but it is probably quite specific.

Few data are available concerning the accuracy of physical examination because stenosis has been widely recognized only in recent years. Diagnostic criteria, indications for surgery, and the natural history are still being elucidated. Increased pain on spine extension is typical of stenosis (whereas flexion is usually most painful with herniated disks), but accuracy data are unavailable. The sensitivity of leg pain is about 85%; neurologic abnormalities, about 60%; and abnormal SLR, about 50%.^{18,53}

Cauda Equina Syndrome

A massive midline disk herniation may cause spinal cord or cauda equina compression, requiring immediate surgical referral. Fortunately, the cauda equina syndrome occurs in only 1% to 2% of all patients with lumbar disk herniations who come to surgery,¹⁷ so its prevalence among all patients with low back pain is about 0.0004. The most consistent finding is urinary retention, with a sensitivity of 0.90.⁵⁴⁻⁵⁶ The most common sensory deficit occurs over the buttocks, posterior-superior thighs, and perineal regions (“saddle anesthesia”), with a sensitivity of about 0.75.⁵⁴⁻⁵⁶ Anal sphincter tone is diminished in 60% to 80% of cases.⁵⁴⁻⁵⁶ Assuming a specificity of about 95%, the predictive value of a negative test result (no urinary retention) would be almost 100%. Unilateral or bilateral sciatica, sensory and motor deficits, and abnormal SLR results are all common, with sensitivities of greater than 0.80.⁵⁴⁻⁵⁶

Indications for Imaging Tests

There is a growing consensus that radiographs are not necessary for every patient with low back pain because of a low yield of useful findings, potentially misleading results, substantial gonadal irradiation, and common interpretive disagreements. The Quebec Task Force on Spinal Disorders suggested that early radiography was necessary only in the face of neurologic deficits, age older than 50 years or younger than 20 years, fever, trauma, or signs of neoplasm.⁵⁷ Table 7-1

indicates screening questions that can exclude neoplasm according to patient medical history alone.¹⁰

MRI and CT can be used even more selectively, usually for surgical planning. The finding of herniated disks and spinal stenosis in many asymptomatic persons^{6,7} indicates that imaging results alone can be misleading, and valid decision making requires correlation with the medical history and physical examination.⁵⁸

IS THERE EVIDENCE OF SOCIAL OR PSYCHOLOGICAL DISTRESS THAT MAY AMPLIFY OR PROLONG PAIN?

Some features of patient medical history influence management regardless of the exact spinal pathology. Chronic pain or depression may be indications for the use of antidepressant medication rather than opiates. Alcohol or drug abuse influences the choice of medications and requires specific intervention. Disability compensation claims or litigation may affect initial evaluation and prognosis, and patients seeking compensation often respond poorly to a variety of treatments.⁵⁹

Patients with chronic low back pain (≥ 3 months) present complex problems, and often, a pathoanatomic cause is not apparent.⁶⁰ Unlike acute pain, chronic pain is often not associated with ongoing tissue injury, serves no biological usefulness, and is not accompanied by the autonomic response of sympathetic overactivity. Vegetative signs, such as sleep disturbance, appetite disturbance, and irritability, appear, and pain is often reinforced or perpetuated by social and psychological factors. Back pain can affect employment, income, family, and social roles, producing psychological distress.^{60,61} Resulting somatic amplification can serve the patient's needs for economic survival and maintenance of self-esteem.⁶¹

In patients with chronic low back pain, the absence of systemic disease and treatable anatomic abnormalities should be confirmed by medical history, physical examination, and review of diagnostic tests. Neurologic abnormalities often prove to be longstanding and may persist after surgical interventions. Evidence of psychological distress should be sought because this may respond to direct intervention and improve the likelihood of response to other treatments. The Minnesota Multiphasic Personality Inventory is impractical in most primary care settings, and shorter depression scales are useful for screening.^{62,63}

Waddell et al⁶⁴ proposed 5 categories of inappropriate or non-organic signs that correlated with other indicators of psychological distress: (1) inappropriate tenderness that is superficial or widespread; (2) pain on simulated axial loading by pressing on the top of the head, or simulated spine rotation (performed by holding the patient's arms to the side while rotating the hips, ensuring that the shoulders and hips rotate together); (3) “distraction” signs, such as inconsistent performance between SLR in the seated position vs the supine position; (4) regional disturbances in strength and sensation that do not correspond with nerve root innervation patterns; and (5) overreaction during the physical examination. The occurrence of any 1 sign was of limited value, but positive findings in 3 of the 5 categories suggested

psychological distress. The precision of nonorganic signs was reported by Waddell et al⁶⁴ to be high, but subsequent evaluation found poor precision in the regional disturbance category (Table 7-2).²³

SUMMARY AND RECOMMENDATIONS

History

1. A few key questions can raise or lower the probability of underlying systemic disease. The most useful items are age, history of cancer, unexplained weight loss, duration of pain, and responsiveness to previous therapy.
2. Intravenous drug use or urinary infection raises the suspicion of spinal infection.
3. Ankylosing spondylitis is suggested by the patient's age and sex (most common in young men), but most clinical findings have limited accuracy.
4. Failure of bed rest to relieve the pain is a sensitive finding for all these systemic conditions, although not specific.
5. Neurologic involvement is suggested by symptoms of sciatica or pseudoclaudication. Pain radiating distally (below the knee) is more likely to represent a true radiculopathy than pain radiating only to the posterior thigh. A history of numbness or weakness in the legs further increases the likelihood of neurologic involvement.
6. Inquiry should be made concerning symptoms of the cauda equina syndrome: bladder dysfunction (especially urinary retention) and saddle anesthesia in addition to sciatica and weakness.
7. The psychosocial history helps to estimate prognosis and plan therapy. The most useful items are a history of failed treatments, substance abuse, and disability compensation. Brief screening questionnaires for depression may suggest important therapeutic opportunities.

Physical Examination

1. Fever suggests the possibility of spinal infection. Vertebral tenderness is a sensitive finding for infection but not specific.
2. The search for soft-tissue tenderness is unlikely to provide reproducible data or demonstrably valid pathophysiologic inferences.^{23,39}
3. Limited lumbar flexion is not highly sensitive or specific for ankylosing spondylitis or other diagnoses. However, limited spinal motion may be useful in planning physical therapy and monitoring response.
4. In a patient with sciatica or possible neurogenic claudication, SLR should be assessed bilaterally, preferably with an inclinometer or goniometer.
5. Neurologic examination emphasizes ankle dorsiflexion strength, great-toe dorsiflexion strength, ankle reflexes, and the sensory examination. A rapid screening sensory examination would test pinprick sensation in the medial, dorsal, and lateral aspects of the foot.
6. For the patient with chronic pain, all the evaluations described herein should be completed. Anatomically "inappropriate" signs may be helpful in identifying psychological

distress as a result of or as an amplifier of low back symptoms. The most reproducible of these signs are superficial tenderness, distracted SLR, and the observation of patient overreaction during the physical examination.

Author Affiliations at the Time of the Original Publication

From the Health Services Research and Development Field Program, (Washington) Veterans Affairs Medical Center (Drs Deyo and Kent) Seattle; the Departments of Medicine (Drs Deyo and Kent) and Health Services (Dr Deyo), University of Washington, Seattle; and the Department of Rehabilitation Medicine, Tufts University School of Medicine, Boston, Massachusetts (Dr Rainville).

Acknowledgments

This report was supported in part by the Northwest Health Services Research and Development Field Program, Seattle Veterans Affairs Medical Center; by grant HS-06344 from the Agency for Health Care Policy and Research (the Back Pain Outcome Assessment Team); and by grant 88298-2H from the John A. Hartford Foundation through the Multi-Institutional Technology Assessment Consortium.

Darlena Bursell assisted in the preparation of the manuscript, and Monica Hayes helped to coordinate study activities.

REFERENCES

1. Cypress BK. Characteristics of physician visits for back symptoms: a national perspective. *Am J Public Health*. 1983;73(4):389-395.
2. Deyo RA, Tsui-Wu JY. Descriptive epidemiology of low-back pain and its related medical care in the United States. *Spine*. 1987;12(3):264-268.
3. Deyo RA, Loeser JD, Bigos SF. Herniated lumbar intervertebral disk. *Ann Intern Med*. 1990;112(8):598-603.
4. White AA, Gordon SL. Synopsis: workshop on idiopathic low-back pain. *Spine*. 1982;7(2):141-149.
5. Nachemson A. The lumbar spine: an orthopedic challenge. *Spine*. 1976;1(1):59-71.
6. Weisel SE, Tsourmas N, Feffer H, Citrin CM, Patronas N. A study of computer-assisted tomography, I: the incidence of positive CAT scans in an asymptomatic group of patients. *Spine*. 1984;9(6):549-551.
7. Boden SD, Davis DO, Dina TS, Patronas NJ, Wiesel SW. Abnormal magnetic resonance scans of the lumbar spine in asymptomatic subjects. *J Bone Joint Surg Am*. 1990;72(3):403-408.
8. Liang M, Komaroff AL. Roentgenograms in primary care patients with acute low back pain: a cost-effectiveness analysis. *Arch Intern Med*. 1982;142(6):1108-1112.
9. Deyo RA. Early diagnostic evaluation of low back pain. *J Gen Intern Med*. 1986;1(5):328-338.
10. Deyo RA, Diehl AK. Cancer as a cause of back pain: frequency, clinical presentation, and diagnostic strategies. *J Gen Intern Med*. 1988;3(3):230-238.
11. Scavone JG, Latshaw RF, Weidner WA. Anteroposterior and lateral radiographs: an adequate lumbar spine examination. *Am J Radiol*. 1981;136(4):715-717.
12. Gehweiler JA, Daffner RH. Low back pain: the controversy of radiologic evaluation. *Am J Radiol*. 1983;140(1):109-112.
13. Brekkan A. Radiographic examination of the lumbosacral spine: an "age-stratified" study. *Clin Radiol*. 1983;34(3):321-324.
14. Carter ET, McKenna CH, Brian DD, Kurland LT. Epidemiology of ankylosing spondylitis in Rochester, Minnesota, 1935-1973. *Arthritis Rheum*. 1979;22(4):365-370.
15. Hawkins BR, Dawkins RL, Christiansen FT, Zilko PJ. Use of the B27 test in the diagnosis of ankylosing spondylitis: a statistical evaluation. *Arthritis Rheum*. 1981;24(5):743-746.
16. Waldvogel FA, Vasey H. Osteomyelitis: the past decade. *N Engl J Med*. 1980;303(7):360-370.

17. Spangfort EV. Lumbar disc herniation: a computer aided analysis of 2504 operations. *Acta Orthop Scand.* 1972;142:1-93.
18. Turner JA, Ersek M, Herron L, Deyo R. Surgery for lumbar spinal stenosis: attempted meta-analysis of the literature. *Spine.* 1992;17(1):1-8.
19. Gran JT. An epidemiological survey of the signs and symptoms of ankylosing spondylitis. *Clin Rheumatol.* 1985;4(2):161-169.
20. Sapico FL, Montgomerie JZ. Pyogenic vertebral osteomyelitis: report of nine cases and review of the literature. *Rev Infect Dis.* 1979;1(5):754-776.
21. Baker AS, Ojemann RG, Swartz MN, et al. Spinal epidural abscess. *N Engl J Med.* 1975;293(10):463-468.
22. Chandrasekar PH. Low back pain and intravenous drug abusers. *Arch Intern Med.* 1990;150(5):1125-1128.
23. McCombe PF, Fairbank JCT, Cockersole BC, Pynsent PB. Reproducibility of physical signs in low-back pain. *Spine.* 1989;14(9):908-918.
24. Bauer RL, Deyo RA. Low risk of vertebral fracture in Mexican American women. *Arch Intern Med.* 1987;147(8):1437-1439.
25. Calin A, Porta J, Fries JF, Schurman DJ. Clinical history as a screening test for ankylosing spondylitis. *JAMA.* 1977;237(24):2613-2614.
26. Van der Linden S, Valkenburg HA, Cats A. Evaluation of diagnostic criteria for ankylosing spondylitis: a proposal for modification of the New York criteria. *Arthritis Rheum.* 1984;27(4):361-368.
27. Calin A, Kaye B, Sternberg M, Antell B, Chan M. The prevalence and nature of back pain in an industrial complex: questionnaire and radiographic and HLA analysis. *Spine.* 1980;5(2):201-205.
28. Reynolds PMG. Measurement of spinal mobility: a comparison of three methods. *Rheumatol Rehabil.* 1975;14(3):180-185.
29. Rae PS, Waddell G, Venner RM. A simple technique for measuring lumbar spinal flexion. *J R Coll Surg Edin.* 1984;29(5):281-284.
30. Moll JMH, Wright V. An objective clinical study of chest expansion. *Ann Rheum Dis.* 1972;31(1):1-8.
31. Potter NA, Rothstein JM. Intertester reliability for selected clinical tests of the sacroiliac joint. *Phys Ther.* 1985;65(11):1671-1675.
32. Russell AS, Maksymowich W, LeClercq S. Clinical examination of sacroiliac joints: a prospective study. *Arthritis Rheum.* 1981;24(12):1575-1577.
33. Blower PW, Griffin AJ. Clinical sacroiliac tests in ankylosing spondylitis and other causes of low back pain. *Ann Rheum Dis.* 1984;43(2):192-195.
34. Mayer TG, Tencer AF, Kristoferson S, Mooney V. Use of noninvasive techniques for quantification of spinal range of motion in normal subjects and chronic low-back pain dysfunction patients. *Spine.* 1984;9:588-595.
35. Alpers BJ. The neurological aspects of sciatica. *Med Clin North Am.* 1953;37(2):503-510.
36. Brieg A, Troup JDG. Biomechanical considerations in the straight-leg-raising test: cadaveric and clinical studies of medial hip rotation. *Spine.* 1979;4(3):242-250.
37. Charnley J. Orthopedic signs in the diagnosis of disc protrusion with special reference to the straight leg raising test. *Lancet.* 1951;1(6648):186-192.
38. Kosteljanetz M, Bang F, Schmidt-Olsen S. The clinical significance of straight-leg-raising (Lasegue's sign) in the diagnosis of prolapsed lumbar disc. *Spine.* 1988;13(4):393-395.
39. Waddell G, Main CS, Morris EW, et al. Normality and reliability in the clinical assessment of backache. *BMJ.* 1982;284(6328):1519-1523.
40. Hoehler FK, Tobis JS. Low back pain and its treatment by spinal manipulation: measures of flexibility and asymmetry. *Rheumatol Rehabil.* 1982;21(1):21-26.
41. Hsieh CY, Walker JM, Gillis K. Straight leg raising test: comparison of three instruments. *Phys Ther.* 1983;63(9):1429-1432.
42. Schwartz RS, Morris JGL, Crimmins D, et al. A comparison of two methods of eliciting the ankle jerk. *Aust N Z J Med.* 1990;20(2):116-119.
43. Kosteljanetz M, Espersen JO, Halaburt H, Miletic T. Predictive value of clinical and surgical findings in patients with lumbago-sciatica: a prospective study (part 1). *Acta Neurochir.* 1984;73(1-2):67-76.
44. Hakelius A, Hindmarsh J. The comparative reliability of preoperative diagnostic methods in lumbar disc surgery. *Acta Orthop Scand.* 1972;43(4):234-238.
45. Hakelius A, Hindmarsh J. The significance of neurological signs and myelographic findings in the diagnosis of lumbar root compression. *Acta Orthop Scand.* 1972;43(4):239-246.
46. Kortelainen P, Puranen J, Koivisto E, Lahde S. Symptoms and signs of sciatica and their relation to the localization of the lumbar disc herniation. *Spine.* 1985;10(1):88-92.
47. Aronson HA, Dunsmore RH. Herniated upper lumbar discs. *J Bone Joint Surg Am.* 1963;45:311-317.
48. Hudgins RW. The crossed straight leg raising test: a diagnostic sign of herniated disc. *J Occup Med.* 1979;21(6):407-408.
49. Shiqing X, Quanzhi Z, Dehao F. Significance of straight-leg-raising test in the diagnosis and clinical evaluation of lower lumbar intervertebral disc protrusion. *J Bone Joint Surg Am.* 1987;69(4):517-522.
50. Blower PW. Neurologic patterns in unilateral sciatica. *Spine.* 1981;6(2):175-179.
51. Keegan JJ. Dermatome hypalgesia associated with herniation of intervertebral disk. *Arch Neurol Psychiatry.* 1943;50:67-83.
52. Morris EW, DiPaola M, Vallance R, Waddell G. Diagnosis and decision making in lumbar disc prolapse and nerve entrapment. *Spine.* 1986;11(5):436-439.
53. Hawkes CH, Roberts GM. Neurogenic and vascular claudication. *J Neurol Sci.* 1978;38(3):337-345.
54. Kostuk JP, Harrington I, Alexander D, Rand W, Evans D. Cauda equina syndrome and lumbar disc herniation. *J Bone Joint Surg Am.* 1986;68(3):386-391.
55. O'Laoire SA, Crockard HA, Thomas DG. Prognosis for sphincter recovery after operation for cauda equina compression owing to lumbar disc prolapse. *BMJ.* 1981;282(6279):1852-1854.
56. Tay ECK, Chacha PB. Midline prolapse of a lumbar intervertebral disc with compression of the cauda equina. *J Bone Joint Surg Br.* 1979;61(1):43-46.
57. Spitzer WO, LeBlanc FE, Dupuis M, et al. Scientific approach to the assessment and management of activity-related spinal disorders: a monograph for clinicians: report of the Quebec Task Force on Spinal Disorders. *Spine.* 1987;12(suppl 7):S16-S21.
58. Deyo RA, Bigos SJ, Maravilla KR. Diagnostic imaging procedures for the lumbar spine. *Ann Intern Med.* 1989;111(11):865-867.
59. Walsh NE, Dumitru D. The influence of compensation on recovery from low back pain. *Occup Med.* 1988;3(1):109-121.
60. Gatchel RJ, Mayer TG, Capra P, Diamond P, Barnett J. Quantification of lumbar function, VI: the use of psychological measures in guiding physical functional restoration. *Spine.* 1986;11(1):36-42.
61. Korbon GA, DeGood DE, Schroeder ME, Schwartz DP, Shutty MS. The development of a somatic amplification rating scale for low-back pain. *Spine.* 1987;12(8):787-791.
62. Burnam AM, Wells KB, Leake B, Landsverk J. Development of a brief screening instrument for detecting depressive disorders. *Med Care.* 1988;26(8):775-789.
63. Rucker L, Frye EB, Cygan RW. Feasibility and usefulness of depression screening in medical outpatients. *Arch Intern Med.* 1986;146(4):729-731.
64. Waddell G, McCulloch JA, Kummel E, Vernner RM. Nonorganic physical signs in low back pain. *Spine.* 1980;5(2):117-125.

Prepared by Ben Stern, MS, DPT,
Richard A. Deyo, MD, MPH,
and James Rainville, MD

Reviewed by Richard S. Bedlack, MD, PhD

CLINICAL SCENARIO

A physically active 61-year-old man presents with complaints of low back pain and occasional pain in his left buttock and upper thigh. His symptoms began approximately 3 weeks ago. In addition, increasing pain in his lower extremities is preventing him from participating in his hobbies and socializing with his usual group of friends. He has no history of weight loss and no changes in bowel or bladder habits. During the physical examination, the patient reports thigh and back pain at 50 degrees during the straight leg raise (SLR) test on the left but no radiation below the knee. He has slight pain in the back of his right leg with SLR testing to 75 degrees. When you test his quadriceps strength, his left side seems a little weaker than the right, but the testing is limited by his discomfort. His single-leg sit-to-stand test result is normal. The ankle reflexes are absent bilaterally. Given the results of this brief history and physical examination, are there other maneuvers you could perform? What diagnosis can you provide for this patient?

UPDATED SUMMARY ON LOW BACK PAIN

Original Review

Deyo RA, Rainville J, Kent DL. What can the history and physical examination tell us about low back pain? *JAMA*. 1992;268(6):760-765.

UPDATED LITERATURE SEARCH

We initially sought articles including the keywords "back pain," "herniated disk," or "sciatica" and "specificity" using the "Clinical Query" mechanism in MEDLINE. In addition, we filtered for human, English-language articles, resulting in 190 citations from 1992 to August 2004. We performed an additional search including the various forms and combinations of the terms "intervertebral disk displacement," "characteristic," "feature," "finding," "marker," "predictor," "sign," "test," "variable," "physical," "exam," and "sensitivity." This search added 28 unique citations to our article pool. We also searched bibliographies and personal files for additional articles. Two reviewers independently examined the abstracts of the articles retrieved by

this search. Articles included in the update were selected by consensus between the 2 reviewers; 6 articles were deemed relevant to this update.

We included 2 systematic reviews^{1,2} and 3 prospective studies³⁻⁵ that focused on the physical examination of individuals with low back pain. We excluded 1 literature synthesis when we could not replicate the data on our review of the original references.⁶ We considered articles related to detecting lumbar radiculopathy or underlying systemic diseases among patients with low back pain. We excluded neck pain or spinal stenosis from our review.

NEW FINDINGS

- More than 90% of normal individuals younger than 60 years have bilateral ankle reflexes, but 5% have 1 absent ankle reflex and 5% have no ankle reflexes. Among those older than 60 years, only 60% have both ankle reflexes, 30% have no ankle reflexes, and 10% have an absent ankle reflex in 1 lower extremity. These age-related deficits reduce the specificity of a diminished ankle reflex as a test for L5 to S1 radiculopathy in older patients.
- When a patient with low back pain is screened for cancer, it may be prudent to inquire about pain at night. Night pain is sensitive among patients with cancer as a cause of back pain, but not specific: sensitivity, 0.92; specificity, 0.46; positive likelihood ratio (LR+), 1.7 (95% confidence interval [CI], 1.2-1.9); and negative likelihood ratio (LR-), 0.17 (95% CI, 0.03-0.73). Thus, the absence of night pain is helpful in reducing the probability of cancer, but its presence is minimally helpful. The absence of night pain should be interpreted together with the absence of other important findings to identify patients at low risk of back pain secondary to malignancy.
- The single-leg sit-to-stand test (described below) may be the most reliable method for detecting quadriceps weakness ($\kappa = 0.85$) which suggests upper lumbar (L3-L4) radiculopathy in patients with low back pain.⁶

IMPROVEMENTS IN THE DATA PRESENTED IN THE ORIGINAL PUBLICATION

No new findings substantially changed the results of those originally reported in the Rational Clinical Examination series.

Table 7-4 Estimated Accuracy of Ipsilateral Straight-leg Raise Test for Lumbar Disk Herniation

Source, Patient Population	LR+ (95% CI) or Range	LR- (95% CI) or Range
Jonsson and Stromqvist, ⁴ surgical series ^a (n = 300 patients)	2.0 (1.7-2.4)	0.21 (0.12-0.36)
van den Hoogen et al, ² surgical series (n = 7 studies)	0.99-1.8	0.04-0.54
Deville et al, ¹ surgical series (n = 10 studies)	1.1 (1.0-1.1)	0.34 (0.28-0.40)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

^aPatients with herniated disk were compared with patients with lateral or central stenosis.

Table 7-5 Estimated Accuracy of Crossed Straight-leg Raise Test for Lumbar Disk Herniation

Source, Patient Population	LR+ (95% CI, when data available)	LR- (95% CI, when data available)
Jonsson and Stromqvist, ⁴ surgical series ^a (n = 300 patients)	5.8 (2.7-12)	0.80 (0.72-0.90)
van den Hoogen et al, ² surgical series ^b (n = 6 studies)	1.6-8.8	0.59-90.0
Deville et al, ¹ surgical series (n = 6 studies)	2.2 (1.8-2.8)	0.81 (0.77-0.87)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

^aPatients with herniated disk were compared to patients with lateral or central stenosis.

^bVarious literature estimates were not pooled in this study.

Table 7-6 Estimated Accuracy of Sit-to-Stand Test for Upper Lumbar Disk (L3 to L4) Herniation With Radiculopathy

Source	LR+ (95% CI)	LR- (95% CI)
Rainville et al, ⁵ nonsurgical series ^a	26 (1.7-413)	0.35 (0.22-0.56)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

^aPatients with L3 to L4 radiculopathy were compared with patients with lower lumbar radiculopathy (L5-S1).

Table 7-7 Presence of Achilles Tendon Reflex in Patients Without a History of Low Back Pain, Sciatica, or Systemic Disease^a

Age, y	Total Patients	Both Present, % (95% CI)	Both Absent, % (95% CI)	One Absent, % (95% CI)
16-20	38	100 (92-100)	0 (0-8)	0 (0-8)
21-30	133	100 (98-100)	0 (0-2)	0 (0-2)
31-40	112	96 (93-100)	0.9 (0.8-3)	3 (0-6)
41-50	140	95 (90-98)	2.9 (0.1-6)	3 (0-6)
51-60	162	88 (83-93)	4 (1-6)	8 (4-12)
61-70	187	63 (56-70)	7 (3-10)	30 (23-60)
71-80	186	54 (47-61)	10 (5-14)	37 (30-43)
81-90	99	40 (31-50)	10 (4-16)	50 (40-59)
91-100	17	18 (0-36)	6 (0-17)	77 (66-87)

Abbreviation: CI, confidence interval.

^aFrequency may not total 100% because of rounding.

CHANGES IN THE REFERENCE STANDARD

The reference standard for a herniated disk causing radiculopathy continues to be surgical findings or the combination of clinical findings, imaging results, electrophysiology, and clinical course. No major new diagnostic techniques have been introduced. However, as suggested in the synthesis of literature of SLR, the choice of reference standard (imaging vs surgical findings) may influence estimates of test performance.

RESULTS OF LITERATURE REVIEW

Univariate Results of Tests for Herniated Lumbar Disk

The methods used in studying low back pain continue to be poor, leading to ambiguous results. As indicated in previous reviews, a clinical diagnosis is generally reached from multiple items of medical history and physical examination, with no single test sensitive and specific enough to make a definitive diagnosis.

Since the original Rational Clinical Examination article on back pain, 1 systematic review and a new surgical series have addressed the sensitivity and specificity of the SLR and crossed straight leg raise (CSLR) tests. These studies result in estimates close to those cited in the original Rational Clinical Examination article (Tables 7-4 and 7-5). The review article suggests a somewhat higher sensitivity of the SLR test (close to 0.90 rather than 0.80), whereas the surgical series reported somewhat greater specificity for the CSLR test (0.96 vs 0.90).

The sit-to-stand test is the most reliable test ($\kappa = 0.85$) for detecting quadriceps weakness, and it may discriminate those with an L3 to L4 herniation from those with an L5 to S1 lesion.⁶ To perform the single-leg sit-to-stand test, the patient attempts to rise from a chair by using only 1 leg. The patient is allowed to place his or her hand in the examiner's for aid with balance, and a negative finding/normal result is recorded if the patient is able to rise successfully (LR+, 26 [95% CI, 1.7-413]; LR-, 0.35 [95% CI, 0.22-0.56]) (Table 7-6). With regard to reflexes, a large study assessed whether absent Achilles reflexes occur in seemingly normal older patients (Table 7-7). The absence of an ankle reflex becomes increasingly common in individuals older than 60 years, suggesting that this finding is most meaningful at younger ages.

EVIDENCE FROM GUIDELINES

The 1994 guidelines on acute low back problems in adults prepared by the Agency for Health Care Policy and Research (now the US Agency for Healthcare Research and Quality) largely reiterated data from the original Rational Clinical Examination article. Guidelines on back pain from New Zealand, Australia, and Holland (published in 1995, 1996, and 2003, respectively) have no discussion on accuracy of the medical history and physical examination but recommend clinical evaluation consistent with the evaluation proposed here.

CLINICAL SCENARIO—RESOLUTION

Although this patient has some symptoms with SLR and has absent ankle reflexes, it is unlikely that he has neurologic deficits related to his low back pain. Some 30% of patients this age (>60 years) have absent ankle reflexes in the absence of low back pathology. The absence of pain radiating below the knee with SLR suggests that this patient's pain is most likely not the result of a lumbar radiculopathy. The absence of a positive CSLR result reinforces this impression. It is sometimes difficult to decide whether a patient is truly weak or whether strength testing effort is reduced by pain. However, this patient's normal sit-to-stand test result confirms normal strength. The combination of findings suggests that he does not have a herniated disc, so ordering additional tests (eg, electromyogram, nerve conduction, magnetic resonance imaging [MRI]) is not necessary.

See next page for the “Make the Diagnosis” section.

REFERENCES FOR THE UPDATE

1. Deville WLJM, van der Windt DAWM, Dzaferagic A, Bezemer PD, Bouter LM. The test of Lasègue: systematic review of the accuracy in diagnosing herniated discs. *Spine*. 2000;25(9):1140-1147.^a
2. van den Hoogen HM, Koes BW, van Eijk JT, Bouter LM. On the accuracy of history, physical examination, and erythrocyte sedimentation rate in diagnosing low back pain in general practice: a criteria-based review of the literature. *Spine*. 1995;20(3):318-327.^a
3. Bowditch MG, Sanderson P, Livesey JP. The significance of an absent ankle reflex. *J Bone Joint Surg Br*. 1996;78(2):276-279.^a
4. Jonsson B, Stromqvist B. Symptoms and signs in degeneration of the lumbar spine: a prospective, consecutive study of 300 operated patients. *J Bone Joint Surg Br*. 1993;75(3):381-385.^a
5. Rainville J, Jouve C, Finno M, Limke J. Comparison of four tests of quadriceps strength in L3 or L4 radiculopathies. *Spine*. 2003;28(21):2466-2471.^a
6. Vroomen PC, de Krom MCTFM, Knotterus JA. Diagnostic value of history and physical examination in patients suspected of sciatica due to disc herniation: a systematic review. *J Neurol*. 1999;246(10):899-906.

^aFor the Evidence to Support the Update on this topic,
see <http://www.JAMAEvidence.com>.

LOW BACK PAIN—MAKE THE DIAGNOSIS

PRIOR PROBABILITY

Because of the weak associations among symptoms, physical findings, imaging results, and electromyograms, a majority of patients with low back pain ($\approx 85\%$) cannot be given a definitive diagnosis. Among asymptomatic individuals, 20% to 30% have evidence of a herniated disk on computed tomography (CT) or MRI. However, only small portions (2%) of individuals with low back pain eventually undergo surgery for disk herniation. Thus, the prevalence of clinically important disk herniations is low.

In the primary care setting, the prevalence of compression fracture and spondylolisthesis is small, at 4% and 3%, respectively, in patients with low back pain. Fortunately, low back pain as a result of spinal malignancy, ankylosing spondylitis, or spinal infection is rare. The prevalence of these conditions among patients with back pain is approximately 0.7%, 0.3%, and 0.01%, respectively.

POPULATION FOR WHOM HERNIATED DISK WITH RADICULOPATHY SHOULD BE CONSIDERED

A herniated disk with radiculopathy should be considered in any adult with back and leg pain. Herniated disks causing sciatica are most common in middle-aged adults (30–55 years) and are somewhat less common in older adults (Table 7-8).

Table 7-8 Utility of the Clinical Examination for Herniated Disk or Cancer Among Patients With Back Pain

	LR+ (95% CI) or Range	LR- (95% CI) or Range
Sit-to-stand test for upper lumbar herniation	26 (1.7–413)	0.35 (0.22–0.56)
Nocturnal pain for cancer-induced back pain	1.7 (1.2–1.9)	0.17 (0.03–0.73)
Crossed straight-leg raise for disk herniation	1.6–5.8	0.59–0.90
Ipsilateral straight-leg raise for disk herniation	0.99–2.0	0.04–0.50

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR–, negative likelihood ratio.

POPULATION FOR WHOM CANCER SHOULD BE CONSIDERED

Although it accounts for less than 1% of patients with back pain, cancer is the most common of systemic causes. Cancer should be considered as a possible cause of low back pain in patients older than 50 years with low back pain and in patients with a history of cancer (especially prostate, lung, or breast carcinoma). In addition, patients who fail to improve after 4 to 6 weeks of conservative therapy should be evaluated for underlying systemic diseases such as cancer (Table 7-8).

REFERENCE STANDARD TESTS

For herniated disks, surgical findings may be a gold standard for diagnosis, but back surgery should never be considered just to confirm the absence of a disk hernia among patients with a negative clinical and imaging examination result. For patients who do not undergo surgery, CT or MRI demonstrating a disk herniation with nerve root impingement might be considered a gold standard. In addition, electromyography may confirm nerve root involvement. However, clinicians must realize that herniated disks on imaging are common among asymptomatic individuals. Thus, the imaging findings must be carefully correlated with clinical history, physical examination, and the time course of illness.

For metastatic cancer or infection, biopsy will be the usual gold standard, but these are performed only in patients with suggestive clinical and imaging findings. Imaging and laboratory test results (such as the erythrocyte sedimentation rate), if negative, are usually sufficient to rule out cancer and infection as a cause of back pain. For compression fractures, the gold standard remains imaging.

EVIDENCE TO SUPPORT THE UPDATE:

Low Back Pain

TITLE The Significance of an Absent Ankle Reflex.**AUTHORS** Bowditch MG, Sanderson P, Livesey JP.**CITATION** *J Bone Joint Surg [Br]*. 1996;78(2):276-279.**QUESTION** What is the prevalence of abnormal ankle reflexes in adults without pathologic causes of reflex loss?**DESIGN** Prospective.**SETTING** Orthopedic outpatient department in 2 hospitals.**PATIENTS** A total of 1074 patients (541 men, 533 women), aged 16 to 99 years, without history of spinal disease, low back pain, sciatica, diabetes mellitus, or neuropathic or systemic medical disease.**DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD**

Patients were examined in 3 positions: sitting with legs hanging over edge of seat, kneeling with feet over edge, and lying in supine and lateral positions. A reflex was considered present if it was elicited in any of the positions and absent if it was not. To determine interexaminer reliability, 50 patients were examined separately by each of the 3 authors ($\kappa = 0.94$).

MAIN OUTCOME MEASURES

The presence or absence of either 1 or both ankle reflexes was noted, and data were displayed according to age range in increments of 10 years. The prevalence and its 95% confidence interval for age group were calculated. The authors also tested for a relationship between prevalence and age with the χ^2 test. Finally, the results of each pair of consecutive groups were compared to determine at what age the largest changes in prevalence occurred.

MAIN RESULTS

See **Tables 7-9**, **7-10**, and **7-11**.

Table 7-9 Presence of Achilles Tendon Reflex in Patients Without a History of Low Back Pain, Sciatica, or Systemic Disease^a

Age, y	Total Patients	Both Present, % (95% CI)	One Absent, % (95% CI)	Both Absent, % (95% CI)
16-20	38	100 (92-100)	0 (0-8)	0 (0-8)
21-30	133	100 (98-100)	0 (0-2)	0 (0-2)
31-40	112	96 (93-100)	0.9 (0.8-3)	3 (0-6)
41-50	140	95 (90-98)	2.9 (0.1-6)	3 (0-6)
51-60	162	88 (83-93)	4 (1-6)	8 (4-12)
61-70	187	63 (56-70)	7 (3-10)	30 (23-60)
71-80	186	54 (47-61)	10 (5-14)	37 (30-43)
81-90	99	40 (31-50)	10 (4-16)	50 (40-59)
91-100	17	18 (0-36)	6 (0-17)	77 (66-87)

Abbreviation: CI, confidence interval.

^aFrequency may not total to 100% because of rounding.**Table 7-10** Significant Changes in Prevalence Between Consecutive Age Groups

Consecutive Age Groups, y	P Value for Both Ankle Reflexes Absent
51-60 vs 61-70	<.001
71-80 vs 81-90	.04

Table 7-11 A “Working” Guide

Age, y	Both Present, %	Both Absent, %	One Absent, %
<60	>90	<5	<5
>60	60	30	10

CONCLUSIONS**LEVEL OF EVIDENCE** Level 1.**STRENGTHS** Large number of participants in a prospective study with high interrater reliability.

LIMITATIONS Examiners were not blinded to the patient's age, although it is hard to do a practical study in which the examiners would have no idea of the patient's age.

The prevalence of ankle reflexes decreases with age. The largest decrements occur when comparing individuals in their 50s with those in their 60s and individuals in their 70s with those in their 80s. When using ankle reflexes to examine a patient for lumbar radiculopathy, the absence of an ankle reflex will be more meaningful in a patient younger than 60 years. Unilateral ankle reflex loss is far less common and is thus a more meaningful clinical sign, especially in younger patients.

Reviewed by Ben Stern, MS, DPT

TITLE The Test of Lasegue: Systematic Review of the Accuracy in Diagnosing Herniated Disks.

AUTHORS Deville WLJM, van der Windt DAWM, Dzaferagic A, Bezemer PD, Bouter LM.

CITATION *Spine*. 2000;25(9):1140-1147.

QUESTION How accurate are the straight leg raise (SLR) and cross straight leg raise (CSLR) tests at diagnosing a herniated disk in patients with low back pain?

DATA SOURCES

A MEDLINE and EMBASE search from an earlier review was extended to include 1992 through 1997 (keywords: "radiculopathy," "backache," "low back," "Lasegue," "straight leg raising," and "cross straight leg raising"). Bibliographies of retrieved studies were also reviewed for relevant material.

STUDY SELECTION

In total, 552 studies were retrieved; 15 met the inclusion criteria. Studies were selected if they used surgery as the reference standard, presented data on sensitivity or specificity, and included more than 10 patients with disease. Review articles were not included. The authors' original review (1995) included 19 studies, 12 of which were included in this review. The extended search through 1997 yielded 12 additional studies, of which 3 were retained for use.

DATA EXTRACTION

Two reviewers independently rated each study in 16 categories, including criteria related to internal and external validity (reference and index application and quality, spectrum of patients, setting, reproducibility, etc). The maximum possible score was 17, with 6 points on internal validity and 11 on

external. In addition, information on disease prevalence at the setting was collected.

MAIN RESULTS

Of the 15 studies included in this review, 7 included patients with previous disk surgery and 2 included patients with bilateral radiculopathy, both of whom had previous disk surgery. None of the studies occurred in a primary care setting. Positive SLR cutoff point was mentioned in 6 of the studies and ranged from less than 70 degrees ($n = 3$) to less than 90 degrees ($n = 2$). The addition of neck flexion or foot dorsiflexion was not evaluated. The median internal validity scores were 50% (range, 33%-66%) and 45% (range, 18%-72%), respectively. Median total validity score was 47% (range, 29%-65%), with 6 studies scoring 50% or better.

The authors included studies that were "sensitivity-only studies" (ie, only diseased patients), along with studies of diagnostic accuracy (patients with and without disk herniation). The pooled sensitivity of SLR was 0.91 (95% confidence interval [CI], 0.82-0.94) and pooled specificity was 0.26 (95% CI, 0.16-0.38). For the CSLR, pooled sensitivity was 0.29 (95% CI, 0.24-0.34) and specificity was 0.88 (95% CI, 0.86-0.90). See [Table 7-12](#).

CONCLUSIONS

LEVEL OF EVIDENCE Systematic review.

STRENGTHS Appropriate study question, literature search, and evaluation for bias.

LIMITATIONS The authors included sensitivity-only studies in their pooled estimates. However, they provide the data for all the studies that allow us to calculate the pooled likelihood ratios.

These data suggest that the SLR and CSLR should be used in combination. Although they are similar in overall accuracy (as evidenced by similar diagnostic odds ratio), the SLR primarily has value when it is absent (lowering the likelihood of a disk herniation), whereas the CSLR primarily has value when it is present (increasing the likelihood of a disk herniation).

Because all the studies included were surgical case series taken from hospitals and not from primary care facilities, an unusually high prevalence existed in these studies (86% for the

Table 7-12 SLR and CSLR as a Test for Disk Herniation^a

Test (n = No. of Studies)	LR+ (95% CI)	LR- (95% CI)
Straight leg raise (n = 10)	1.1 (1.0-1.1)	0.34 (0.28-0.40)
Crossed straight leg raise (n = 6)	2.2 (1.8-2.8)	0.81 (0.77-0.87)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

^aWe calculated the pooled likelihood ratio with random-effects measures. We used only studies that had sensitivity and specificity data. We excluded the outlier study noted by the authors.

SLR studies and 92% for the CSLR studies). The diagnostic odds ratio of the SLR decreased with designs of higher validity, homogeneity of case mix, and exclusion of patients with history of disk surgery. Both findings need better validation in populations of patients with a lower prevalence of disk herniation, such as those treated in primary care settings.

Reviewed by Ben Stern, MS, DPT

TITLE Symptoms and Signs in Degeneration of the Lumbar Spine.

AUTHORS Jonsson B, Stromqvist B.

CITATION *J Bone Joint Surg [Br]*. 1993;75(3):381-385.

QUESTION What are the frequencies of symptoms and neurologic disturbances among patients with spinal stenosis and lumbar disk herniation?

DESIGN Prospective study of patients consecutively admitted for lumbar spine surgery.

SETTING Inpatient surgery.

PATIENTS Three hundred patients admitted for lumbar disk or lumbar decompression surgery (100 disk herniation, 100 lateral stenosis, and 100 central stenosis).

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

Diagnosis was established with myelography, computed tomography, or magnetic resonance imaging and occasionally supplemented by nerve root block.

MAIN OUTCOME MEASURES

Sensitivity and specificity for a herniated lumbar disc: the likelihood ratios (LRs) represent the likelihood of a herniated disk (as opposed to central stenosis). When a finding is abnormal, the associated positive LR (LR+) of more than 1.0 makes a herniated disk more likely, whereas an LR+ of less than 1.0 makes central stenosis more likely. When a finding is normal, a negative LR (LR-) of more than 1.0 increases the likelihood of a disk herniation, whereas an LR- of less than 1.0 increases the likelihood of central stenosis.

MAIN RESULTS

See Table 7-13.

Table 7-13 Likelihood of Disk Herniation Versus Central Stenosis

	Disk Herniation	
	LR+ (95% CI) ^a	LR- (95% CI) ^b
SLR	2.0 (1.7-2.4)	0.21 (0.12-0.36)
CSLR	5.8 (2.7-12)	0.80 (0.72-0.90)
Patellar reflex	0.40 (0.22-0.73)	1.2 (1.1-1.4)
Ankle reflex	0.96 (0.75-1.2)	1.0 (0.82-1.30)
Sensory disturbance	1.3 (1.1-1.6)	0.70 (0.52-0.95)
No relief with rest	1.1 (1.0-1.3)	0.59 (0.35-1.0)

Abbreviations: CI, confidence interval; CSLR, crossed straight leg raise; LR+, positive likelihood ratio; LR-, negative likelihood ratio; SLR, straight leg raise.

^aLR+ greater than 1 favors disk herniation, whereas LR+ less than 1 favors central stenosis.

^bLR- greater than 1 favors disk herniation, whereas LR- less than 1 favors central stenosis.

CONCLUSIONS

LEVEL OF EVIDENCE Level 4.

STRENGTHS Differential diagnostic test evaluated among patients with known disease status (herniated disk vs central spinal stenosis).

LIMITATIONS The clinicians knew that all the patients had lesions.

Patients without evidence of spinal pathology were not included in this study, so it is difficult to know whether the results generalize to patients who have not yet had an imaging study or surgery. However, the data suggest that a positive response to CSLR increases the likelihood of disk herniation rather than central stenosis. A normal conventional SLR response favored central stenosis over disk herniation, whereas abnormal patellar reflexes decreased the likelihood of disk herniation (perhaps a counterintuitive finding). Given the limitation imposed by the study population, in which all patients had either lumbar stenosis or central stenosis, the other clinical results had limited or no ability to distinguish between the 2 diagnoses.

Reviewed by Ben Stern, MS, DPT

TITLE Comparison of Four Tests of Quadriceps Strength in L3 or L4 Radiculopathies.

AUTHORS Rainville J, Jouve C, Finno M, Limke J.

CITATION *Spine*. 2003;28(21):2466-2471.

QUESTION In adults with demonstrable L3 or L4 nerve root compression via computed tomography (CT) or magnetic resonance imaging (MRI), which of 4 tests of quadriceps strength best reflects evidence of a lesion at L3 or L4 vs L5 to S1? In other words, which tests best distinguish an upper lumbar radiculopathy from the far more common lower lumbar radiculopathy?

DESIGN Prospective, nonconsecutive patients with uncertainty in the independence of the clinical findings.

SETTING Outpatient physician office.

PATIENTS One group of participants recruited from a hospital spine center if they had lumbar radiculopathy and radiographically demonstrated evidence of displaced or compressed L3 or L4 nerve root on symptomatic side ($n = 33$: L3, $n = 10$; L4, $n = 23$). Another group of patients with L5 or S1 nerve root compression evidence via CT or MRI was asked to participate in the study as a comparison group ($n = 19$: L5, $n = 8$, S1, $n = 11$). Patients with bilateral radiculopathy, neurologic or muscular disease affecting the lower extremity (LE), evidence of symptom magnification, LE arthritis, cancer, cognitive dysfunction, and nonambulatory status were excluded. Average age of participants was 53 years, with an average duration of symptoms of approximately 2.8 months.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

The CT or MRI results served as the reference standard.

In addition to a routine physical examination, 4 tests of quadriceps strength were performed on each patient: (1) single-leg sit-to-stand, (2) step-up test, (3) knee-flexed manual muscle testing, and (4) knee-extended manual muscle test. For each maneuver, quadriceps strength was graded as abnormal (a positive result suggesting an L3-L4 lesion).

Patients with a normal result (negative likelihood ratio) were less likely to have an L3 to L4 lesion.

To perform the single-leg sit-to-stand, the participant attempted to rise from a chair by using only 1 leg. The participant was allowed to place her or his hands in the examiner's for aid with balance, and a score of normal was recorded if the participant was able to rise successfully. The step-up test was performed by asking the patient to step up on a 7-in stool (such as those built in to the end of an examining table). If the participant was able to step onto the stool successfully, a score of normal was recorded. The knee-flexed manual muscle test was performed in the supine position. The patient's leg was held distally near the ankle while the hip was flexed to 90 degrees and the knee was flexed to end range. The participant was then asked to straighten the leg toward the end of the table. Ability to straighten the leg against maximum resistance was recorded as normal. The knee-extended manual muscle test was also performed while the patient was supine. For this test, the examiner placed one hand above the participant's distal ankle and the other forearm under the participant's distal femur. The participant then straightened the knee, resulting in the heel's rising off the table. After this, the examiner attempted to bend the knee and touch the heel to the table while the participant offered maximum resistance. Ability to maintain the knee in extension was recorded as normal.

When available, a second examiner (blinded to the previous results) performed the tests on the participants again (39 of 53 participants).

In addition, patients completed questionnaires including items related to quadriceps weakness.

MAIN OUTCOME MEASURES

Frequency of detection of L3 to L4 vs L5 to S1 disk herniation as evidenced by imaging studies. The patients were evaluated for frequency of quadriceps weakness in L5 and S1 radiculopathies. In addition, κ values were used to determine inter-rater reliability of the 4 tests.

MAIN RESULTS

Thirty-three patients had an L3 to L4 lesion, whereas 19 had L5 to S1 nerve compression (Table 7-14).

Table 7-14 Quadriceps Strength as an Indicator of an L3 to L4 Lesion Among Patients With Nerve Root Compression

Quadriceps Strength for Each Maneuver	κ (Interobserver Agreement) (%)	LR+ for an L3-L4 Lesion (95% CI)	LR- (95% CI)
Sit to stand	0.85 (92)	26 (1.7-413)	0.35 (0.22-0.56)
Step up on stool	0.83 (95)	11 (0.69-182)	0.74 (0.59-0.92)
Manual muscle test, knee flexed	0.66 (84)	4.0 (1.0-16)	0.64 (0.46-0.90)
Manual muscle test, knee straight	0.08 (87)	4.1 (0.22-76)	0.92 (0.80-1.0)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

CONCLUSIONS

LEVEL OF EVIDENCE

Level 4.

STRENGTHS Differential diagnostic test evaluated among patients with known disease status (L3 to L4 vs L5 to S1 nerve root compression). An evaluation of the interobserver reliability was conducted.

LIMITATIONS It is not clear whether the authors were blinded to the level of nerve root compression. They did know that all patients had lesions. Height of chair was not specified for sit-to-stand test.

Controls without evidence of spinal pathology were not included in this study; thus, it is difficult to generalize to patients who have not yet had an imaging study or surgery. However, the excellent agreement among observers on watching the patient go from sit to stand or step up on a stool suggests that this may be a better way of evaluating quadriceps weakness than manual muscle testing.

Reviewed by Ben Stern, MS, DPT

DATA EXTRACTION

Studies were independently rated for methodology by 2 reviewers, with differences in rating resolved by consensus. Ratings for each study consisted of scores in categories for index and reference test quality, reference test application, independence, clinical description, study population, sample size, and data presentation. Sensitivity and specificity were calculated for each diagnostic test.

MAIN OUTCOME MEASURES

The mean total quality score for all studies was 55 of 100 (range, 20-85). The lowest scores fell in the categories of reference and index test quality, independence, clinical description, and study population. Only studies with scores greater than 55 were reviewed for diagnostic accuracy.

MAIN RESULTS

The data presented in Tables 7-15, 7-16, and 7-17 are the findings not reported in the original Rational Clinical Examination article on low back pain.¹

CONCLUSIONS

LEVEL OF EVIDENCE

Systematic review.

STRENGTHS Comprehensive review of articles with pre-defined selection criteria and a method for assessing quality.

TITLE On the Accuracy of History, Physical Examination, and Erythrocyte Sedimentation Rate in Diagnosing Low Back Pain in General Practice.

AUTHORS Van den Hoogen HM, Koes BW, van Eijk JT, Bouter LM.

CITATION *Spine*. 1995;20(3):318-327.

QUESTION How accurate are the medical history, physical examination, and erythrocyte sedimentation rate (ESR) in diagnosing various causes of low back pain?

DESIGN Systematic review.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

MEDLINE search was done (1986-1992) using the terms “backache,” or “low back,” and “sciatica,” “cancer,” or “spondylitis,” and bibliographies of retrieved studies were reviewed.

STUDY SELECTION

Studies were selected if they presented data on sensitivity or specificity of items in the medical history, physical examination, and ESR for radiculopathy, vertebral cancer or metastasis, and ankylosing spondylitis. Review articles and studies including fewer than 10 patients were excluded; 540 studies were retrieved, and 36 were included in this review (19 radiculopathy, 9 vertebral cancer, and 8 ankylosing spondylitis).

Table 7-15 Likelihood Ratios for Diagnosing Radiculopathy

Finding (No. of Studies)	Reference Standard	LR+ (95% CI) or Range	LR- (95% CI) or Range
Sciatica (Knutsson ² ; n = 205; patients with low back pain [21 with no radiculopathy])	Operative findings	0.92 (0.70-1.1)	1.5 (0.5-4.4)
Paresthesia (n = 2)	Operative findings	0.71-0.86	1.2-1.4
SLR (n = 7)	Operative findings	0.99-1.8	0.04-0.54
CSLR (n = 6)	Operative findings	1.6-8.8	0.59-0.90

Abbreviations: CI, confidence interval; CSLR, crossed straight leg raise; LR+, positive likelihood ratio; LR-, negative likelihood ratio; SLR, straight leg raise.

Table 7-16 Likelihood Ratios for Diagnosing Vertebral Cancer

Finding (No. of Studies)	LR+ Range	LR- Range
Spinal tenderness (n = 3)	0.38-3.6	0.26-1.4

Abbreviations: LR+, positive likelihood ratio; LR-, negative likelihood ratio.

Table 7-17 Likelihood Ratios for Diagnosing Ankylosing Spondylitis

Source	Finding	Sensitivity (95% CI)	Specificity (95% CI)	LR+ (95% CI)	LR- (95% CI)
Gran ³ (n = 449); patients with low back pain (27 with ankylosing spondylitis)	Out of bed at night	0.65 (0.48-0.81)	0.79 (0.75-0.83)	3.2 (2.3-4.4)	0.42 (0.25-0.72)
	No relief lying down	0.80 (0.63-0.92)	0.49 (0.44-0.54)	1.6 (1.3-2.0)	0.40 (0.17-0.84)
	Pain duration ≥ 3 mo	0.71 (0.52-0.84)	0.54 (0.49-0.59)	1.5 (1.2-2.0)	0.55 (0.30-1.0)
	Age at onset ≤ 35 y	0.92 (0.77-0.98)	0.30 (0.26-0.35)	1.3 (1.2-1.5)	0.25 (0.06-0.94)
	Morning stiffness	0.63 (0.44-0.79)	0.55 (0.51-0.60)	1.4 (1.0-1.9)	0.67 (0.41-1.1)
Mau et al ⁴ (n = 54); suspected of having ankylosing spondylitis (32 positive)	ESR raised	0.69	0.68		

Abbreviations: CI, confidence interval; ESR, erythrocyte sedimentation rate; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

LIMITATIONS Lack of specificity data of many included studies. The majority (33/36) of the studies included only hospital-based patients, thus limiting ability to generalize results.

None of the individual items in the medical history or physical examination were sufficiently useful in diagnosing ankylosing spondylitis, radiculopathy, or vertebral cancer. Rather than using single tests, clinicians must instead rely on the diagnostic value of a combination of the available clinical data.

Reviewed by Ben Stern, MS, DPT

REFERENCES FOR THE EVIDENCE

1. Deyo RA, Rainville J, Kent DL. What can the history and physical examination tell us about low back pain? *JAMA*. 1992;268(6):760-765.
2. Knutsson B. Comparative value of electromyographic, myelographic and clinical-neurologic examinations in diagnosis of lumbar root compression syndrome. *Acta Orthop Scand*. 1961;49(suppl):1-135.
3. Gran JT. An epidemiological survey of the signs and symptoms of ankylosing spondylitis. *Clin Rheumatol*. 1985;4(2):161-169.
4. Mau W, Zeidler H, Mau R, et al. Evaluation of early diagnostic criteria for ankylosing spondylitis in a 10 year follow-up. *Z Rheumatol*. 1990;49(2):82-87.

CHAPTER 8

Does This Patient Have Breast Cancer?

The Screening Clinical Breast Examination: Should It Be Done? How?

Mary B. Barton, MD, MPP

Russell Harris, MD, MPH

Suzanne W. Fletcher, MD, MSc

CLINICAL SCENARIOS

CASE 1 On annual examination of a 64-year-old woman, you observe an 8-mm mass in her right breast. She says she never noticed the mass before. Her screening mammogram result 7 months ago was normal.

CASE 2 A 42-year-old woman comes to see you because she is upset. "I want a breast examination, doctor. My coworker was just diagnosed with breast cancer." She practices breast self-examination regularly. She has observed no changes in her breasts.

WHY PERFORM A BREAST EXAMINATION?

The clinical breast examination (CBE), like any part of the physical examination, can be used either for screening (to detect breast cancer in asymptomatic women) or diagnosis (to evaluate breast complaints, primarily to rule out cancer). In primary care, screening CBEs are more commonly performed than diagnostic CBEs. Of a total of 14 859 CBEs performed on a cohort of 2400 women during a 10-year period, 73% were for screening and 27% were diagnostic¹ (Joann G. Elmore, MD, MPH, Harborview Medical Center, Seattle, Washington, written communication, November 1998). This review concentrates on the screening CBE because most research has been directed to screening rather than for diagnostic CBE. Because the screening CBE involves the search for cancer, there may be legal reasons, as well as medical reasons, for performing it well. Failure to diagnose breast cancer is a leading reason for malpractice claims, and lawsuits against primary care clinicians account for half the indemnity payments made.² Clinicians who do not perform careful screening may be more liable. Also, some women are more willing to accept screening CBE than mammography,³ in which case screening CBE is particularly important.

Anatomic Basis of the Breast Examination

The female breast consists of glandular and fibrous tissue and fat. Lobules of milk-producing glandular tissue radiate from the nipple, centrally supported by fibrous strands. Breast tissue, surrounded by superficial fascia, is attached to both the skin and the pectoral fascia by supporting ligaments. Fat surrounds the lobules of the breast, predominating in the superficial and peripheral portions. Breast tissue extends from the sternum medially to the midaxillary line laterally and from the clavicle superiorly to the "bra line" inferiorly, a rectangular rather than a circular area. The normal breast does not have a homogeneous texture but usually is somewhat lumpy on palpation.

Common distortions of the breast architecture include cysts, which are thought to arise from obstructed collecting ducts, and fibroadenomas, which are caused by an overgrowth of

periductal stromal connective tissue within the lobules of the breast. Other benign processes within the ductal system may cause a mass or nipple discharge such as mammary duct ectasia and intraductal papilloma. Most of these benign lesions carry no increased risk of breast cancer. One pathologic lesion, atypical hyperplasia, does increase risk by 3 to 5 times.⁴⁻⁶ Each of these benign processes may cause symptoms or signs that mimic malignancy.

Breast cancer is an unrestrained proliferation of cells arising in tissue of the ducts or lobules. Cancer arising from either type of tissue may be contained without spreading into surrounding stroma (ductal carcinoma in situ, and lobular carcinoma in situ) or may spread to contiguous tissues, through lymph channels, or hematogenously. Although ductal carcinoma in situ is a precursor lesion to invasive cancer, controversy surrounds its prognostic significance.^{7,8} Lobular carcinoma in situ is less common and is understood to be a marker for increased risk of development of invasive cancer, rather than a precursor lesion.⁹ Invasive breast cancer carries a 15.3% 5-year mortality rate¹⁰; advances in screening and treatment have contributed to a decrease in the mortality rate since 1989.^{11,12}

Risk Factors for Breast Cancer

Breast cancer is expected to occur in approximately 12% of American women during their lifetime.¹³ Breast cancer risk in the general population is most affected by age and family history. The annual incidence at age 70 years (1 in 200) is 20 times higher than that at age 30 years (1 in 4000) (Table 8-1).¹⁴ A woman with 2 first-degree relatives diagnosed as having breast cancer at an early age has a relative risk more than 4 times that of a woman without such a family history.¹⁵ Other risk factors are related to estrogen exposure (age of menarche, first pregnancy and menopause, parity, and estrogen replacement therapy¹⁵). Gail et al¹⁶ have developed a model to estimate the breast cancer risk of individual women according to known risk factors. Among a few women, genetic mutations in the *BRCA1* gene and, less commonly, *BRCA2* gene confer a high risk of breast cancer (50%-80% during a lifetime)¹⁷⁻¹⁹; women with these mutations account for only 3% of all breast cancer cases.²⁰

Clinically, strong risk factors affect the likelihood that any abnormality on CBE is cancer. For example, an abnor-

mal finding is more likely to be malignant in an older woman than in a younger woman. The Canadian National Breast Screening Study (NBSS)²¹ reported the positive predictive value for CBE to be twice as high in women from 50 through 59 years than in women from 40 through 49 years. In the Breast Cancer Detection Demonstration Project (BCDDP),²² the ratio of benign to malignant biopsy results decreased from 16.4 among women from 35 through 39 years to 3.2 for women from 60 through 69 years.

METHODS

We sought articles on effectiveness and test characteristics of the CBE. We identified potential English-language sources from the MEDLINE database for 1966 through 1997, using the search terms "physical examination," "palpation," "breast," "breast diseases," "diagnosis," "diagnostic tests," and "sensitivity and specificity." We reviewed all potentially relevant articles and the reference lists of these articles. In addition, other articles known to us and their references were reviewed. We contacted investigators of several studies for further clarification and, in some cases, for unpublished data. All authors reviewed and agreed on the studies selected for inclusion in the pooled analysis.

For information on the effectiveness of the CBE, we included all controlled trials and case-control studies in which CBE was at least a part of the screening modality.

Data on CBE techniques included information from both clinical studies and studies using silicone models of the breast. The data synthesis on test characteristics of screening CBE in human populations used the following criteria: (1) CBE performed on asymptomatic population, (2) all screening outcomes reported (ie, total numbers of screens and positive screens), (3) breast cancer outcome determined for all screens, within a defined follow-up period, and (4) all breast cancers histologically confirmed.

Summary measures for the sensitivity and specificity of the CBE and for likelihood ratios (LRs) of a positive or negative examination used published raw data from the reported trials that met our criteria. A random-effects model was used to generate conservative summary measures and confidence intervals (CIs).^{23,24}

EFFECTIVENESS OF CBE

Determining the effectiveness of screening CBE is difficult because no clinical trial has compared CBE alone with no screening. One randomized trial and one case-control study compared the combination of screening CBE and mammography with no screening and demonstrated statistically significant decreased breast cancer mortality rates of 20% and 71%, respectively, in women between the ages of 40 and 64 years^{25,26} (Table 8-2). These results, along with the evidence from randomized trials^{34,35} and case-control studies^{36,37} that screening mammography alone decreases breast cancer mortality rates, make designing a clinical trial in which the con-

Table 8-1 Incidence of Breast Cancer Within 1 Year for Women at a Given Age^a

Age, y	Breast Cancer Incidence
30	1 in 4000
40	1 in 800
50	1 in 400
60	1 in 300
70	1 in 200
80	1 in 200

^aData are from the United States and include all ethnicities from 1973-1995.¹⁴

Table 8-2 Studies of Breast Cancer Screening That Included Clinical Breast Examination

Study	Years	Examiners	Age of Women at Entry, y	No. of Women		Screening Modality		No. of Rounds	Years Followed Up	Mortality Reduction, RR (95% CI)						
				Intervention	Comparison	Intervention	Comparison									
Trials Comparing Screening Group With an Unscreened Group																
Randomized Controlled Trials																
HIP of New York ²⁵	1963-1966	Surgeons	40-64	30131	30565	CBE yearly; M yearly	None	4	18	0.77 (0.62-0.97)						
Edinburgh randomized trial of breast screening ²⁷	1979-1988	Physicians, nurses	45-64	22944	21344	CBE yearly; M alternate years	None	7	10	0.82 (0.61-1.1)						
Nonrandomized Controlled Trial																
UK Trial ^{28,29,a}	1979-1988	Physicians, nurses	45-64	45956	127109	CBE yearly; M alternate years	None	7	10	0.86 (0.73-1.0)						
Case-Control Study																
The DOM Project ^{30,31}	1974-1981	Medical assistants	50-64	14796 Invited: 54 cases 162 controls	... ^b	CBE yearly; M yearly	None	4	8	0.29 (0.14-0.62)						
Trials Comparing 2 Screening Strategies																
Canadian NBSS 1 ³²	1980-1988	Nurses	40-49	25214	25216	CBE yearly; M yearly	CBE 1 time only	5	7	1.4 (0.84-2.2)						
Canadian NBSS 2 ³³	1980-1988	Nurses	50-59	19711	19694	CBE yearly; M yearly	CBE yearly	5	7	0.97 (0.62-1.5)						

Abbreviations: CBE, clinical breast examination; CI, confidence interval; HIP, Health Insurance Plan; M, mammography; NBSS, National Breast Screening Study; RR, relative risk; UK, United Kingdom.

^aUK Trial includes data from the Edinburgh randomized trial.

^bEllipses indicate not applicable.

trol group members receive no screening unethical. It is unlikely that CBE alone will ever be compared with no screening in a randomized trial; therefore, we must use less direct evidence.

Meta-analyses of trials^{25-27,34-38} demonstrated that CBE or screening mammography decreases breast cancer mortality rates by about one-fourth in women from 50 through 69 years³⁹ and by 18% in women in their 40s.⁴⁰ In several of these studies, breast cancer was detected using a combination of CBE and mammography²⁵⁻²⁸ (Table 8-2). These studies that compared a combination screening strategy with no screening are the strongest scientific evidence for an effect of screening CBE.

Other evidence comes from the randomized Canadian NBSS 2,³³ in which women from 50 through 59 years were offered either a standardized CBE alone or a CBE and mammography annually for 5 years. The 7-year breast cancer-specific mortality rate for women in these 2 groups was similar,³³ suggesting that mammography may not offer mortality rate advantages over a careful screening CBE, at least for women in their 50s.⁴¹

Additional evidence comes from the Health Insurance Plan (HIP) study,⁴² conducted during mammography's infancy, in which most cancers were found by CBE. Mortality reduction after 10 years in the HIP trial of 29% was similar to a 30% reduction in the Swedish Two-County Trial,^{43,44}

which used mammography alone. The similarity in the percentage of reduced mortality rates found in these 2 approaches, along with the NBSS described above, argues for the effectiveness of carefully conducted CBE.

Finally, we compared the sensitivity of CBE and mammography in the trials that used both methods. In most cases, mammography outperformed CBE (Table 8-3). However, the sensitivity of the combined method was greater than that of mammography alone because CBE detected cancers that had been missed by mammography. The proportion of cancers detected by CBE alone ranged from 3.4% in the Edinburgh trial⁴⁵ to 45% in the HIP study.²⁵ Proportions of breast cancers found by CBE but missed by mammography in other studies⁴⁷⁻⁵⁸ range from 5.2%⁵⁸ to 29%.⁵¹ In one series, among women younger than 35 years, 23% of cancers were reported to be silent on mammography.⁵⁶

The value of detecting breast cancers by CBE that are not detected by mammography is not known. The results of randomized trials using both modalities did not demonstrate improved results over those using only mammography; however, the many other differences in the trials make comparisons difficult. The mortality rate in women in whom breast cancer is missed by mammography and detected by CBE was higher than that in women whose cancers were detected by mammography.^{25,32,33,59} However, these

Table 8-3 Proportion of Cancers Detected by CBE and Mammography Screening

Study	Years	No. of Cancers	Method of Detection, %		
			Mammography	CBE Only	Both
Randomized Controlled Trials					
HIP of New York ²⁵	1963-1966	132	33	45	22
Edinburgh randomized trial of breast screening ⁴⁵	1978-1981 ^a	88	26	3	71
Canadian NBSS 1 ³²	1980-1988	255	40	24	36
Canadian NBSS 2 ³³	1980-1988	325	53	12	35
Demonstration Projects					
BCDDP ²²	1973-1981	2045	40	9	50
West London ⁴⁶	1973-1977	29	34	31	34

Abbreviations: BCDDP, Breast Cancer Detection Demonstration Project; CBE, clinical breast examination; HIP, Health Insurance Plan; NBSS, National Breast Screening Study.

^aData are from prevalence screen only.

women still may have benefited compared with women not screened by CBE.

Bottom Line for Effectiveness

The strongest evidence for breast cancer mortality rate reduction after screening CBE comes from studies in which both CBE and mammography were part of breast cancer screening. The individual contribution of CBE cannot be established. In every study, CBE contributed to cancer detection independently of mammography. In one randomized trial, the 7-year breast cancer mortality rate was similar among women receiving a standardized CBE and women receiving both CBE and mammography.

Test Characteristics

Summarizing the precision and accuracy of CBE is difficult for several reasons. First, the examination is not well described in the majority of studies, and it is known that conduct of CBE varies widely.⁶⁰ Second, available studies included women differing in age, history of symptoms (symptomatic and asymptomatic), and practice settings (primary care or surgical). Third, the reported test characteristics of CBE were determined sometimes with and sometimes without accompanying mammography screening. The best standardized data come from studies of CBE on silicone models, but the applicability of these studies to women being screened is unknown.

Precision of Examination

Clinical breast examination, even when performed in large-scale studies, has generally not been standardized; only 1 trial

(NBSS) reported any description of the examination technique.⁶¹ The lack of attention to a standardized CBE technique may partly account for the interobserver variation found in studies among clinicians performing CBE.

Thomas et al⁶² compared findings in 103 women screened by 2 nurses and 2 surgeons independently. Agreement between the 2 nurses for any breast abnormality had a κ of 0.22, whereas the 2 surgeons' κ was 0.38. Chamberlain et al⁶³ studied agreement between a nurse and a physician performing independent screening CBE, with a κ of 0.43. Boyd et al⁶⁴ reported that 4 surgeons found 37 to 74 of 100 women screened to have abnormal findings; in only 25 women did all 4 agree on the findings. The κ value for agreement between any 2 of the 4 surgeons was between 0.34 and 0.59. None of these studies described the CBE technique used by examiners.

Precision varies by the particular physical finding. Ten surgeons examining 242 women had varying indices of agreement (which reflects the chance of agreement using the method of Kendall and Stuart⁶⁵) for specific findings: the index of agreement for nipple discharge was 14%; skin findings such as dilated veins, 22%; *peau d'orange*, 24%; ulceration, 62%; and visibility of lesion, 68%.⁶⁶ For a lump ("saturated nodule") the index of agreement was 59%.

Bottom Line for Precision

Clinicians using unstandardized CBE methods have demonstrated moderate degrees of agreement beyond that expected by chance. A standardized examination would likely improve precision.

ACCURACY

To determine its accuracy as a screening test, CBE must be compared with a criterion standard. Mammography cannot be that standard because cancers that are missed by mammography can be found on CBE. Histology alone also cannot be the standard because tissue will never be obtained from all women whose abnormalities are detected by CBE. Even less likely is the histologic examination of breasts that are normal on examination to determine specificity. A compromise criterion standard is to follow up all screened women for a defined period; women diagnosed as having breast cancer must have histologic proof, and all cases of breast cancer among women screened during the follow-up period must be counted. This admittedly imperfect standard nevertheless is so stringent that few studies of breast cancer screening^{22,25,32,33,67,68} meet it.

We defined sensitivity as the number of women who had cancer found on CBE, divided by the sum of screen-detected cancers (found by CBE or mammography) and those interval cancers diagnosed in the year after screening. Specificity was defined as the number of women who had normal CBE results and did not develop breast cancer during follow-up, divided by all the women without cancer at the end of the follow-up period.

The data show that sensitivity of CBE is far from perfect. Pooled data from human studies give an overall estimate for

the sensitivity of the CBE of 54% (95% CI, 48%-60%) (**Table 8-4**). Clinical breast examination sensitivity was higher than 60%^{32,33,67} when screening rounds included only physical examination but was lower when both CBE and mammography were used in the screening. This difference may reflect the enhanced case-finding capacity of mammography. However, 2 of the 3 studies with higher sensitivity also were the only ones using a well-described and standardized method of CBE.^{32,33} It is possible that CBE sensitivity was higher because of superior CBE technique.

The same trials provide data on the specificity of the CBE. Individual trial specificity ranged from 86% to 99%, with a pooled estimated specificity of 94% (95% CI, 90%-97%).

The combined data, pooled using a random-effects model to adjust for heterogeneity, indicate that the LR of a positive CBE result is 11 (95% CI, 5.8-19), whereas the LR of a negative test result is 0.47 (95% CI, 0.40-0.56). The positive LR is more discriminating than the negative LR, which is to say, a positive finding on examination conveys more information about an increased chance of cancer than does the finding of a benign examination offer certainty about the absence of breast cancer. This would be expected, given what we know about the frequent discovery by mammography of impalpable cancers.

Clinical breast examination is associated with a relatively high false-positive rate and an even higher false-negative rate. There are no data on the effect of the false-positive outcomes in terms of subsequent health care use or on women's psychological status, both of which have been issues for false-positive mammography results.^{1,69,70}

Lumps embedded in silicone breast models provide their own standard. Clinical breast examination sensitivity as measured in silicone models (40%-71%) was similar to that found in population studies.^{60,71-75} On the other hand, specificity measured in models was lower than in population studies (41%-77%).⁷¹⁻⁷⁵

The sensitivity of the CBE is approximately 54%. The specificity of the examination is about 94%.

Bottom Line for Accuracy

The sensitivity of the CBE is approximately 54%. The specificity of the examination is about 94%.

Examiner Factors

Studies in humans and silicone models demonstrate several factors, pertaining to both examiner and woman, that influence the accuracy of the CBE.

Duration of the Examination

Clinical breast examination duration correlated significantly with lump detection accuracy in experiments involving silicone breast models. In 5 studies, mean examination duration was always longer for examiners with higher sensitivity (**Table 8-5**). The highest recorded sensitivity in human studies (69%) was achieved in the NBSS, in which examiners took between 5 and 10 minutes to complete examination of both breasts.²¹

Technique

The use of correct CBE technique (a systematic search pattern, thoroughness, varying palpation pressure, 3 fingers, finger pads, and circular motion) also correlated with better examination sensitivity in silicone models (**Table 8-5**). The number of correct techniques was greater among examiners with higher CBE sensitivity.

Examiner Experience

Experience with abnormal breast lumps may be important. Even after controlling for technique differences, medical residents found more lumps in silicone models than lay women did before special training.⁷⁴ Almost none of the women had

Table 8-4 Sensitivity and Specificity of Clinical Breast Examination in Human Studies^a

Study	Years	Age, y	Screening Modality	No. of Rounds	CBE Sensitivity, %	CBE Specificity, %	LR+ (95% CI) ^b	LR- (95% CI) ^b
HIP of New York ²⁵	1963-1966	40-64	CBE and M	4	49	99	46 (39-54)	0.51 (0.44-0.59)
UK Trial ^{67,68}	1979-1988	45-64	CBE only	3	64	95	14 (12-16)	0.37 (0.29-0.48)
			CBE and M	4	51	... ^c		
Canadian NBSS 1 ³²	1980-1988	40-49	CBE only	1	69	86	4.8 (4.2-5.5)	0.36 (0.27-0.49)
			CBE and M	5	48	92	6.1 (5.4-6.8)	0.57 (0.50-0.63)
NBSS 2 ³³	1980-1988	50-59	CBE only	5	63	94	11 (9.6-12)	0.39 (0.33-0.46)
			CBE and M	5	40	94	7.2 (6.3-8.2)	0.63 (0.58-0.69)
BCDDP ⁵⁹	1973-1981	35-74	CBE and M	5	52
West London ^{45,d}	1973-1977	≥40	CBE and M	4	56	89
Pooled result (95% CI)					54 (48-60)	94 (90-97)	11 (5.8-19)	0.47 (0.40-0.56)

Abbreviations: BCDDP, Breast Cancer Detection Demonstration Project; CBE, clinical breast examination; CI, confidence interval; HIP, Health Insurance Plan; LR+, positive likelihood ratio; LR-, negative likelihood ratio; M, mammography; NBSS, National Breast Screening Study.

^aCase definition includes all cancers found at screening (by either method) and interval cancers found within 12 months of screening, except where noted otherwise.

^bAn LR is the probability that persons with a disease have a particular test result divided by the probability that persons without the disease have that result. The LR+ is determined by dividing the sensitivity by the probability of an abnormal CBE result among women without breast cancer (1 - specificity). The LR- is calculated as (1 - sensitivity)/specificity.

^cEllipses indicate not applicable.

^dSpecificity data based on first round only, with 6 months' follow-up.

Table 8-5 The Relationship Between Clinical Breast Examination Sensitivity and Duration or Techniques Used on Silicone Models^a

Study Participants	No. of Participants	Median Sensitivity, %	Mean Duration, min		Mean No. of Correct Techniques Used ^b	
			Sensitivity < Group Median	Sensitivity ≥ Group Median	Sensitivity < Group Median	Sensitivity ≥ Group Median
Women patients ⁷¹	260	44	1.5	1.9	2.9	3.7
Medical students ⁷⁶	151	100	2.3	2.8	2.7	3.7
Medical residents ⁷²	60	61	1.7	2.5	2.9	3.4
Practicing physicians ^c	60	55	1.9	2.4	2.3	2.7
Total ^d	531		1.8	2.3	2.8	3.6

^aIn each study, examiners were divided into 2 groups: those with examination sensitivity at or above the group median and those with sensitivity below the group median. Mean values for duration and numbers of correct techniques used are presented for these 2 groups.

^bOf a total of 6 correct techniques: systematic search pattern, thorough examination, varying palpation pressure, 3 fingers, pads of fingers, and small circular motion.

^cRussell Harris, MD, MPH, University of North Carolina at Chapel Hill, written communication, February 1999.

^dP<.001 for pooled differences in both duration and number of techniques.

ever felt either a real or simulated breast lump before the testing session, whereas 77% of the physicians had. Among the residents, previous experience also predicted higher sensitivity. After practice with silicone models containing embedded lumps, the women's abilities approached that of physicians.⁷¹ However, 2 other studies found no differences in sensitivity across categories thought to correlate with experience.^{60,77}

Bottom Line for Examiner Influence on Accuracy

Spending adequate time on the CBE and using the proper techniques improve breast lump detection.

Patient Factors

Age

On average, younger women have denser breasts that make lump detection more difficult, whereas in older women, the breast becomes more fatty, making lump detection easier.⁷⁸ In one referral population, examiners' sensitivity was 86% among women aged 20 through 49 years and 96% among women aged 50 years and older.⁵⁹ Silicone models simulating postmenopausal breast tissue improved sensitivity over that in models simulating premenopausal breast tissue (64% vs 51%).⁷⁵ Two large trials came to a different conclusion, albeit among women in narrowly defined age ranges. The BCDDP found CBE sensitivity of 53% among women between 40 and 49 years and 48% among women between 50 and 59 years.²² The NBSS⁷⁹ reported higher CBE sensitivity in women aged 40 through 49 years (68%) compared with those aged 50 through 59 years (63%), among women receiving both mammography and CBE. Further study is needed on this issue.

Breast Characteristics

Clinical breast examination sensitivity is slightly lower in women with larger breasts.⁸⁰ Women's breasts also vary in the amount of background glandular nodularity that is a normal characteristic of breast tissue.⁸¹ Many women have ill-defined fibrocystic changes that make their breasts feel particularly lumpy; anecdotally, clinicians (and women) find it more difficult to detect breast cancer in lumpy breasts.

Cancer Characteristics

Breast cancers vary in size, hardness, mobility, and location in the breast. Clinical breast examination sensitivity probably varies according to these characteristics of cancers. Prognosis generally follows cancer size at diagnosis, so it is important to determine the accuracy of CBE for small cancers, that is, 2 cm or less. In the BCDDP, sensitivity for noninfiltrating cancers was 35%; for infiltrating cancers smaller than 1 cm, 36%; and for infiltrating cancers at least 1 cm, 52%.²²

To date, most information about CBE accuracy by lump characteristic comes from experiments carried out on silicone breast models with embedded lumps varying in size, hardness, and placement. These experiments found sensitivity increased with lump size (from 14% for 3-mm lumps to 79% for 1-cm lumps) and hardness (from 42% for 20-durometer lumps to 72% for 60-durometer lumps). Durometers are a measure of hardness; 20 durometers corresponds to a soft to medium-hardness grape, whereas a 60-durometer mass is almost as hard as calcified bone. Medium or deep placement of the lump in a model did not alter sensitivity.^{59,72,74}

Bottom Line for Patient Effects on Accuracy

A woman's age and the size and lumpiness of her breasts may affect the ability of examiners to detect cancer. Size and hardness of breast cancers also affect CBE sensitivity.

Suggested Approach

Many physical diagnosis textbooks give directions for carrying out a breast examination.⁸²⁻⁸⁵ They all involve palpation and inspection, but research has stressed palpation. The approach outlined below is derived from a review of the research literature and owes much to the work of Baines,^{3,86} Baines et al,²¹ Baines and Miller⁷⁹ and others⁸⁷⁻⁹¹ because of their work in standardizing the examination. Our recommendation incorporates practices from the Mammacare Method because its components have been validated in independent investigations of CBE technique.^{71,72,92}

Palpation

Variables important in palpating the breast correctly are patient position; breast boundaries; examination pattern; finger position, movement, and pressure; and duration of the examination.

Patient Position

Clinical breast examination requires flattening breast tissue against the patient's chest; she should be supine during the examination. The importance of maneuvers to flatten the breast depends on breast size; they are particularly useful in women with large breasts. To flatten the lateral part of the breast, have the patient roll onto her contralateral hip, rotate her shoulders back into a supine position, and place her ipsilateral hand on her forehead (Figure 8-1). To flatten the medial part of the breast, the woman should lie flat on her back and move her elbow up until it is level with her shoulder (Figure 8-1).

Breast Boundaries

Breast tissue extends laterally toward the axilla and superiorly toward the clavicle. To be sure that all breast tissue is examined, it is best to cover a rectangular area bordered by the clavicle superiorly, the midsternum medially, the midaxillary line laterally, and the bra line inferiorly.

Examiner Pattern

Palpation begins in the axilla and extends in a straight line down the midaxillary line to the bra line (Figure 8-1). The fingers then move medially, and palpation continues up the chest in a straight line to the clavicle. The entire breast is covered in this manner, going up and down between the clavicle and the bra line. To examine all breast tissue, rows should be overlapping. This vertical strip pattern (or lawnmower technique) was found to be more thorough than concentric circles or a radial spoke pattern.⁹² In one study, two-fifths of physicians used no discernible pattern at all.⁶⁰

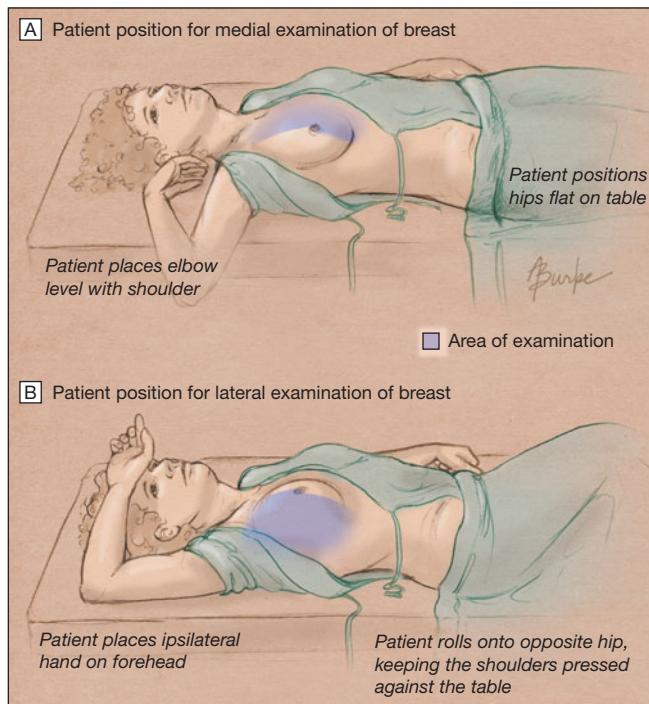


Figure 8-1 Position of Patient and Direction of Palpation for the Clinical Breast Examination

The figure shows the positioning of the patient for examining the (A) medial and (B) lateral portions of the breast. See "Suggested Approach" section for complete description.

Fingers

Most texts scarcely describe what the fingers should do during palpation, an ironic situation because the fingers must detect and differentiate abnormal lumps in breast tissue. Behavioral psychologists have shown that the finger can detect a soft (20-durometer) 2-mm lump in simulated breast tissue when specific techniques are used.^{88,90,93} These researchers developed a breast palpation technique (the Mammacare Method) combining the vertical strip pattern and specific finger techniques, taught using discrimination skill practice (with the use of silicone breast models) to enhance lump detection. Their method is described below.

The 3 middle fingers are held together, with the metacarpal-phalangeal joint slightly flexed. The pads (not tips) of the fingers (Figure 8-2) are the examining surface. (Confusion

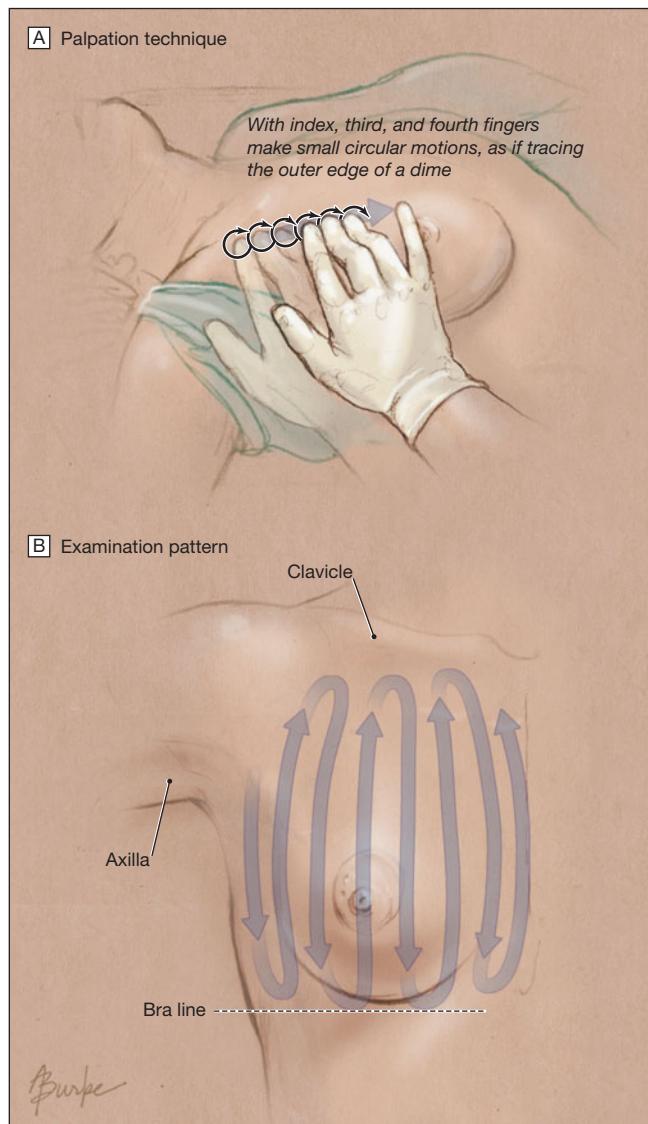


Figure 8-2 Palpation Technique

Pads of the index, third, and fourth fingers make small circular motions (A), as if tracing the outer edge of a dime. A vertical strip pattern (B) ensures an examination of the entire breast.

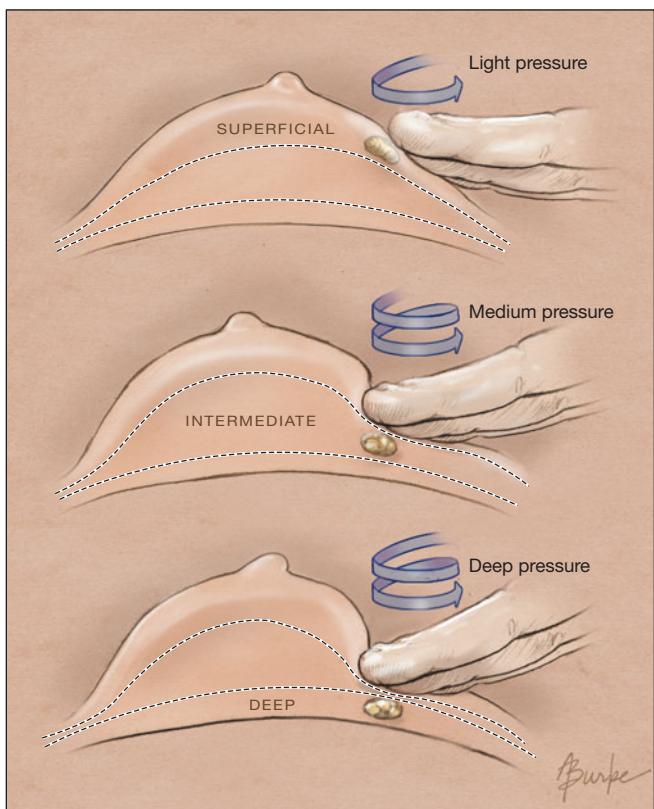


Figure 8-3 Levels of Pressure for Palpation of Breast Tissue Shown in a Sagittal View of the Right Breast

The examiner should make 3 circles with the finger pads, increasing the level of pressure (light, medium, and deep) with each circle.

regarding the definition of the finger pad exists even among experienced examiners.⁸⁶ Each area is palpated by making small circles as if following the edge of a dime (Figure 8-2). At each spot, 3 circles with different pressures—light, medium, and deep—are made to ensure palpation of all levels of tissue (Figure 8-3).

Duration

A careful examination of an average-sized breast (brassiere size B) takes at least 3 minutes (6 minutes for both breasts). This is much longer than the average 1.8 minutes physicians spent in one study examining both breasts and giving instructions for breast self-examination.⁹⁴ If it seems awkward to spend this amount of time, clinicians should discuss with patients the time needed to do a complete examination and discuss the procedure during the examination.

Other Issues

Palpation of the supraclavicular and axillary regions to detect adenopathy is a standard part of the CBE, though untested. Breast cancer was found in a minority of women with isolated axillary lymphadenopathy and normal CBE results in 2 series (12% and 29%, respectively).^{95,96}

Palpation of the nipple area is performed in the same manner as the rest of the breast. Although some texts call for squeezing the nipple to express discharge,^{44,82,83,97} among 448 women complaining of nipple discharge, expression of fluid

was not a useful prognostic sign for cancer. Of the women with otherwise normal CBE findings, 3 (2%) of the 151 women with spontaneous discharges were diagnosed as having cancer, whereas none (0%) of the 178 women with discharges only apparent by expression were diagnosed as having cancer.⁹⁸

Inspection

The importance of inspection is unproved. Most commonly, directions for inspection suggest that the woman face the examiner with her arms at her side. The breasts are then inspected for nipple abnormalities, dimpling, and retraction or tethering of the skin. No adequate data support recommendations of some authorities^{61,99,100} to examine women in a variety of other positions, such as raising her hands over her head, putting her hands on her hips and bearing down (to contract the pectoral muscles), or leaning forward to allow the breasts to hang out from the chest.

In a series of 296 breast cancers found on breast examination,¹⁰¹ 96% were discovered on palpation, only 1% by retraction alone, and another 3% by visible nipple abnormalities. The women's position when these visual cues were elicited was not reported. Inspection and positioning the patient for inspection take time. Given these facts and given the press of time, we suggest that in asymptomatic women clinicians should concentrate on careful breast palpation, all the while, of course, using their eyes. If the patient is symptomatic, or if an abnormality is discovered during palpation of an asymptomatic patient, careful inspection should be added.

Bottom Line of the Suggested Approach

Use a vertical strip pattern to cover all the breast tissue. Make circular motions with the pads of the middle 3 fingers and examine each breast area with 3 different pressures. Spend at least 3 minutes on each breast.

Teaching the Technique

What is the evidence that using the Mammacare Method improves lump detection abilities and that the technique can be taught?

In one study, 20 lay women taught according to the Mammacare Method doubled their detection of known breast lumps in other volunteer women, although they also increased the number of false-positive detections after training.⁸⁹ Three randomized trials using silicone breast models evaluated training of internal medicine residents, graduate nurses, medical students, and female patients.⁷¹⁻⁷³ All showed that training improved CBE sensitivity when measured on silicone models. Pooling the results, the training improved sensitivity by 13 percentage points (95% CI, 10%-16%) from 46% to 59%, whereas the specificity declined nonsignificantly by a mean of 4 points (95% CI, -8.9 to 0.7) from 61% to 57%.

Does the effect of teaching persist? In one study, 91 patients were taught the Mammacare Method and, 1 year later, were able to find more lumps in silicone breast models than women either taught the traditional (circular) CBE pattern or not taught at all.⁷¹ Similar results occurred in randomized studies using silicone models with medical students and nurses,^{72,76} with the effect persisting from 4 to 6 months.

In most cases, sensitivity improved without adverse effects on specificity. However, among medical residents, higher sensitivity was at the expense of specificity in silicone model testing. A 6-month medical record review of patients cared for by these physicians did not demonstrate any deterioration in CBE specificity in patients.⁷²

Are Lumps Ever Normal?

Normal breasts are often lumpy; the clinician's job is to distinguish normal from abnormal (cancerous) lumps. Cancers classically are characterized as hard, fixed, and irregular, whereas benign breast lumps are the opposite: soft or cystic, movable, and regular. However, many cancers do not conform to the classic picture, and benign masses can mimic cancers. LRs for the presence of these signs (calculated from HIP data,¹⁰² after Mushlin¹⁰³) are unimpressive, except for fixed lesions (LR, 2.4) and lumps greater than 2 cm (LR, 1.9); none of the LRs fall in the range considered discriminating (Table 8-6). Table 8-6 also shows the resulting succession of probabilities if a 64-year-old woman had a mass on CBE and if the mass had the listed positive findings. (It is assumed that the findings are independent, although there is not information about the independence of the findings.) In 2400 women undergoing 10905 screening CBEs in a community setting during a 10-year period, an abnormal CBE result was associated with an LR of 2.1 (Joann G. Elmore, MD, MPH, Harborview Medical Center, Seattle, Washington, written communication, June 1998). A positive screening CBE result in an average-risk woman conveys less risk of cancer than does a woman presenting with a breast lump (LR, 55)¹⁰⁴ or an abnormal screening mammogram result (LR, 26).¹⁰⁵

Because the characteristics of cancerous lumps overlap with those of noncancerous lumps, clinicians rarely diagnose breast cancer with CBE. Careful CBE can locate abnormalities. Further evaluation with other tests is then required.¹⁰⁶⁻¹⁰⁸

THE BOTTOM LINE

Screening CBEs should be conducted for women who are at risk for breast cancer and for whom breast cancer screening has been shown effective. Presently, this includes women older than 40 years. A well-conducted CBE can detect at least 50% of asymptomatic cancers and may contribute to mortality rate reduction in women screened.

Resolution of Scenarios

The discovery of a breast mass in a 64-year-old patient conveys an increased risk of cancer. Her pretest probability of invasive cancer in the coming year is 0.35% (347 cases per 100 000 women¹⁴). Your finding on CBE gives a posttest probability of 0.73% (Table 8-6). If the mass is greater than 2 cm and has all the other malignant characteristics, the probability of cancer increases to 8.8% (Table 8-6).

The 42-year-old woman with no breast symptoms has a pretest probability of breast cancer of 0.12%, or 119 per 100 000.¹⁴ A normal CBE result would decrease her risk of

Table 8-6 Breast Cancer Probabilities in a 64-Year-Old Woman Assessed After Each of a Succession of Positive Findings^a

Prior Probability of Breast Cancer, %	Prior Odds	Finding	LR+ ^b	Successive Posterior Odds ^c	Successive Posterior Probability, %
0.35	0.0035	Mass	2.1	0.007	0.73
		Fixed	2.4	0.018	1.7
		Hard	1.6	0.028	2.8
		Irregular	1.8	0.051	4.9
		≥2-cm Lump	1.9	0.097	8.8

Abbreviation: LR+, positive likelihood ratio.

^aThe effect of a particular finding is expressed in the following way: prior odds × likelihood ratio (LR) = posterior odds. Probabilities and odds are interconverted according to these formulae: prior odds = prior probability/(1 – prior probability); posterior probability = posterior odds/(1 + posterior odds).

^bLRS are calculated from data on cases diagnosed through June 1970 in the Health Insurance Plan Breast Cancer Screening Study,¹⁰² after Mushlin.¹⁰³

^cThe LR for each positive finding is applied to the posterior odds from the line above, using an assumption that the findings contribute independently to the odds of breast cancer.

breast cancer to 0.11%, but with such a low baseline risk, the difference is hard to appreciate. An explanation of her low pretest probability may suffice; however, the psychological reassurance she may gain from a CBE could increase the value of this maneuver.

Priorities for Research

Standardization of CBE is sorely needed. Numerous studies suggest that the Mammacare Method improves the performance characteristics of CBE on silicone models; further work should be done to determine whether the Mammacare technique (or other standardized methods) can improve CBE sensitivity and specificity in patient populations. The contribution of visual inspection has been found to be associated with better outcomes in women who use it as part of breast self-examination.¹⁰⁹ This should be investigated as to its contribution to the CBE.

Screening CBE may be particularly useful in women older than 70 years because fatty changes in the breast make lump detection easier, and older women do not accept mammography as readily as younger women.¹¹⁰ Comparison of test characteristics of standardized CBE with mammography in older women is needed. At the other end of the age spectrum, because mammography misses substantial numbers of breast cancers in women younger than 50 years, studies are needed to determine whether standardized CBE can contribute to decreasing breast cancer mortality rates in this age group.

The cost-effectiveness of CBE screening deserves study if it is to be compared with other maneuvers available for breast cancer screening and compared with other primary care maneuvers that it may displace in a 15-minute visit. Similarly, cost-effectiveness of programs to teach providers how to perform the examination should be evaluated.

Although some argue that the CBE adds nothing to regular mammography screening, an overall view of the evidence suggests that a carefully performed CBE detects cancers that are potentially curable. If research confirms that CBE is as effective as mammography in reducing breast cancer mortality rates for older women, then physicians will want to perform CBE regularly and perform it well.

Author Affiliations at the Time of the Original Publication

Department of Ambulatory Care and Prevention, Harvard Pilgrim Health Care and Harvard Medical School, Boston, Massachusetts (Drs Barton and Fletcher); and Department of Medicine, University of North Carolina, Chapel Hill, North Carolina (Dr Harris).

Acknowledgments

This work was supported by the Harvard Pilgrim Health Care Foundation, Boston, Massachusetts.

We thank Joanne T. Piscitelli, MD, and Joann G. Elmore, MD, MPH, for thoughtful comments and manuscript review; Sara Moore, MPH, for research assistance.

REFERENCES

1. Elmore JG, Barton MB, Moceri VM, Polk S, Arena PJ, Fletcher SW. Ten-year risk of false-positive screening mammograms and clinical breast examinations. *N Engl J Med.* 1998;338(16):1089-1096.
2. Physician Insurers Association of America. *Breast Cancer Study—June 1995.* Washington, DC: Physician Insurers Association of America; 1995.
3. Baines CJ. Physical examination of the breasts in screening for breast cancer. *J Gerontol.* 1992;47:63-67.
4. Dupont WD, Page DL. Risk factors for breast cancer in women with proliferative breast disease. *N Engl J Med.* 1985;312(3):146-151.
5. Dupont WD, Page DL, Parl FF, et al. Long-term risk of breast cancer in women with fibroadenoma. *N Engl J Med.* 1994;331(1):10-15.
6. London SJ, Connolly JL, Schnitt SJ, Colditz GA. A prospective study of benign breast disease and the risk of breast cancer. *JAMA.* 1992;267(7):941-944.
7. Welch HG, Black WC. Using autopsy series to estimate the disease “reservoir” for ductal carcinoma in situ of the breast. *Ann Intern Med.* 1997;127(11):1023-1028.
8. Fonseca R, Hartmann LC, Petersen IA, Donohue JH, Crotty TB, Gisvold JJ. Ductal carcinoma in situ of the breast. *Ann Intern Med.* 1997;127(11):1013-1022.
9. Wood WC. Management of lobular carcinoma in situ and ductal carcinoma in situ of the breast. *Semin Oncol.* 1996;23(4):446-452.
10. Ries LAG, Kosary CL, Hankey BF, et al. *SEER Cancer Statistics Review, 1973-1996.* Bethesda, MD: National Cancer Institute; 1999: Table IV-3. http://seer.cancer.gov/csr/1973_1996/breast.pdf. Accessed May 19, 2008.
11. Chu KC, Tarone RE, Kessler LG, et al. Recent trends in US breast cancer incidence, survival and mortality rates. *J Natl Cancer Inst.* 1996;88(21):1571-1579.
12. Wingo PA, Ries LAG, Rosenberg HM, Miller DS, Edwards BK. Cancer incidence and mortality, 1973-1995. *Cancer.* 1998;82(6):1197-1207.
13. Feuer EJ, Wun LM, Boring CC, et al. The lifetime risk of developing breast cancer. *J Natl Cancer Inst.* 1993;85(11):892-897.
14. Ries LAG, Kosary CL, Hankey BF, Miller BA, Edwards BK. *SEER Cancer Statistics Review, 1973-1995.* Bethesda, MD: National Cancer Institute; 1998: Table IV-2. http://seer.cancer.gov/csr/1973_1995/breast.pdf. Accessed May 19, 2008.
15. Hulka B, Stark AT. Breast cancer: cause and prevention. *Lancet.* 1995; 346(8979):883-887.
16. Gail MH, Brinton LA, Byar DP, et al. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *J Natl Cancer Inst.* 1989;81(24):1879-1886.
17. Krainer M, Silva-Arrieta S, FitzGerald MG, et al. Differential contributions of BRCA1 and BRCA2 to early-onset breast cancer. *N Engl J Med.* 1997;336(20):1416-1421.
18. Easton DF, Bishop DT, Ford D, Crockford GP. Genetic linkage analysis in familial breast and ovarian cancer. *Am J Hum Genet.* 1993;52(4):678-701.
19. Rowell S, Newman B, Boyd J, King MC. Inherited predisposition to breast and ovarian cancer. *Am J Hum Genet.* 1994;55(5):861-865.
20. Whittemore AS, Gong G, Itnyre J. Prevalence and contribution of BRCA1 mutations in breast cancer and ovarian cancer. *Am J Hum Genet.* 1997;60(3):496-504.
21. Baines CJ, Miller AB, Bassett AA. Physical examination: its role as a single screening modality in the Canadian National Breast Screening Study. *Cancer.* 1989;63(9):1816-1822.
22. Baker LH. The Breast Cancer Detection Demonstration Project. *CA Cancer J Clin.* 1982;32(4):194-225.
23. Eddy DM, Hasselblad V, Shachter RD. *Meta-analysis by the Confidence Profile Method: The Statistical Synthesis of Evidence.* San Diego, CA: Academic Press; 1992:93-98.
24. Eddy DM, Hasselblad V. *Fast*Pro V1.8: Software for Meta-analysis by the Confidence Profile Method.* San Diego, CA: Academic Press; 1992.
25. Shapiro S, Venet W, Strax P, Venet L. *Periodic Screening for Breast Cancer: The Health Insurance Plan Project and Its Sequelae, 1963-1986.* Baltimore, MD: Johns Hopkins University Press; 1988.
26. Collette HJA, Rombach JJ, Day NE, de Waard F. Evaluation of screening for breast cancer in a non-randomised study (the DOM project) by means of a case-control study. *Lancet.* 1984;1(8388):1224-1226.
27. Alexander FE, Anderson TJ, Brown HK, et al. The Edinburgh randomised trial of breast cancer screening. *Br J Cancer.* 1994;70(3):542-548.
28. UK Trial of Early Detection of Breast Cancer Group. Breast cancer mortality after 10 years in the UK trial of early detection of breast cancer. *Breast.* 1993;2:1-8.
29. Moss SM, Coleman DA, Ellman R, et al. Interval cancers and sensitivity in the screening centres of the United Kingdom trial of early detection of breast cancer. *Eur J Cancer.* 1993;29A(2):255-258.
30. Collette HJA. Attempts to evaluate a non-randomized breast cancer screening programme (the DOM-project). *Maturitas.* 1985;7(1):43-50.
31. Collette HJA, de Waard F, Rombach JJ, et al. Further evidence of benefits of a (non-randomised) breast cancer screening programme: the DOM project. *J Epidemiol Community Health.* 1992;46(4):382-386.
32. Miller AB, Baines CJ, To T, Wall C. Canadian National Breast Screening Study, 1. *CMAJ.* 1992;147(10):1459-1476.
33. Miller AB, Baines CJ, To T, Wall C. Canadian National Breast Screening Study, 2. *CMAJ.* 1992;147(10):1477-1488.
34. Tabar L, Fagerberg CJ, Gad A, et al. Reduction in mortality from breast cancer after mass screening with mammography. *Lancet.* 1985;1(8433):829-832.
35. Andersson I, Aspegren K, Janzon L, et al. Mammographic screening and mortality from breast cancer. *BMJ.* 1988;297(6654):943-948.
36. Palli D, Del Turco MR, Buiatti E, et al. A case-control study of the efficiency of a non-randomized breast cancer screening program in Florence (Italy). *Int J Cancer.* 1986;38(4):501-504.
37. Verbeek ALM, Holland R, Sturmans F, Hendriks JHCL, Mravunac M, Day NE. Reduction of breast cancer mortality through mass screening with modern mammography. *Lancet.* 1984;1(8388):1222-1224.
38. UK Trial of Early Detection of Breast Cancer Group. First results on mortality reduction in the UK trial of early detection of breast cancer. *Lancet.* 1988;2(8608):411-416.
39. Kerlikowske K, Grady D, Rubin SR, Sandrock C, Ernster VL. Efficacy of screening mammography. *JAMA.* 1995;273(2):149-154.
40. Hendrick RE, Smith RA, Rutledge JH, Smith CR. Benefit of screening mammography in women aged 40-49. *Natl Cancer Inst Monogr.* 1997; 22:87-92.
41. Mittra I. Breast screening: the case for physical examination without mammography. *Lancet.* 1994;343(8893):342-344.
42. Thomas LB, Ackerman LV, McDivitt R. *Report of NCI Ad Hoc Pathology Working Group to Review the Gross and Microscopic Findings of Breast Cancer Cases in the HIP Study.* Bethesda, MD: National Cancer Institute; 1977. NIH Publication No. 77-1400.
43. Tabar L, Fagerberg CJG, South MC, Day NE, Duffy SW. The Swedish Two-County Trial of mammographic screening for breast cancer. In: Miller AB, Chamberlain J, Day NE, Hakama M, Prorok PC, eds. *Cancer Screening.* Cambridge, England: Cambridge University Press; 1991:23-44.

44. Tabar L, Fagerberg G, Duffy SW, Day NE. The Swedish Two-County Trial of mammographic screening for breast cancer. *J Epidemiol Community Health*. 1989;43(2):107-114.
45. Roberts MM, Alexander FE, Anderson TJ, et al. Edinburgh trial of screening for breast cancer: mortality at seven years. *Lancet*. 1990;335 (8684):241-246.
46. Chamberlain J, Clifford RE, Nathan BE, Price JL, Burn I. Error-rates in screening for breast cancer by clinical examination and mammography. *Clin Oncol*. 1979;5(2):135-146.
47. Di Pietro S, Fariselli G, Bandieramonte G, et al. Diagnostic efficacy of the clinical-radiological-cytological triad in solid breast lumps. *Eur J Surg Oncol*. 1987;13(4):335-340.
48. Edeiken S. Mammography and palpable cancer of the breast. *Cancer*. 1988;61(2):263-265.
49. Meyer KK. Diagnostic error in breast disease. *Am Surg*. 1975;41(12):774-785.
50. Hermansen C, Skovgaard Poulsen H, Jensen J, et al. Diagnostic reliability of combined physical examination, mammography, and fine needle puncture ("triple-test") in breast tumors: a prospective study. *Cancer*. 1987;60(8):1866-1871.
51. Day PJ, O'Rourke MGE. The diagnosis of breast cancer. *Med J Aust*. 1990;152(12):635-639.
52. Bender HG, Schnurch HG, Beck L. Breast cancer detection: age-related significance of findings on physical exam and mammography. *Gynecol Oncol*. 1988;31(1):166-172.
53. Van Dam PA, Van Goethem MLA, Kersschot E, et al. Palpable solid breast masses: retrospective single- and multimodality evaluation of 201 lesions. *Radiology*. 1988;166(2):435-439.
54. Butler JA, Vargas HI, Worthen N, Wilson SE. Accuracy of combined clinical mammographic-cytologic diagnosis of dominant breast masses. *Arch Surg*. 1990;125(7):893-896.
55. Ashley S, Royle GT, Corder A, et al. Clinical, radiological and cytological diagnosis of breast cancer in young women. *Br J Surg*. 1989;76(8):835-837.
56. Ciatto S, Smith AH, DiMaggio C, et al. Breast cancer diagnosis under the age of forty years. *Tumori*. 1987;73(5):457-461.
57. Hansell DM, Cooke JC, Parsons CA. The accuracy of mammography alone and in combination with clinical examination and cytology in the detection of breast cancer. *Clin Radiol*. 1988;39(2):150-153.
58. Burns PE, May C, Gutter Z, et al. Relative accuracy of clinical examination and mammography in a breast clinic in Alberta. *J Can Assoc Radiol*. 1978;29(1):22-27.
59. Seidman H, Gelb SK, Silverberg E, LaVerda N, Lubera JA. Survival experience in the Breast Cancer Detection Demonstration Project. *CA Cancer J Clin*. 1987;37(5):258-290.
60. Fletcher SW, O'Malley MS, Bunce LA. Physicians' abilities to detect lumps in silicone breast models. *JAMA*. 1985;253(15):2224-2228.
61. Bassett AA. Physical examination of the breast and breast self-examination. In: Miller AB, ed. *Screening for Cancer*. Orlando, FL: Academic Press Inc; 1985:271-291.
62. Thomas DC, Spitzer WO, MacFarlane JK. Inter-observer error among surgeons and nurses in presymptomatic detection of breast disease. *J Chronic Dis*. 1981;34(12):617-626.
63. Chamberlain J, Rogers P, Price JL, Ginks S, Nathan BE, Burn I. Validity of clinical examination and mammography as screening tests for breast cancer. *Lancet*. 1975;2(7943):1026-1030.
64. Boyd NF, Sutherland HJ, Fish EB, et al. Prospective evaluation of physical examination of the breast. *Am J Surg*. 1981;142(3):331-334.
65. SAS Institute, Inc. *SAS/STAT User's Guide, Version 6, Fourth Edition, Volume 1*. Cary, NC: SAS Institute, Inc; 1989:867.
66. Yorkshire Breast Cancer Group. Observer variation in recording clinical data from women presenting with breast lesions. *BMJ*. 1977;2(6096): 1196-1199.
67. UK Trial of Early Detection of Breast Cancer Group. Specificity of screening in UK Trial of Early Detection of Breast Cancer. *BMJ*. 1992; 304:346-349.
68. Chamberlain J, Coleman D, Ellman R, et al. Sensitivity and specificity of screening in the UK Trial of Early Detection of Breast Cancer. In: Miller AB, Chamberlain J, Day NE, Hakama M, Prorok PC, eds. *Cancer Screening*. Cambridge, England: Cambridge University Press; 1991:3-17.
69. Welch HG, Fisher ES. Diagnostic testing following screening mammography in the elderly. *J Natl Cancer Inst*. 1998;90(18):1389-1392.
70. Barton MB, Moore S, Polk S, et al. Anxiety and health care utilization after false-positive mammograms [abstract]. *J Gen Intern Med*. 1999;14(suppl):9.
71. Fletcher SW, O'Malley MS, Earp JL, Morgan TM, Lin S, Degnan D. How best to teach women breast self-examination. *Ann Intern Med*. 1990;112 (10):772-779.
72. Campbell HS, Fletcher SW, Lin S, Pilgrim CA, Morgan TM. Improving physicians' and nurses' clinical breast examination. *Am J Prev Med*. 1991;7(1):1-8.
73. Campbell HS, McBean M, Mandin H, Bryant H, A'Hern RP. Teaching medical students how to perform a clinical breast examination. *Acad Med*. 1994;69(12):993-995.
74. Fletcher SW, O'Malley MS, Pilgrim CA, Gonzalez JJ. How do women compare with internal medicine residents in breast lump detection? *J Gen Intern Med*. 1989;4(4):277-283.
75. McDermott MM, Dolan NC, Huang J, Reifler D, Rademaker AW. Lump detection is enhanced in silicone breast models simulating postmenopausal breast tissue. *J Gen Intern Med*. 1996;11(2):112-114.
76. Pilgrim C, Lannon C, Harris RP, Cogburn W, Fletcher SW. Improving clinical breast examination training in a medical school. *J Gen Intern Med*. 1993;8(12):685-688.
77. Lee KC, Dolan NC, Dunlop D. Do clinical breast examination skills improve during medical school [abstract]? *J Gen Intern Med*. 1997;12 (suppl 1):93.
78. Stomper PC, D'Souza DJ, DiNitto PA, Arredondo MA. Analysis of parenchymal density on mammograms in 1353 women 25-79 years old. *AJR Am J Roentgenol*. 1996;167(5):1261-1265.
79. Baines CJ, Miller AB. Mammography versus clinical examination of the breasts. *J Natl Cancer Inst*. 1997;22:125-129.
80. Strax P, Venet L, Shapiro S, Gross S. Mammography and clinical examination in mass screening for cancer of the breast. *Cancer*. 1967;20(12):2184-2188.
81. Love SM, Gelman RS, Silen W. Fibrocystic "disease" of the breast—a nondisease? *N Engl J Med*. 1982;307(16):1010-1014.
82. Seidel HM, Ball JW, Dains JE, Benedict GW. *Mosby's Guide to Physical Examination*. 2nd ed. St. Louis, MO: Mosby; 1991:371-402.
83. Willms JL, Schneiderman H, Algranati PS. *Physical Diagnosis: Bedside Evaluation of Diagnosis and Function*. Baltimore, MD: Lippincott Williams & Wilkins; 1994:221-241.
84. Bickley LS. *Bate's Guide to Physical Examination and History Taking*. Philadelphia, PA: JB Lippincott Co; 1999:340-347.
85. Sapiro JD. *The Art & Science of Bedside Diagnosis*. Baltimore, MD: Urban & Schwarzenberg; 1990:239-243.
86. Baines CJ. Breast palpation technique: what is the finger pad? *J Chronic Dis*. 1987;40(4):361-362.
87. Stephenson HS, Adams CK, Hall DC, Pennypacker HS. Effects of certain training parameters on detection of simulated breast cancer. *J Behav Med*. 1979;2(3):239-250.
88. Bloom HS, Criswell EL, Pennypacker HS, Catania AC, Adams CK. Major stimulus dimensions determining detection of simulated breast lesions. *Percept Psychophys*. 1982;32(3):251-260.
89. Hall DC, Adams CK, Stein GH, et al. Improved detection of human breast lesions following experimental training. *Cancer*. 1980;46(2):408-414.
90. Adams CK, Hall DC, Pennypacker HS, et al. Lump detection in simulated human breast. *Percept Psychophys*. 1976;20:163-167.
91. Madden MC, Hench LL, Hall DC, et al. Model breasts for use in teaching breast self-examination. *J Biomed Eng*. 1978;2(5):427-435.
92. Saunders KJ, Pilgrim CA, Pennypacker HS. Increased proficiency of search in breast self-examination. *Cancer*. 1986;58(11):2531-2537.
93. Hall DC, Goldstein MK, Stein GH. Progress in manual breast examination. *Cancer*. 1977;40(1):364-370.
94. Kahn KL, Goldberg RJ. Screening for breast cancer in the ambulatory setting [abstract]. *Clin Res*. 1984;32:649A.
95. Leibman AJ, Kossoff MB. Mammography in women with axillary lymphadenopathy and normal breasts on physical examination. *Am J Roentgenol*. 1992;159(3):493-495.
96. de Andrade JM, Marana HR, Sarmento Filho JM, Murta EF, Velludo MA, Bighetti S. Differential diagnosis of axillary masses. *Tumori*. 1996;82(6): 596-599.
97. Orel SG. High-resolution MR imaging for the detection, diagnosis, and staging of breast cancer. *Radiographics*. 1998;18(4):903-912.
98. Gulay H, Bora S, Kilictepe S, Hamaloglu E, Goksel HA. Management of nipple discharge. *J Am Coll Surg*. 1994;178(5):471-474.
99. Winchester DP. Physical examination of the breast. *Cancer*. 1992;69(7 suppl):1947-1949.

100. Lawrence HC. History, physical examination, and education in breast self-examination. *Clin Obstet Gynecol.* 1994;37(4):881-886.
101. Mahoney L, Csima A. Efficiency of palpation in clinical detection of breast cancer. *CMAJ.* 1982;127(8):729-730.
102. Venet L, Strax P, Venet W, Shapiro S. Adequacies and inadequacies of breast examinations by physicians in mass screening. *Cancer.* 1971;28(6):1546-1551.
103. Mushlin AI. Diagnostic tests in breast cancer. *Ann Intern Med.* 1985; 103(1):79-85.
104. Barton MB, Elmore JG, Fletcher SW. Breast complaints among women enrolled in a health maintenance organization. *Ann Intern Med.* 1999; 130(8):651-657.
105. Kerlikowske K, Grady D, Barclay J, Sickles EA, Ernster V. Likelihood ratios for modern screening mammography. *JAMA.* 1996;276(1):39-43.
106. Donegan WL. Evaluation of a palpable breast mass. *N Engl J Med.* 1992;327(13):937-942.
107. The palpable breast lump: information and recommendations to assist decision-making when a breast lump is detected. *CMAJ.* 1998;158 (suppl 3):S3-S8.
108. Fletcher SW, Barton MB. Evaluation of breast lumps. In: *UpToDate in Medicine* [book on CD-ROM]. Boston, MA: UpToDate Inc; 1999.
109. Harvey BJ, Miller AB, Baines CJ, Corey PN. Effect of breast self-examination techniques on the risk of death from breast cancer. *CMAJ.* 1997;157(9):1205-1212.
110. White E, Urban N, Taylor V. Mammography utilization, public health impact, and cost-effectiveness in the United States. *Annu Rev Public Health.* 1993;14:605-633.

Prepared by Mary B. Barton, MD

Reviewed by Kathryn A. Myers, MD, EdM

CLINICAL SCENARIO

A 55-year-old woman without a family history of breast or ovarian cancer and without a personal history of mantle radiation, suggesting average risk of breast cancer, comes to your office requesting a magnetic resonance imaging (MRI) for screening for breast cancer. Will the findings on a clinical breast examination (CBE) affect the likelihood of breast carcinoma?

UPDATED SUMMARY ON BREAST CANCER

Original Review

Barton MB, Harris R, Fletcher SW. Does this patient have breast cancer? the screening clinical breast examination: should it be done? how? *JAMA*. 1999;282(13):1270-1280.

UPDATED LITERATURE SEARCH

We searched the PubMed database for the period October 1998 to September 2004, using the terms "breast" and "palpation," in combination with the original search strategy, including the terms "physical exam," "professional competence," "medical history taking," "sensitivity," "specificity," "observer variation," "reproducibility of results," "diagnostic tests," and "Bayes theorem." The search was limited to articles in English and indexed as human studies. Seventy-five articles were identified, and their abstracts were reviewed. Seventeen potentially eligible articles were retrieved according to their abstracts, and the articles and their reference lists were reviewed. In addition, the titles of 43 articles that had referenced the original review were reviewed and, of these, the abstracts of an additional 9 were considered. A total of 23 articles were read for salience and quality. For the purpose of updating the information synthesis on the characteristics of CBE in human screening populations, only 1 article contained data for both the sensitivity and specificity of the CBE,¹ and an additional article provided data on sensitivity only.² No studies have been published with relevant information on the effectiveness of CBE during this interval.

Several articles with information relevant to the teaching of the CBE are included in this summary review.

NEW FINDINGS

- Although finding a breast lesion on clinical examination increases the likelihood of cancer (likelihood ratio [LR], approximately 9), in community-based settings, the positive predictive value was low (2.9%-4.3%).
- The maximum expected sensitivity in asymptomatic women in current general community practice is 36%.
- About 5% of all breast cancer cases were detected by CBE alone.

Details of the Update

Although no major advances in knowledge about the pathophysiology of breast cancer have been made, the public level of concern and the controversy around breast cancer screening continue to be high. In the last 5 years, there have been scientific debates on the utility of mammography³ and newspaper exposés on the variability in quality of mammography reading in the United States.⁴ The publication of negative data from 2 trials of breast self-examination^{5,6} resulted in a repeated "insufficient evidence" recommendation from the US Preventive Services Task Force (USPSTF)⁷ and led to downgrading the recommendation of teaching this practice to "not recommended" by the Canadian Task Force.⁸ National Health Interview Survey data indicate that use of the CBE has decreased during the last 10 years, whereas substantially more women reported recent mammography in 2000 than in 1990 (Table 8-7).⁹

Table 8-7 Mammography Screening Is Increasing as Clinical Breast Examination Is Decreasing

Age, y	% Reported CBE in 1990	% Reported CBE in 2000	% Reported Mammography in 1990	% Reported Mammography in 2000
40-49	83	76	55	65
50-64	78	79	56	79
65+	71	68	43	68

Abbreviation: CBE, clinical breast examination.

At the same time, new studies of MRI for breast cancer screening in high-risk populations¹⁰⁻¹⁴ have generated public interest in this technology. MRI in an average-risk population has not been studied and would likely not be feasible because of the low positive predictive value that would accompany the use of such a highly sensitive test in a population with a low prevalence of breast cancer.

A new large series of CBEs has been published by Bobo et al¹ (see Table 8-8). This series of 752081 CBEs reported an overall sensitivity of 59%; 5.1% of all cancers diagnosed were found only by CBE (ie, were found in women with an abnormal CBE result and a normal mammogram result). The sensitivity must be viewed with 3 caveats: many women presenting for examination in the Bobo et al study did so because of concern due to patient-observed palpable findings (discovered on self-examination or by accident) or skin or nipple changes; the sensitivity of the CBE in these women was 85% vs 36% for women without symptoms (eg, a true screening population). Second, although women with an abnormal CBE or mammogram result were followed carefully to the resolution of the finding, there was no systematic follow-up of women with normal examination results. Only about 25% of these women returned the following year; for this reason, the sensitivity estimate of the screening CBE must be seen as an upper limit of the true sensitivity in this population. Third, the technique for CBE was not standardized across the many study sites, nor were any efforts at ensuring the quality of the examination described. Because of these limitations, we did not revise the LR estimates for the clinical examination in detecting breast cancer during screening evaluations. These caveats aside, the main finding of this study, that CBE in the community could contribute to breast cancer detection, is supported and is important from an effectiveness point of view.

Oestreicher et al² reported on 468 women with breast cancer who had taken part in a managed care organization's breast cancer screening program. In that program, CBE detected 35% of tumors diagnosed within 1 year of screening, and 5.8% of the cancers were diagnosed by CBE and not detected by mammography. Factors significantly associated in a multivariable model with lower sensitivity of the CBE were age younger than 50 years or older than 80 years and increased body weight (defined as ≥ 135 lb [61.2 kg]). Better sensitivity was associated with Asian race (compared with white) and tumors greater than 1 cm. Although this sample is

Table 8-8 Clinical Breast Examination Characteristics Change When Women Have Breast Symptoms¹

Finding	LR+ (95% CI)	LR- (95% CI)
Asymptomatic women	9.5 (8.9-10)	0.66 (0.64-0.69)
Symptomatic women ^a	2.5 (2.4-2.6)	0.22 (0.20-0.25)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

^aSymptoms in the breast that cause a woman to present to a physician include pain, finding of a lump, nipple discharge, or other change in the nipple or the skin, each of which is associated with an increased risk of breast cancer.

small compared with that in the Bobo et al¹ study, there are striking similarities in both the sensitivity of the screening CBE and the proportion of cancers found only by CBE.

Costanza et al¹⁵ described the results of a trial using standardized patients to teach CBE skills to practicing clinicians; those completing a 5-hour training session had improved performance on each of 7 separate components of CBE technique. Vetto et al¹⁶ provided CBE training with silicone models to 205 practicing primary care physicians and found in a pretest-posttest design that lump detections increased significantly (proportion finding from 3 to 5 of 5 lumps went from 59% to 94%; $P < .001$) and false-positive detections decreased significantly (27% with 2 or more before training, and 15% after training, $P < .004$).

IMPROVEMENTS IN THE DATA PRESENTED IN THE ORIGINAL PUBLICATION

None.

CHANGES IN THE REFERENCE STANDARD

None.

RESULTS OF LITERATURE REVIEW

See Table 8-8.

EVIDENCE FROM GUIDELINES

Recent guidelines regarding the CBE remain as they were in 1999: an "I" recommendation (ie, the USPSTF concludes that the evidence is insufficient to recommend for or against routinely providing the service) from the USPSTF¹⁷ and consensus-based recommendations for annual screening from the American Cancer Society (every 3 years for women aged 20-39 years and annually thereafter)¹⁸ and the American College of Obstetricians and Gynecologists (annually for all women).¹⁹ Breast self-examination by patients is now "not recommended" by the USPSTF and Canadian Task Force.

CLINICAL SCENARIO—RESOLUTION

A 55-year-old woman at average risk of breast cancer should have a careful medical history taken to elicit symptoms, be advised of the benefits and risks of screening mammography,²⁰ and be offered a CBE. She should not be offered MRI for screening according to the data available at this time. If a CBE is performed, the LRs would suggest the following according to the findings of the examination: with a baseline risk of cancer of 1 in 350 (or 2.8 per 1000) in the coming year, a normal examination result (LR, 0.47) suggests a decrease in her risk to 1 in 744. An abnormal examination result (LR+, 11) would suggest an increase in risk to 1 in 33, and she should be referred for further investigations and treatment.

BREAST CANCER—MAKE THE DIAGNOSIS

PRIOR PROBABILITY

The risk of breast cancer increases as a function of age. The lifetime risk for US women is 12%. The annual risk is shown in **Table 8-9**.

Table 8-9 Breast Cancer Risk Increases With Age²¹

Age, y	Incidence, %
30	0.025
40	0.125
50	0.25
60	0.33
70	0.5
80	0.5

POPULATION FOR WHOM A SCREENING CBE SHOULD BE CONSIDERED

- Women who would be considered for mammography screening (eg, women 40 years and older) should be offered a CBE (**Table 8-10**).
- Women with a positive family history for breast cancer may benefit from breast cancer screening starting at a younger age.

Table 8-10 Detecting the Likelihood of Breast Cancer

Finding ^a	LR+ (95% CI)	LR- (95% CI)
CBE	11 (5.8-19)	0.47 (0.40-0.56)

Abbreviations: CBE, clinical breast examination; CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

^aPooled results based on 7 studies.²²⁻²⁸

REFERENCE STANDARD TESTS

The reference standard for determining the outcomes of screening for breast cancer continues to be temporal follow-up, preferably 1 year after any negative testing result to ascertain true-negative status.

REFERENCES FOR THE UPDATE

1. Bobo JK, Lee NC, Thames SF. Findings from 752,081 clinical breast examinations reported to a national screening program from 1995 through 1998. *J Natl Cancer Inst.* 2000;92(12):971-976.^a
2. Oestreicher N, White E, Lehman CD, Mandelson MT, Porter PL, Taplin SH. Predictors of sensitivity of clinical breast examination (CBE). *Breast Cancer Res Treat.* 2002;76(1):73-81.^a
3. Gotzsche PC, Olsen O. Is screening for breast cancer with mammography justifiable? *Lancet.* 2000;355(9198):129-134.
4. Moss M. Spotting breast cancer: doctors are weak link. *New York Times.* June 27, 2002:A-1.
5. Semiglavov VF, Moiseenko VM, Manikahas AG. Interim results of a prospective randomized study of self-examination for early detection of breast cancer (Russia/St. Petersburg/WHO). *Voprosy Onkologii.* 1999;45(35):265-271.
6. Thomas DB, Gao DL, Ray RM, et al. Randomized trial of breast self-examination in Shanghai: final results. *J Natl Cancer Inst.* 2002;94(19):1445-1457.
7. US Preventive Services Task Force. Screening for breast cancer: recommendations and rationale. *Ann Intern Med.* 2002;137(5 pt 1):344-346.
8. Baxter N, and the Canadian Task Force on Preventive Health Care. Preventive health care, 2001 update: should women be routinely taught breast self-examination to screen for breast cancer? *CMAJ.* 2001;164(13):1837-1846.
9. Meissner HI, Breen N, Yabroff KR. Whatever happened to clinical breast examinations? *Am J Prev Med.* 2003;25(3):259-263.
10. Krieg M, Brekelmans CTM, Boetes C, et al. Efficacy of MRI and mammography for breast-cancer screening in women with a familial or genetic predisposition. *N Engl J Med.* 2004;351(5):427-437.
11. Warner E, Plewes DB, Hill KA, et al. Surveillance of BRCA1 and BRCA2 mutation carriers with magnetic resonance imaging, ultrasound, mammography, and clinical breast examination. *JAMA.* 2004;292(11):1317-1325.
12. Kuhl CK, Schmutzler RK, Leutner CC, et al. Breast MR imaging screening in 192 women proved or suspected to be carriers of a breast cancer susceptibility gene: preliminary results. *Radiology.* 2000;215(1):267-279.
13. Podo F, Sardanelli F, Canese R, et al. The Italian multi-centre project on evaluation of MRI and other imaging modalities in early detection of breast cancer in subjects at high genetic risk. *J Exp Clin Cancer Res.* 2002;21(3 suppl):115-124.
14. Tilanus-Linthorst MM, Obdeijn IM, Bartels KC, de Koning HJ, Oudkerk M. First experiences in screening women at high risk for breast cancer with MR imaging. *Breast Cancer Res Treat.* 2000;63(1):53-60.
15. Costanza ME, Luckmann R, Quirk ME, Clemow L, White MJ, Stoddard AM. The effectiveness of using standardized patients to improve community physician skills in mammography counseling and clinical breast exam. *Prev Med.* 1999;29(4):241-248.
16. Vetto JT, Petty JK, Dunn N, Prouser NC, Austin DF. Structured clinical breast examination (CBE) training results in objective improvement in CBE skills. *J Cancer Educ.* 2002;17(3):124-127.
17. Humphrey LL, Helfand M, Chan BKS, Woolf SH. Breast cancer screening: a summary of the evidence for the US Preventive Services Task Force. *Ann Intern Med.* 2002;137(5):347-367.
18. Smith RA, Saslow D, Sawyer KA, et al. American Cancer Society guidelines for breast cancer screening: update 2003. *CA Cancer J Clin.* 2003;53(3):141-169.
19. ACOG practice bulletin: breast cancer screening: number 42, April 2003. *Int J Gynaecol Obstet.* 2003;81(3):313-323.
20. Fletcher SW, Elmore JG. Mammographic screening for breast cancer. *N Engl J Med.* 2003;348(17):1672-1680.
21. Ries LAG, Kosary CL, Hankey BF, et al. *SEER Cancer Statistics Review, 1973-1996.* Bethesda, MD: National Cancer Institute; 1999: Table IV-3. http://seer.cancer.gov/csr/1973_1996/breast.pdf. Accessed May 19, 2008.
22. Shapiro S, Venet W, Strax P, Venet L. *Periodic Screening for Breast Cancer: The Health Insurance Plan Project and Its Sequelae, 1963-1986.* Baltimore, MD: Johns Hopkins University Press; 1988.
23. United Kingdom Trial of Early Detection of Breast Cancer Group. Specificity of screening in United Kingdom trial of early detection of breast cancer. *BMJ.* 1992;304:346-349.
24. Chamberlain J, Coleman D, Ellman R, et al. Sensitivity and specificity of screening in the UK Trial of Early Detection of Breast Cancer. In: Miller AB, Chamberlain J, Day NE, Hakama M, Prorok PC, eds. *Cancer Screening.* Cambridge, England: Cambridge University Press; 1991:3-17.

25. Miller AB, Baines CJ, To T, Wall C. Canadian National Breast Screening Study, 2: breast cancer detection and death rates among women aged 50 to 59 years. *CMAJ.* 1992;147(10):1477-1488.
26. Miller AB, Baines CJ, To T, Wall C. Canadian National Breast Screening Study, 1: breast cancer detection and death rates among women aged 40 to 49 years. *CMAJ.* 1992;147(10):1459-1476.
27. Seidman H, Gelb SK, Silverberg E, LaVerda N, Lubera JA. Survival experience in the breast cancer detection demonstration project. *CA Cancer J Clin.* 1987;37(5):258-290.
28. Chamberlain J, Clifford RE, Nathan BE, Price JL, Burn I. Error-rates in screening for breast cancer by clinical examination and mammography. *Clin Oncol.* 1979;5:135-146.

^aFor the Evidence to Support the Update for this topic,
see <http://www.JAMAevidence.com>.

8

EVIDENCE TO SUPPORT THE UPDATE:

Breast Cancer

TITLE Findings From 752081 Clinical Breast Examinations Reported to a National Screening Program From 1995 Through 1998.

AUTHORS Bobo JK, Lee NC, Thames SF.

CITATION *J Natl Cancer Inst.* 2000;92(12):971-976.

QUESTION What are the sensitivity, specificity, and positive predictive value of clinical breast examinations (CBE) performed in community settings?

DESIGN A national program designed to provide cancer screening to low-income women paid for examinations performed in a variety of settings. Records provided by those providers included documentation of CBE findings, as well as results of diagnostic evaluations for women with abnormal CBE or mammogram findings. Complete follow-up and ascertainment of interval cancers were not available for all women.

SETTING United States: facilities including university and community-based hospitals and clinics, health department clinics, mobile mammography units, and private-practice offices.

PATIENTS A total of 564708 adult women who presented for 752081 breast examinations. Of the examinations, 87815 were done on women who were known to have breast symptoms at the examination; 589048 were performed on asymptomatic women.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

CBE technique was not dictated or described. Concurrent mammography was provided in nearly all CBEs. Interval cancers could be determined only for patients with more than 1 screening record in the study period (~25% of the population).

MAIN OUTCOME MEASURES

Sensitivity, specificity, and positive predictive value.

MAIN RESULTS

See [Table 8-11](#).

CONCLUSIONS

LEVEL OF EVIDENCE Analysis of a large database.

STRENGTHS This study contains valuable data on current practice outside of research settings. This national undertaking to provide screening services to low-income women had the forethought to require documentation of findings in a consistent manner. It is the largest such report of a multicenter database of nonresearch clinical examinations.

LIMITATIONS The data include more than 750000 CBEs, but the number done in asymptomatic women is lower. The sensitivity and the likelihood ratios associated with the examination differ, depending on whether a woman has symptoms or not. Symptoms in the breast that bring a woman to

Table 8-11 Clinical Breast Examination Characteristics Change When Women Have Breast Symptoms

Clinical Breast Examination	Sensitivity, %	Specificity, %	Positive Predictive Value, % (95% CI)	LR+ (95% CI)	LR- (95% CI)
Asymptomatic patients	36	96	2.9 (2.6-3.1) ^a	9.5 (8.9-10)	0.66 (0.64-0.69)
Symptomatic patients	85	73	5.6 (5.3-5.9) ^a	2.5 (2.4-2.6)	0.22 (0.20-0.25)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

^aCalculated from data provided in the report.

present to a physician include pain, finding of a lump, nipple discharge, or other change in the nipple or the skin, each of which is associated with an increased risk of breast cancer.¹ Although 87% of the examinations in the series were done on asymptomatic women, 47% of the cancers detected were found in women who came to the program with symptoms.

With regard to screening CBE in the asymptomatic population, several points are worthy of note. First, the measured sensitivity of the CBE must be seen as an upper limit to the true sensitivity because there was no systematic follow-up of women who had normal examination results, and one must allow that interval cancers occurred in the group of women lost to follow-up, which are not recorded. Second, the lack of standardized procedures used in the performance of the CBE causes one to wonder for this study, as for most of the screening studies reviewed in the original article, whether the performance characteristics of the CBE would improve with trained examiners following a standard protocol.

REFERENCE FOR THE EVIDENCE

- Barton MB, Elmore JG, Fletcher SW. Breast symptoms among women enrolled in a health maintenance organization: frequency, evaluation, and outcome. *Ann Intern Med.* 1999;130(8):651-657.

Reviewed by Mary B. Barton, MD

TITLE Predictors of Sensitivity of Clinical Breast Examination.

AUTHORS Oestreicher N, White E, Lehman CD, Mandelson MT, Porter PL, Taplin SH.

CITATION *Breast Cancer Res Treat.* 2002;76(1):73-81.

QUESTION What factors influence the sensitivity of the clinical breast examination (CBE) in screening for breast cancer?

DESIGN Analysis of data linkage between a breast cancer screening program involving both CBE and mammography and a population-based cancer registry.

SETTING Breast cancer screening program of a large health maintenance organization in Washington State.

PATIENTS Women who had undergone screening and who were diagnosed with a first breast cancer within 12 months of the screening examination were potentially eligible to be included ($n = 474$). Four of these women were excluded because of the presence of breast implants and 1 each because of symptoms at the screening visit and at the request of the patient.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

CBE technique is not described, but the training of examiners is described and the authors state that examiners were CBE-certified by the American Cancer Society. Concurrent mammography was provided in all CBEs, but in most cases, the results of the mammogram were not available to the examiner. Breast cancer diagnoses were determined from the Surveillance, Epidemiology and End-Results cancer registry of Seattle-Puget Sound.

MAIN OUTCOME MEASURE

Sensitivity of CBE.

MAIN RESULTS

The sensitivity of the breast examination was 0.35 (95% CI 0.31-0.39). The authors found in multivariable analyses that CBE sensitivity was significantly higher for women with larger tumors at diagnosis, for Asian women compared with white women, and for women with normal body mass index or weight compared with women with increased body mass index. The sensitivity of CBE was lower in women at extremes of age (ie, 40-49 years or ≥ 80 years) compared with that in women aged 50 to 59 years.

CONCLUSIONS

LEVEL OF EVIDENCE Level 4 (“sensitivity-only” study).

STRENGTHS This study used a comprehensive breast cancer screening program in a stable managed-care population and linked these data to a population-based cancer registry to ascertain cancer outcomes of women screened.

LIMITATIONS Although the data were somewhat old (all cancers diagnosed 8 or more years before the date of publication), the technique of CBE had not changed during that time. Because of the nature of the analysis, the study could confidently assess sensitivity of CBE only among women in whom cancers were diagnosed and could not assess the specificity or the positive predictive value of CBE.

The authors observed that their study is one of effectiveness, not efficacy, in comparing their findings with those of the Canadian NBBS studies. Although this may be true, one imagines that, even in an actual clinical setting, the use of standardized best practice procedures in the performance of the CBE could not hurt, and might help the performance characteristics of the CBE.

Reviewed by Mary B. Barton, MD

CHAPTER 9

Does This Patient Have a Clinically Important Carotid Bruit?

Jean-Stéphane Sauvé, MD

Andreas Laupacis, MD

Truls Østbye, MD

Brian Feagan, MD

David L. Sackett, MD

CLINICAL SCENARIOS

CASE 1 A 50-year-old man undergoes a general physical examination for his insurance policy. A left-sided, focal, systolic carotid bruit is identified. There is no history of stroke or transient ischemic attack (TIA).

CASE 2 A 50-year-old man undergoes a preoperative examination the evening before he is to undergo coronary artery bypass surgery. A bruit identical to that found in the first patient is heard. There is no history of cerebrovascular symptoms.

CASE 3 A 50-year-old man presents to the emergency department with a history of a transient (less than 1 hour) slurring of speech and right-arm weakness. There is no history of cerebrovascular disease, and the physical examination reveals a focal, left-sided, systolic carotid bruit.

THE IMPORTANCE OF CLINICAL EXAMINATION

The clinical significance of the identical-sounding bruits is vastly different in these patients. In each of them, the coupling of a thoughtful history with a competent physical examination will lead to different prognostic predictions and differing courses of appropriate clinical action.

THE CAROTID ARTERY AS A CAUSE FOR BRUITS IN THE NECK

The right common carotid artery arises from the brachiocephalic artery (the first branch of the aortic arch), and the left arises directly from the aortic arch. The common carotid arteries run upward and backward through the neck, from the sternoclavicular joint to the upper border of the thyroid cartilage, where they divide into the external and internal carotid arteries (Figure 9-1). The external carotid artery terminates in the substance of the parotid gland, where it divides into the superficial temporal and mandibular arteries. The internal carotid artery ascends to the base of the skull and enters the cranium through the carotid canal in the temporal bone.

Although bruits of the carotid artery have been reported in approximately 20% of children younger than 15 years, they occur in about 1% of healthy adults.¹ Carotid bruits can be heard in states of increased vascular flow such as thyrotoxicosis, anemia, and arteriovenous fistulas. A relatively common example of the latter occurs with the creation of a forearm fistula in patients receiving hemodialysis.² In a convenience sample of 15 long-term hemodialysis patients, Messert et al² found bilateral carotid bruits in 5 patients and a unilateral bruit in 6 patients. The bruit was usually louder on the side

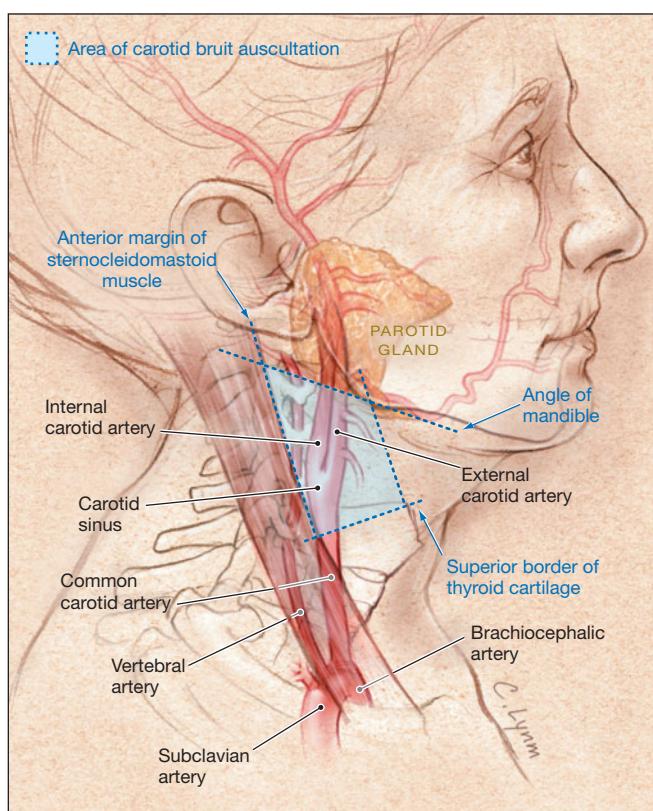


Figure 9-1 Anatomy of the Right Carotid Artery

Carotid bruits are heard best in the polygonal area (shaded in blue). This area is bounded superiorly by the angle of the jaw, inferiorly by the upper border of the thyroid cartilage, and posteriorly by the sternocleidomastoid muscle.

of the fistula and was often associated with a subclavian bruit (in 13 of 15 patients). Carotid artery stenosis, typically caused by atherosclerosis, is the underlying condition to be considered when one hears a carotid bruit, and the accuracy of this sign is discussed below. However, a bruit may be heard over the bifurcation of the carotid artery when the associated angiogram shows either a normal or a completely occluded internal carotid artery; in these cases, the bruit may arise from a stenosed external carotid artery.³

HOW TO HEAR CAROTID BRUITS

In a quiet room, with the patient relaxed, it is conventional to use the bell of the stethoscope and to listen for carotid bruits over an area beginning from just behind the upper end of the thyroid cartilage to just below the angle of the jaw (Figure 9-1).^{4,5} No method of auscultation has been demonstrated to be superior to another. Most carotid bruits are heard only in systole, but some are heard in both systole and diastole, the significance of which is unclear, given the poor clinical agreement on the assessment of the duration of carotid bruits.⁶

Carotid bruits make up but a portion of all neck bruits. Systolic heart murmurs transmitted to the neck usually can be differentiated from carotid bruits because they are louder over the precordium than over the neck.

Venous hums, caused by flow in the internal jugular vein, have been reported to occur in approximately 25% of young adults.⁷ They are easily distinguishable from carotid bruits, being most prominent in diastole, with the patient sitting and the head turned away from the side of auscultation. Venous hums are rarely heard with the patient lying down and are always abolished either by the compression of the ipsilateral internal jugular vein cephalad to the stethoscope or by Valsalva maneuver.^{8,9}

PRECISION OF AUSCULTATION FOR CAROTID BRUITS

Among 55 patients examined independently by 2 neurologists (both of whom had normal audiogram results), the agreement beyond chance for the presence of a bruit was substantial, with a κ of 0.67. However, agreement regarding the intensity, pitch, or duration of the bruit was only fair ($\kappa < 0.40$).⁶

THE IMPORTANCE OF CAROTID BRUITS IN DIFFERENT CLINICAL PRESENTATIONS

Case 1: The Asymptomatic Ambulatory Bruit

How Often Should We Expect to Find an Asymptomatic Carotid Bruit?

In a community-based study, Heyman et al¹⁰ found the prevalence of asymptomatic cervical bruits (bruits heard in the supraclavicular area or anterior to the sternocleidomastoid muscle) to increase with age, from 2.3% in the age group of 45 to 54 years to a high of 8.2% in the age group of 75 years or older. Bruits were more common in women and hypertensive patients.

If No Bruit Is Found at This Examination, What Are the Chances of Developing a Bruit De Novo During the Following Years?

The incidence of de novo bruits also increases with age. Wolf et al¹¹ estimated that of a cohort of 100 adults aged 65 years or older, approximately 1% per year (7 during the next 8 years) will develop a new carotid bruit, a rate twice that of individuals aged 45 to 54 years.

What Are the Prognostic Implications of Discovering an Asymptomatic Carotid Bruit During a General Physical Examination in a 50-Year-Old Man?

Asymptomatic carotid bruits are associated with increased incidence of both cerebrovascular and cardiac events in this age group. For example, Wiebers et al¹² conducted a 5-year prospective, population-based study of 2 unmatched but generally similar cohorts, one of which had carotid bruits (566 individuals) and one of which did not (428 individuals). The average annual stroke rates were 3 times as high in patients with bruits (1.5%) compared with those without (0.5%), and similar ratios were also found for TIAs (0.9% vs 0.2%). Most strokes and TIAs occurred on the same side as the bruit. The prognosis was not different for the various types of carotid bruits (diffuse vs localized, isolated systolic vs systolic and diastolic). In a second prospective, popula-

tion-based cohort, Heyman et al¹⁰ followed up 1620 asymptomatic adults aged 45 years or older for 6 years and again found a higher incidence of strokes in patients with cervical bruits (odds ratio [OR], 4.2). The association appeared stronger in men (OR, 7.5) than in women (OR, 1.6). Heyman et al¹⁰ also found a 3.4-fold higher risk of death from ischemic heart disease in men with asymptomatic cervical bruits (90% confidence interval [CI], 1.1–11), and a 1.9-fold higher risk in women (90% CI, 0.7–5).

A randomized trial of carotid endarterectomy in asymptomatic carotid stenoses of at least 50% reported a decrease in TIAs after surgery.¹³ However, there was no decrease in disabling or fatal stroke after surgery, and most clinicians would not refer such patients for angiography.

In the elderly (older than 75 years), there may not be an increased risk of stroke with asymptomatic carotid bruits. Among nursing home residents, the 3-year incidence of TIA or stroke was 10% when a bruit was present and 9% when it was absent, a relative risk of only 1.1 (95% CI, 0.45–2.7).¹⁴

Case 2: The Asymptomatic Preoperative Bruit

How Often Should We Expect to Find an Asymptomatic Carotid Bruit on Routine Preoperative Assessment?

The prevalences reported in the 4 surgical cohort studies that assessed for the presence of bruits preoperatively range from a low of 6% (Ivey et al¹⁵) to a high of 16% (Evans and Cooperman¹⁶), with an overall average of approximately 10%. These figures are significantly higher than those in the general population (average, 4.4%), probably because 3 of the 4 surgical series were patients undergoing major vascular procedures, in which the prevalence of systemic atherosclerosis is increased.

Are Patients With Asymptomatic Preoperative Bruits at Higher Risk of Perioperative Stroke?

As shown in Table 9-1, only Barnes et al¹⁷ of the 4 studies^{15–18} found an increased incidence of permanent neurologic complications after surgery among patients with preoperative asymptomatic carotid bruits. When combined with the other 3 studies, the difference becomes a nonsignificant trend favoring fewer strokes among patients with carotid bruits (pooled rate difference,¹⁹ –0.6% [95% CI, –1.6% to 0.4%]; pooled OR,²⁰ 0.94 [95% CI, 0.22–3.9]).

On the other hand, Ivey et al¹⁵ found an increase (11% vs 2%; $P < .001$) in transient, nonfocal neurologic abnormalities (such as intellectual and behavioral changes) in patients with asymptomatic bruits who underwent cardiac procedures.

Case 3: The Symptomatic Bruit

Should Further Diagnostic or Therapeutic Procedures Be Carried Out in Patients With Symptomatic Carotid Bruits?

Two randomized controlled trials^{21,22} demonstrated that carotid endarterectomies markedly decrease mortality and stroke in patients with symptomatic, high-grade (70%–99%) carotid stenosis. Accordingly, the onus is on the physician to rule in or rule out high-grade carotid stenosis in all patients with anterior-circulation TIAs or minor strokes, regardless of bruits.

Does the Presence or Absence of a Carotid Bruit Accurately Reflect the Degree of Underlying Carotid Artery Stenosis in Symptomatic Patients?

The relationship between carotid bruits in patients with cerebrovascular symptoms and angiographically determined carotid stenoses is summarized in Table 9-2.^{23–26} The 2 studies that reported data specifically about high-grade stenoses found an association with carotid bruits.^{25,26} The likelihood ratios for high-grade carotid stenoses were 3.2 and 1.6 when bruits were present and 0.3 and 0.6 when bruits were absent.

Unfortunately, however, this relationship is not strong enough for the clinician to be able to use the presence of a bruit to rule in, or the absence of a bruit to rule out, high-grade carotid stenosis. For example, in the North American Symptomatic Carotid Endarterectomy Trial (NASCET),²¹ more than one-third of patients with high-grade stenoses had no detectable bruits, and the presence of a focal carotid bruit increased the probability of underlying high-grade (70%–99%) carotid stenosis by only 11%, from a preexamination probability of 52% to a postexamination probability of 63%. Furthermore, the NASCET also showed that no other bruits (supraclavicular, ophthalmic, or contralateral) added to the accuracy of the finding.

THE BOTTOM LINE

1. Asymptomatic carotid bruits are relatively common. Their prevalence increases with age. They are associated

Table 9-1 Risk of Perioperative Stroke in Patients With Preoperative Carotid Bruits

Studies	Types of Patients	No. of Patients With Perioperative Stroke/ Total No. of Patients With Bruits (%)	No. of Patients With Perioperative Stroke/ Total No. of Patients Without Bruits (%)	P^a
Barnes et al ¹⁷	Coronary artery bypass graft and vascular surgery	2/44 (4.5)	3/405 (0.7)	.02
Evans and Cooperman ¹⁶	Major vascular surgery	0/92 (0)	4/496 (0.8)	.39
Ivey et al ¹⁵	Cardiac surgeries	0/82 (0)	9/1339 (0.7)	.46
Ropper et al ¹⁸	All elective surgeries for those >55 y	0/82 (0)	4/592 (0.7)	.46

^aUsing the χ^2 test.

Table 9-2 Ability of Carotid Bruits to Indicate Various Degrees of Angiographic Carotid Stenosis in Patients With Symptoms

Studies	Types of Patients	Degree of Stenosis Predicted, %	Sensitivity	Specificity	LR+	Pretest P	Posttest P
Ingall et al ²³	Various symptoms	50-99	0.37	0.94	5.7	.25	.65
Ziegler et al ²⁴	TIA	>50	0.29	0.88	2.4	.08	.17
Hankey and Warlow ²⁵	Anterior-circulation TIA	75-99	0.76	0.76	3.2	.16	.37
North American Symptomatic Carotid Endarterectomy Trial Collaborators ²⁶	Anterior-circulation TIA	70-99	0.62	0.61	1.6	.52	.63

Abbreviations: LR+, positive likelihood ratio; TIA, transient ischemic attack.

with a long-term increase in cerebrovascular and cardiac events, except perhaps in individuals older than 75 years.

- Asymptomatic preoperative bruits are not predictive of increased risk of perioperative stroke. However, they may be harbingers of transient postoperative cognitive and behavioral abnormalities.
- Although the presence of a carotid bruit in a patient with carotid-territory cerebrovascular symptoms increases the probability that the underlying stenosis is high grade (and therefore amenable to endarterectomy), the accuracy of this physical finding is low. Accordingly, the presence of a carotid bruit cannot be used to rule in, nor can its absence be used to rule out, surgically amenable carotid artery stenosis in symptomatic patients.

Author Affiliations at the Time of the Original Publication

From the Departments of Medicine and Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario, Canada (Drs Sauvé and Sackett); the Departments of Medicine and Epidemiology and Community Medicine, University of Ottawa, Ottawa, Ontario, Canada (Dr Laupacis); the Departments of Epidemiology and Biostatistics and Family Medicine, University of Western Ontario, London, England (Dr Østbye); and the Department of Medicine, University of Western Ontario, London, England (Dr Feagan).

Acknowledgments

We thank Jeff Turnbull, MD, for his early contribution in developing this article.

REFERENCES

- Wadia NH, Monckton G. Intracranial bruits in health and disease. *Brain*. 1957;80(4):492-509.
- Messert B, Marra TR, Zerofsky RA. SuprACLAVICULAR AND CAROTID BRUITS IN HEMODIALYSIS PATIENTS. *Ann Neurol*. 1977;2(6):535-536.
- Matthews WB. Observations on the carotid bruit. *J Neurol Neurosurg Psychiatry*. 1961;24:161-166.
- Ratcheson RA. Clinical diagnosis of atherosclerotic carotid artery disease. *Clin Neurosurg*. 1982;29:464-481.
- Harrison MJ. Cervical bruits and asymptomatic carotid stenosis. *Br J Hosp Med*. 1984;32(2):80-81.
- Chambers BR, Norris JW. Clinical significance of asymptomatic neck bruits. *Neurology*. 1985;35(5):742-745.
- Jones FL. Frequency, characteristics and importance of the cervical venous hum in adults. *N Engl J Med*. 1962;267:658-660.
- Hurst JW, Hopkins LC, Smith RB. Noises in the neck. *N Engl J Med*. 1980;302(15):862-863.
- Fowler NO, Gause R. The cervical venous hum. *Am Heart J*. 1964; 67:135-136.
- Heyman A, Wilkinson WE, Heyden S, et al. Risk of stroke in asymptomatic persons with cervical arterial bruits: a population study in Evans County, Georgia. *N Engl J Med*. 1980;302(15):838-841.
- Wolf PA, Kannel WB, Sorlie P, McNamara P. Asymptomatic carotid bruit and risk of stroke: the Framingham Study. *JAMA*. 1981;245(14):1442-1445.
- Wiebers DO, Whisnant JP, Sandok BA, O'Fallon WN. Prospective comparison of a cohort with asymptomatic carotid bruit and a population-based cohort without carotid bruit. *Stroke*. 1990;21(7):984-988.
- Hobson RW, Weiss DC, Fields WS, et al. Efficacy of carotid endarterectomy for asymptomatic carotid stenosis. *N Engl J Med*. 1993;328(4):221-227.
- Van Ruiswyk J, Noble H, Sigmann P. The natural history of carotid bruits in elderly persons. *Ann Intern Med*. 1990;112(5):340-343.
- Ivey TD, Strandness E, Williams DB, Langlois Y, Misbach GA, Kruse AP. Management of patients with carotid bruit undergoing cardiopulmonary bypass. *J Thorac Cardiovasc Surg*. 1984;87(2):183-189.
- Evans WE, Cooperman M. The significance of asymptomatic unilateral carotid bruits in preoperative patients. *Surgery*. 1978;83(5):521-522.
- Barnes RW, Liebman PR, Marszalek PB, Kirk CL, Goldman MH. The natural history of asymptomatic carotid disease in patients undergoing cardiovascular surgery. *Surgery*. 1981;90(6):1075-1081.
- Ropper AH, Wechsler LR, Wilson LS. Carotid bruit and risk of stroke in elective surgery. *N Engl J Med*. 1982;307(22):1388-1390.
- DerSimonian R, Laird N. Meta-analysis in clinical trials. *Control Clin Trials*. 1986;7(3):177-188.
- Yusuf S, Peto R, Lewis J, Collins R, Sleight P. Beta blockade during and after myocardial infarction: an overview of the randomized trials. *Prog Cardiovasc Dis*. 1985;27(5):335-371.
- North American Symptomatic Carotid Endarterectomy Trial Collaborators. Beneficial effect of carotid endarterectomy in symptomatic patients with high-grade carotid stenosis. *N Engl J Med*. 1991;325(7):445-453.
- European Carotid Surgery Trialists' Collaborative Group. NRC European Carotid Surgery Trial: interim results for symptomatic patients with severe (70-99%) or mild (0-29%) carotid stenosis. *Lancet*. 1991; 337(8752):1235-1243.
- Ingall TJ, Homer D, Whisnant JP, Baker HL, O'Fallon WN. Predictive value of carotid bruit for carotid atherosclerosis. *Arch Neurol*. 1989; 46(4):418-422.
- Ziegler DR, Zileli T, Dick A, Sebaugh JL. Correlation of bruits over the carotid artery with angiographically demonstrated lesions. *Neurology*. 1971;21(8):860-865.
- Hankey GJ, Warlow CP. Symptomatic carotid ischaemic events: safest and most cost effective way of selecting patients for angiography, before carotid endarterectomy. *BMJ*. 1990;300(6738):1485-1491.
- Sauvé S, Sackett DL, Taylor DW, Barnett HJM, Haynes RB, Fox A; for the NASCET. Can bruits distinguish high-grade from moderate symptomatic carotid stenosis [abstract]? *Clin Res*. 1992;40:304A.

Prepared by David L. Simel, MD, MHS

Reviewed by Larry Goldstein, MD

CLINICAL SCENARIO

A 65-year-old woman returns to your office to review her home blood pressure recordings. She frequently has a systolic pressure of approximately 130 to 145 mm Hg. While examining her, you slip the stethoscope onto her neck and hear a focal, unilateral bruit. The patient notices a change in your facial expression while you are listening to her neck, so she asks, "Did you hear something?" You tell her you heard a "squeaky noise" and then immediately wonder whether she (or you) needed that information. You realize you need to know whether the presence of a bruit suggests that the patient might have a carotid stenosis severe enough to warrant a surgical evaluation. She reminds you that her father, after being healthy all his life, had a stroke when he was 72 years old.

UPDATED SUMMARY ON THE CLINICAL EXAMINATION FOR CAROTID BRUITS

Original Review

Sauve JS, Laupacis A, Ostbye T, Feagan B, Sackett DL. Does this patient have a clinically important carotid bruit? *JAMA*. 1993;270(23):2843-2845.

UPDATED LITERATURE SEARCH

Much has been written on carotid disease, particularly the role of carotid endarterectomy for individuals with symptoms of a stroke, transient ischemic attack (TIA), or transient monocular blindness. However, there are also new data for individuals who are asymptomatic. We focused our updated literature review on the role of the carotid bruit in detecting patients who have an ipsilateral carotid stenotic lesion because these patients might benefit from carotid endarterectomy.

Our literature search included the years 1992 through July 2004 and combined the text words "bruit and carotid" with "asymptomatic and carotid" to yield 85 English-language articles. We excluded case reports and then reviewed

the abstracts of 76 articles to identify 24 promising articles. When possible, electronic copies of the articles were obtained and searched for the text word "bruit." Articles were retained when they were prospective studies of adults that included both sensitivity and specificity data of level 3 quality or greater. We also retained articles that were studies of the positive predictive value of carotid bruits at a threshold of at least 70% stenosis because these are the patients who will likely benefit from endarterectomy. The reference lists for each article were reviewed, yielding 1 additional study. The reference list of the original Rational Clinical Examination article was reviewed, and previously cited literature was obtained to assess whether the data could be reanalyzed. Eight articles were ultimately included in this update.

NEW FINDINGS

Symptomatic Patients

The presence of a carotid bruit increases the likelihood of a 70% to 99% carotid stenosis (likelihood ratio [LR], 3). However, newer studies confirm that the absence of a bruit is not sufficient to prove that the carotids are normal.

Asymptomatic Patients

Newer studies allow us to deduce that the presence of a bruit in asymptomatic patients appreciably increases the likelihood of carotid stenosis (LR, 4-10).

IMPROVEMENTS IN THE DATA PRESENTED IN THE ORIGINAL PUBLICATION

Confidence intervals were added by retrieving original references and extracting the results. We also reviewed cited studies to assess whether they had information about the predictive value for bruits.

CHANGES IN THE REFERENCE STANDARD

A meta-analysis of noninvasive carotid artery tests showed that carotid duplex, carotid Doppler, and magnetic reso-

nance angiography had excellent sensitivity (89%-94%) and specificity (85%-92%) for detecting carotid stenosis of 70% occlusion or greater, compared with carotid angiography.¹ This makes them useful as the next screening test after the clinical examination, when more information is required.

RESULTS OF LITERATURE REVIEW

Symptomatic Patients

Studies now address the role of carotid bruits in predicting carotid stenosis for a broader array of both symptomatic and asymptomatic patients (Table 9-3). For symptomatic patients, the studies were done in enough detail to allow us to estimate the sensitivity and specificity of the carotid bruit for predicting a surgical stenosis. These studies demonstrate much better specificity than sensitivity and that the presence of a carotid bruit increases the likelihood of a stenotic lesion in symptomatic patients. However, the newer information confirms that the absence of a bruit in symptomatic patients is not adequate to prove that the carotids are normal.

Table 9-3 Results for Predicting a Significant Carotid Stenosis of 70% to 99%

Study	Stenosis, %	LR+, (95% CI)	LR-, (95% CI)
Mead et al ⁶ (symptomatic, referred for neurology evaluation)	70-99	5.5 (4.1-7.2)	0.48 (0.38-0.60)
Hankey and Warlow ² (symptomatic, ^a referred for neurology evaluation for endarterectomy)	75-99	3.2 (2.4-4.2)	0.31 (0.18-0.50)
Sauve et al ³ (symptomatic, enrolled in endarterectomy trial)	70-99	1.6 (1.4-1.8)	0.61 (0.54-0.68)
Magyar et al ⁷ (57% had symptoms; referred to neurology clinic)	70-99	6.0 (3.2-10)	0.48 (0.25-0.74)
Hill et al ⁸ (asymptomatic before cardiac surgery)	>80	8.6 (4.3-15)	0.24 (0.07-0.60)
de Virgilio et al ⁹ (asymptomatic referred for peripheral vascular disease evaluation)	50-99	4.2 (2.3-7.2)	0.55 (0.34-0.77)
Summary LRs			
Symptomatic (n = 3 studies, ^b 2292 patients)		3.0 (1.3-7.1)	0.49 (0.36-0.67)
Asymptomatic (n = 2 studies, ^c 275 patients)		6.0 (2.6-14)	0.45 (0.22-0.92)
Asymptomatic deduced from the positive predictive values (Table 9-4)	4.0-10	Uncertain	

Abbreviations: CI, confidence interval; LR, likelihood ratio; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

^aAll patients were symptomatic. After the arteriogram, 298 carotid arteries were considered "symptomatic" and 124 arteries were considered "asymptomatic." Retrospectively, bruits were heard in 95 of 298 symptomatic arteries and 41 of 124 asymptomatic arteries. Data are not provided to allow calculation of separate sensitivity and specificity.

^bMead et al,⁶ Hankey and Warlow,² and Sauve et al.³

^cHill et al⁸ and de Virgilio et al.⁹

An interesting finding is revealed by looking at 2 of the studies cited in the original Rational Clinical Examination article. These studies^{2,3} included the most selective population of patients for whom the sensitivity and specificity were reported. The study by Hankey and Warlow² included only those symptomatic patients who were suitable candidates for endarterectomy. The North American Symptomatic Carotid Endarterectomy Trial⁴ included only patients with carotid stenosis, and then only those who were randomized to endarterectomy versus medical therapy. These studies exhibit verification bias, which typically creates underestimates of the specificity and value of hearing a carotid bruit. Thus, it should not be surprising that these studies also have the lowest positive likelihood ratio (LR+) of those we reviewed (Table 9-3). On the other hand, this reappraisal confirms that the absence of a bruit in symptomatic patients does not have enough diagnostic power in the symptomatic patient to rule out an important stenotic lesion.

Asymptomatic Patients

By asymptomatic, we mean asymptomatic for cerebrovascular disease. Two studies, one in preoperative cardiac patients and the other in peripheral vascular disease patients, allow us to calculate both the sensitivity and specificity of the carotid bruit. Although the studies used slightly different thresholds to characterize patients as having carotid stenosis, the predictive values for bruit are statistically similar among the asymptomatic studies (Table 9-4). However, we can also look at the predictive value for the presence of a bruit and determine whether it varies (Table 9-4), which is useful because the studies that allowed us to calculate sensitivity and specificity may not generalize to an age-matched general medical patient. The positive predictive value for symptomatic patients is approximately 50% and about half that (22%) for patients with no cerebrovascular symptoms. Because we know the predictive value, we can make inferences about the LR+. This follows from the equation:

$$\text{Posterior odds} = \text{Prior odds} \times \text{LR}$$

From epidemiologic studies, the prevalence should range from approximately 0.5% for patients aged 50 years or older to approximately 10% for patients aged 80 years or older.⁵ These values establish a range of reasonable prior odds. The data from the positive predictive value studies allow us to develop a range for the posterior odds. We can then solve for the LR for both symptomatic patients (50% posterior probability) and asymptomatic patients (22% posterior probability) (Figure 9-2).

The likelihood ratio for a bruit to predict significant carotid stenosis varies with the prior probability. Figure 9-2 shows that as the prior probability of stenosis increases (x-axis), the importance of a carotid bruit becomes less. If your population of asymptomatic patients is recognizably similar to those who were included in the baseline summary estimate from Table 9-4, then you would use the asymptomatic probability line and see that across a reasonable range of prior probabilities (about 3%-8% on the x-axis) for carotid

Table 9-4 Results for the Positive Predictive Value for the Presence of a Bruit in Predicting Ipsilateral Carotid Stenosis

Study	Degree of Stenosis, %	Stenosis/All Patients	Positive Predictive Value, % (95% CI)
Sauve et al ³ (symptomatic, part of endarterectomy trial)	70-99	420/667	63 (60-68)
Mead et al ⁶ (symptomatic, referred to neurologist)	70-99	54/119	45 (36-55)
Hankey and Warlow ² (symptomatic, referred to neurologist for evaluation of endarterectomy)	75-99	35/95	37 (28-47)
Hill et al ⁸ (asymptomatic before cardiac surgery)	>80	7/23	30 (16-51)
Chambers and Norris ¹⁰ (asymptomatic referred for bruit)	>75		23 (19-26)
Lewis et al ¹¹ (asymptomatic referred for bruit)	80-99		21 (18-24)
Summary Predictive Value			
Symptomatic (n = 3 studies, 868 patients) ^a	70-99		50 (35-64)
Asymptomatic (n = 3 studies, 1303 patients) ^b	>75		22 (20-24)

Abbreviation: CI, confidence interval.

^aStudies combined from Mead et al,⁶ Sauve et al,³ and Hankey and Warlow.²

^bStudies combined from Hill et al,⁸ Chambers and Norris,¹⁰ and Lewis et al.¹¹ The results are homogenous ($P = .42$).

stenosis, finding a bruit has a useful LR of 4 to 10 (from the y-axis). Fortunately, we can feel more confident about this because the results are similar to the summary LRs for asymptomatic patients from Table 9-3.

EVIDENCE FROM GUIDELINES

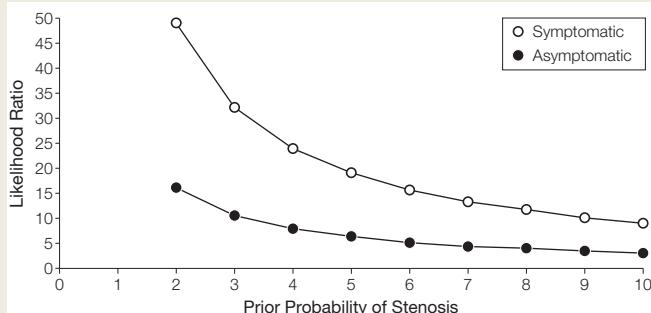
Symptomatic patients with TIAs who are surgical candidates should be evaluated for carotid stenosis, whether or not they have a bruit.¹²

The US Preventive Services Task Force reviewed screening for asymptomatic carotid artery stenosis in 1996 and found insufficient evidence to make a recommendation about listening for carotid bruits.⁵ The Task Force observed that the annual incidence of stroke unheralded by any TIA symptoms ipsilateral to a bruit is 1% to 3%. The interpretation of data presented in this update were not available to the Task Force and have not been incorporated into the 1996 recommendations. There are still no data that assess the effect of screening for an asymptomatic bruit, confirming stenosis, and then

performing an endarterectomy on patients with surgically significant lesions. The Canadian Task Force recommended that clinicians not listen for carotid bruits in asymptomatic patients.¹³ There does seem to be consensus that the presence of an asymptomatic bruit is a marker of atherosclerotic risk.

CLINICAL SCENARIO—RESOLUTION

You listened for a bruit with the plan that you would emphasize risk-reduction strategies for your hypertensive patient, but now, she has asked you to use the findings to help decide whether to assess her for carotid stenosis. A variety of studies suggest that the LR+ for carotid stenosis when a bruit is heard is 4 to 10. Let us say you estimate that her prior probability of carotid stenosis is approximately 3%, which agrees with epidemiologic data. Finding a bruit increases her probability of carotid stenosis to approximately 11%, but it might be as high as 20%. Hence, you probably have identified a patient at higher risk of carotid stenosis. The issue is not whether you *can* identify stenosis with ultrasonography, but whether you *should*. Studies of diagnostic tests give you only the likelihood of the target disorder. You will need to review the natural history of patients with asymptomatic carotid stenosis to help this patient decide whether to pursue further testing.

**Figure 9-2** Likelihood Ratio of Carotid Bruit as a Function of Symptoms and Prior Probability of Stenosis

The likelihood ratio of a carotid bruit in predicting carotid stenosis depends on whether the patient is symptomatic or asymptomatic and on the prior probability of carotid stenosis. However, for both groups of patients the positive likelihood ratio decreases in value as the prior probability of carotid stenosis increases.

REFERENCES FOR THE UPDATE

1. Blakeley DD, Oddone EZ, Hasselblad V, Simel D, Matchar DB. Noninvasive carotid artery testing: a meta-analytic review. *Ann Intern Med*. 1995;122(5):360-367.
2. Hankey GJ, Warlow CP. Symptomatic carotid ischaemic events: safest and most cost-effective way of selecting patients for angiography, before carotid endarterectomy. *BMJ*. 1990;300(6738):1485-1491.⁴
3. Sauve JS, Thorpe KE, Sackett DL, et al. Can bruits distinguish high-grade from moderate symptomatic carotid stenosis? *Ann Intern Med*. 1994;120(8):633-637.³

CAROTID STENOSIS—MAKE THE DIAGNOSIS

It is hard for physicians to resist auscultating the neck. Perhaps no physical finding in adults causes as much confusion as the presence of the carotid bruit in asymptomatic patients. Most clinical research suggests that there is a clear benefit to carotid endarterectomy for patients with symptoms and a benefit (although likely small) for asymptomatic patients.

PRIOR PROBABILITY FOR CAROTID STENOSIS

Symptomatic Patients

Prior Probability

After ruling out patients for whom endarterectomy would not be considered, 10% to 30% will have surgically amenable carotid stenosis. There is variability in the estimates of the remaining patients who will prove to have surgically correctable carotid stenosis. The variability depends on the patient population, criteria for determining surgical risk, and the threshold for defining an “important” stenosis.

Asymptomatic Patients

Prior Probability

For patients 60 years or older, there is 1% to 10% probability for carotid stenosis.

The prevalence of carotid stenosis increases from approximately 0.5% for patients 50 years of age to approximately 10% by age 90 years.¹⁴ For patients older than 65 years, 5% to 7% of women and 7% to 10% of men will have a carotid stenosis of 50% or higher. For more significant degrees of stenosis, 2 prospective, population-based samples show that 1% to 2.3% of women and 1% to 4.1% of men older than 60 years will have a stenosis of 75% to 99%.^{15,16}

POPULATION FOR WHOM THE CAROTID BRUIT MIGHT BE AUSCULTATED

- Patients with cerebrovascular symptoms compatible with a nondeilitating stroke or TIA
- Older patients, as part of an assessment for cardiovascular risk

DETECTING THE LIKELIHOOD OF CAROTID STENOSIS

The presence of a carotid bruit does increase the likelihood of an important stenotic lesion, but the absence of a bruit (especially in patients with atherosclerotic risk factors) does not rule out carotid stenosis (see Tables 9-5 and 9-6).

Table 9-5 Do Carotid Bruits Predict Stenosis in Symptomatic Patients?

	LR for Carotid Stenosis, 70%-99% (95% CI)
Ipsilateral bruit	3.0 (1.3-7.1)
No ipsilateral bruit	0.49 (0.36-0.67)

Abbreviations: CI, confidence interval; LR, likelihood ratio.

Table 9-6 Do Carotid Bruits Increase the Likelihood of Carotid Stenosis in Asymptomatic Patients?

	LR for Carotid Stenosis, 70%-99%
Ipsilateral bruit	4.0-10
No ipsilateral bruit	Uncertain

Abbreviation: LR, likelihood ratio.

REFERENCE STANDARD TESTS

- Carotid duplex ultrasonography
- Carotid Doppler ultrasonography
- Magnetic resonance angiogram

- North American Symptomatic Carotid Endarterectomy Trial (NASCET) Steering Committee. North American Symptomatic Carotid Endarterectomy Trial. *Stroke*. 1991;22(6):711-720.
- US Preventive Services Task Force. *Screening for Asymptomatic Carotid Artery Stenosis: Guide to Clinical Preventive Services*. 2nd ed. Baltimore, MD: Lippincott Williams & Wilkins; 1996:53-61.
- Mead GE, Wardlaw JM, Lewis SC, McDowall M, Dennis MS. Can simple clinical features be used to identify patients with severe carotid stenosis on Doppler ultrasound? *J Neurol Neurosurg Psychiatry*. 1999;66(1):16-19.^a
- Magyar MT, Nam E, Csiba L, Ritter MA, Ringelstein EB, Droste DW. Carotid artery auscultation—anachronism or useful screening procedure? *Neurol Res*. 2002;24(7):705-708.^a
- Hill AB, Obrand D, Steinmetz OK. The utility of selecting screening for carotid stenosis in cardiac surgery patients. *J Cardiovasc Surg*. 1999;40(6):829-836.^a
- de Virgilio C, Toosie K, Arnett T, et al. Asymptomatic carotid artery stenosis screening in patients with lower extremity atherosclerosis: a prospective study. *Ann Vasc Surg*. 1997;11(4):374-377.^a
- Chambers BR, Norris JW. Outcome in patients with asymptomatic neck bruits. *N Engl J Med*. 1986;315(14):860-865.^a
- Lewis R, Abrahamowicz M, Cote R, Battista R. Predictive power of duplex ultrasonography in asymptomatic carotid disease. *Ann Intern Med*. 1997;127(1):13-20.^a
- Albers GW, Hart RG, Lutsep HL, Newell DW, Sacco RL. AHA scientific statement: supplement to the guidelines for the management of transient ischemic attacks: a statement from the Ad Hoc Committee on Guidelines for the Management of Transient Ischemic Attacks, Stroke Council, American Heart Association. *Stroke*. 1999;30(11):2502-2511.
- Mackey A, Cote R, Battista RN. Asymptomatic carotid disease. In: Canadian Task Force on the Periodic Health Examination, eds. *Canadian Guide to Clinical Preventive Health Care*. Ottawa, Ontario, Canada: Health Canada; 1994:692-704.
- Goldstein LB, Adams R, Becker K, et al. Primary prevention of ischemic stroke. *Circulation*. 2001;103(1):163-182.
- Willeit J, Kiechl S. Prevalence and risk factors of asymptomatic extracranial carotid artery atherosclerosis. *Arterioscler Thromb*. 1993;13(5):661-668.
- Leary DHO, Polak JF, Kronmal RA, et al. Distribution and correlates of sonographically detected carotid artery disease in the cardiovascular health study. *Stroke*. 1992;23(12):1752-1760.

^aFor the Evidence to Support the Update for this topic, see <http://www.JAMAevidence.com>.

EVIDENCE TO SUPPORT THE UPDATE:

Carotid Bruit

TITLE Outcome in Patients With Asymptomatic Neck Bruits.

AUTHORS Chambers BR, Norris JW.

CITATION *N Engl J Med.* 1986;315(14):860-865.

QUESTION Does a bruit predict the presence or absence of carotid stenosis?

DESIGN Baseline data collected as part of a prospective cohort of patients enrolled in a study of asymptomatic neck bruits.

SETTING Single site, stroke unit in Toronto.

PATIENTS Among 659 patients referred for Doppler ultrasonography, 500 were asymptomatic and were enrolled in a prospective cohort.

The patients include those in whom physicians might consider the presence of carotid stenosis. They had a mean age of 64 years, 74% were men, 58% had hypertension, 58% had heart disease, 57% had peripheral vascular disease, 13% had diabetes, 73% had smoking history or currently smoked, and 35% had hypercholesterolemia.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

Patients were examined at enrollment. Carotid Doppler ultrasonography was performed without knowledge of the auscultatory findings. The ultrasonographers had demonstrated proficiency when their findings were compared with angiography.

MAIN OUTCOME MEASURE

Positive predictive value at different degrees of stenosis.

MAIN RESULTS

See Table 9-7.

Table 9-7 The Predictive Value of a Carotid Bruit for Identifying Various Levels of Carotid Stenosis

Stenosis, No. (Degree of Stenosis, %)	Positive Predictive Value of a Carotid Bruit (95% CI)
113 (>75)	23 (19-26)
157 (30-74)	31 (27-36)
230 (0-29)	46 (42-50)

Abbreviation: CI, confidence interval.

CONCLUSIONS

LEVEL OF EVIDENCE Positive predictive value studies.

STRENGTHS Prospective with careful screening and confirmed proficiency of ultrasonographers.

LIMITATIONS The results generalize only to populations with the same prevalence of carotid stenosis among patients with carotid bruits. No patient who lacked a carotid bruit was included, so the sensitivity and specificity cannot be determined.

This study included a large cohort of asymptomatic patients, evaluated solely because they had a bruit. The cohort seems typical of a group of patients at risk for cerebrovascular or atherosclerotic disease. To apply these data to your own patients, you would need to know whether the study patients were similar to your patients because the predictive value is affected by the prevalence of disease.

Reviewed by David L. Simel, MD, MHS

TITLE Do Carotid Bruits Predict Disease of the Internal Carotid Arteries?

AUTHORS Davies KN, Humphrey PRD.

CITATION *Postgrad Med J.* 1994;70(824):433-435.

QUESTION Do bruits identify patients with carotid stenosis?

DESIGN Prospective, consecutive patients.

SETTING Single site, cerebrovascular clinic in the United Kingdom.

PATIENTS All patients were referred for evaluation. The underlying prevalence of cardiovascular risk factors is not described.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

The presence of a bruit was taken from the referral note but was not confirmed at study entry. The history was confirmed in regard to symptoms.

MAIN OUTCOME MEASURE

Carotid stenosis of 70% to 99%.

MAIN RESULTS

See Table 9-8.

CONCLUSIONS

LEVEL OF EVIDENCE Level 4.

STRENGTHS Pragmatic study from the perspective of a vascular laboratory that would take the information from the referral note.

LIMITATIONS The presence of a bruit was not confirmed in a standardized manner. It is not stated whether the ultrasonography was done blinded to the clinical findings.

Although interesting from the perspective of clinicians in a vascular laboratory, the presence or absence of a bruit was not systematically confirmed by the study clinicians. The data were

Table 9-8 Likelihood Ratio of a Carotid Bruit for Carotid Stenosis of at Least 70%

Test	Sensitivity	Specificity	LR+ (95% CI)	LR- (95% CI)
Bruit	0.57	0.70	1.9 (1.4-2.6)	0.61 (0.41-0.83)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

taken from the referral requests, which may not have been consistently thorough. Thus, it is likely that some patients recorded as not having a bruit may have actually had a cervical bruit and vice versa.

Reviewed by David L. Simel, MD, MHS

TITLE Asymptomatic Carotid Artery Stenosis Screening in Patients With Lower Extremity Atherosclerosis: A Prospective Study.

AUTHORS de Virgilio C, Toose K, Arnell T, Lewis RJ, Donayre CE, Baker JD, Melany M, White RA.

CITATION *Ann Vasc Surg.* 1997;11(4):374-377.

QUESTION Does a bruit predict ipsilateral carotid stenosis among patients with peripheral vascular disease who have no cerebrovascular symptoms?

DESIGN Prospective.

SETTING Vascular surgery clinic, West Los Angeles Veterans Affairs medical center.

PATIENTS Men (n = 89) who were referred for surgical evaluation for peripheral vascular disease. Patients were excluded if they had any symptoms of cerebrovascular disease. Ninety percent of the patients had typical claudication, 88% were smokers, 60% had hypertension, and 42% had diabetes.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

Auscultation of the carotids and a carotid duplex ultrasonography were performed on each carotid by a radiologist blinded to the clinical status of the patient.

MAIN OUTCOME MEASURE

Presence of carotid stenosis greater than 50%. Data are presented for numbers of arteries imaged (n = 178).

MAIN RESULTS

See Table 9-9. Of 89 patients, 18 had a bruit (in 14 of 18, the bruit was bilateral). Of 32 carotid arteries with bruits, 13 had a stenosis of at least 50%. This study used a threshold value different from those used by other studies on the sensitivity and specificity for a carotid bruit. However, traditionally we like to think of the screening test as having the same sensitivity and specificity independent of the prevalence. Likelihood ratios (LRs) for this study are similar to those among asymptomatic cardiac surgery patients.

Table 9-9 Likelihood Ratios for a Carotid Bruit to Predict a Carotid Stenosis of at Least 50%

Test	Sensitivity	Specificity	LR+ (95% CI)	LR- (95% CI)
Bruit	0.52	0.88	4.2 (2.3-7.2)	0.55 (0.34-0.77)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

CONCLUSIONS

LEVEL OF EVIDENCE Level 2.

STRENGTHS Prospective study.

LIMITATIONS Small sample size. The study used a lower carotid stenosis threshold (50%) than other studies for reporting the association with bruits. The number of arteries with a carotid stenosis of greater than 75% in this study was small (6.7%; 12 of 178).

This is a small but sound study. The population studied seems typical of male patients with claudication. It is not clear whether the patients were consecutive patients or just those for whom peripheral vascular surgery was considered. Nonetheless, we can derive some information about the predictive value in patients with claudication.

By reporting the data at a lower threshold for defining disease (50% as opposed to 75%), there should be proportionally more patients with disease as opposed to "normal." This would not necessarily affect the sensitivity and specificity if the importance of a bruit is independent of the prevalence of disease. In fact, traditionally Bayesian analysis predicts that the sensitivity and specificity will not change with the prevalence of disease. Despite using a different threshold for defining carotid stenosis, the LRs were almost identical to most of the studies using a 70% to 75% cut point. Unfortunately, we cannot combine these data with studies using a different cut point for assessing the predictive value.

Reviewed by David L. Simel, MD, MHS

TITLE Prospective Evaluation of Carotid Bruit as a Predictor of First Stroke in Type 2 Diabetes: The Fremantle Diabetes Study.

AUTHORS Gillett M, Davis WA, Jackson D, Bruce DG, Davis TME.

CITATION *Stroke*. 2003;34(9):2145-2151.

QUESTION Among patients with diabetes who are asymptomatic for cerebrovascular ischemia, does the presence of a carotid bruit identify those who will have stroke?

DESIGN Prospective, observational study of the natural history of diabetes. Patients had a baseline assessment and then yearly follow-up (recruitment, 1993-1996; follow-up, until 2000) or until they had a qualifying event. The mean follow-up was 6.5 ± 2.2 years.

SETTING Community based in Fremantle, Western Australia.

PATIENTS Patients in a defined region of Australia were recruited from the community to participate in the Fremantle Diabetes Study. The current study includes 1181 patients from the registry who had no history of cerebrovascular disease at recruitment into the study. Fifty-three patients had bruits compared with 1128 patients without bruits.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

The presence of a carotid bruit was assessed at entry into the observational study. The presence of preexisting cerebrovascular disease was inferred from the lack of patient symptoms or history of an event. At annual follow-up, qualifying events were determined from patient's self-reported strokes or transient ischemic attack (TIA) symptoms, or a neurologic examination. Details of admissions for stroke or death were reviewed. It is not clear whether the assessment of a qualifying event was made with the knowledge of a baseline bruit. Deaths were reviewed without knowledge of carotid bruit status.

MAIN OUTCOME MEASURE

TIA or stroke.

MAIN RESULTS

See **Table 9-10**. Eighteen patients with bruits had strokes (18 of 53; 34%) vs 116 strokes in patients without bruits at entry (116 of 1128; 10%). Of the 18 patients with bruits and stroke, complete clinical data were available for 10 patients and revealed that 9 of 10 patients had a stroke ipsilateral to the bruit.

Table 9-10 Likelihood Ratios That a Bruit Predicts a Subsequent Stroke

Test	Outcome	LR+ (95% CI)	LR- (95% CI)
Bruit	Stroke in the first 2 y after entry into the study	6.6 (3.6-12)	0.78 (0.64-0.89)
Bruit	Stroke from entry to end of study	4.0 (2.3-6.8)	0.90 (0.82-0.95)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

The patients with bruits were older on entry into the study compared with those without bruits (mean age, 71 vs 63 years; $P < .001$), had a longer history of diabetes (5.0 vs 3.8 years; $P = .009$), had a higher blood pressure (mean systolic, 164 vs 149 mm Hg; $P < .001$) that more frequently led to blood pressure treatment (76% vs 47%; $P < .001$), and had less adiposity (waist circumference, 96 vs 100 cm; $P = .004$). At entry, there was a low frequency of aspirin therapy (26% of those without bruits vs 19% of those with bruits). Of the 4.9% of patients with atrial fibrillation, only 17% without bruits were taking warfarin, whereas none of the patients with bruits were taking warfarin ($P > .99$). During follow-up, 25 patients underwent carotid endarterectomy; all but 3 had qualifying endpoint symptoms.

On proportional hazards modeling, there was a difference in the effect of risk factors for the first 2 years of enrollment compared with the duration of the study. From baseline to year 2, the important risk factors for a stroke were the presence of a carotid bruit (hazard ratio [HR], 6.1; 95% confidence interval [CI], 3.1-12), age (HR, 1.5 for each 10-year increase), and diastolic blood pressure (HR, 1.4 for each 1-mm Hg increase). However, after 2 years, the influence of a carotid bruit at baseline lost statistical significance.

CONCLUSIONS

LEVEL OF EVIDENCE

Level 4.

STRENGTHS Community-based study of patients who are asymptomatic for cerebrovascular disease but who have a risk factor for atherosclerotic disease (diabetes). The prevalence of bruits in these asymptomatic patients with diabetes (4.5%) is approximately what we would expect in a general, community population.

LIMITATIONS The assessment of previous outcomes at baseline or during follow-up (stroke or TIA) relied on patient self-report or the follow-up examination. Thus, not all patients with events were hospitalized or examined when they had their TIA or stroke. The clinicians would have been aware that the patients had bruits (or not) when assessing outcomes.

Using a diagnostic test to establish prognosis can lead to errors when the prognosis depends on whether there were interventions. In this particular study, there may not have been large differences in interventions between the 2 groups

even though there was no standardized approach to care. Some statisticians would take the opportunity to do a propensity analysis to sort this out further and determine whether a bruit was associated with any treatments.

Given these caveats, can we use these data? The notion that the carotid bruit may lose “importance” over time does make sense but needs to be confirmed in other studies and in patients with different atherosclerotic risk factors. An alternative explanation may be that the stroke risk was higher early in the study because the patients were not at currently recommended levels of systolic blood pressure control. Obviously, these data could apply only to patients with diabetes who already have other risk factors for stroke and atherosclerotic disease. What they seem to suggest is that carotid bruits, at the least, are important “by the company they keep.”

Acknowledgment

Timothy M. E. Davis, FRACP, graciously provided the raw data for the event rates during the first 2 years after patient enrollment.

Reviewed by David L. Simel, MD, MHS

TITLE Symptomatic Carotid Ischaemic Events: Safest and Most Cost Effective Way of Selecting Patients for Angiography Before Carotid Endarterectomy.

AUTHORS Hankey GJ, Warlow CP.

CITATION BMJ. 1990;300(6738):1485-1491.

QUESTION Among patients considered for endarterectomy after a symptomatic cerebrovascular event, does a carotid bruit predict those who will have carotid stenosis?

DESIGN Consecutive patients under evaluation for a carotid endarterectomy who were referred to a neurologist.

SETTING Single site, Western General Hospital in Edinburgh, Scotland.

PATIENTS Four hundred eighty-five consecutive patients were referred for evaluation. Because a decision was made not to pursue possible endarterectomy, 189 patients were excluded, leaving 296 patients for analysis. Of the 296 patients, 32% had a bruit, and 70% were men with a mean age of 61 years. The excluded patients also had a prevalence of 32% bruits, and 60% were men with a mean age of 70 years. The investigators state that the decision to pursue possible surgery was independent of the presence of a bruit.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

Each patient was clinically evaluated by the neurologist. The reference standard was carotid arteriography.

Table 9-11 Likelihood Ratios of Bruit for Carotid Stenosis of at Least 75%

Test	Sensitivity	Specificity	LR+ (95% CI)	LR- (95% CI)
Ipsilateral bruit	0.76	0.76	3.2 (2.4-4.2)	0.31 (0.18-0.50)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

MAIN OUTCOME MEASURE

Carotid stenosis of 75% to 99%.

MAIN RESULTS

See Table 9-11.

CONCLUSIONS

LEVEL OF EVIDENCE Level 3.

STRENGTHS Carotid arteriogram was the reference standard test.

LIMITATIONS The study population includes only patients for whom surgery was considered an option. It is unclear whether the presence of a bruit affected the decision to pursue ultrasonography, but the proportion of patients with bruits was the same between included and excluded groups.

This population of patients is most similar to that reported from the North American Symptomatic Carotid Endarterectomy Trial (NASCET).^{1,2} However, the NASCET report on bruits included only patients who were randomized to endarterectomy instead of medical treatment. The study reviewed here includes patients a step before that. Thus, it is less selective because it included patients for whom surgery was being considered rather than only those for whom endarterectomy was planned. The study was affected by verification bias. However, the percentage of patients with bruits was identical to the percentage of patients without bruits. If the authors are correct that the presence of a bruit did not affect the decision to use arteriography, then the effect of verification bias is negligible. The data also allow us to calculate the predictive value of a bruit for different threshold levels for stenosis.

REFERENCES FOR THE EVIDENCE

1. Sauve JS, Thorpe KE, Sackett DL, et al. Can bruits distinguish high-grade from moderate symptomatic carotid stenosis? *Ann Intern Med*. 1994;120(8):633-637.
2. North American Symptomatic Carotid Endarterectomy Trial (NASCET) Steering Committee. North American Symptomatic Carotid Endarterectomy Trial. *Stroke*. 1991;22(6):711-720.

Reviewed by David L. Simel, MD, MHS

TITLE The Utility of Selective Screening for Carotid Stenosis in Cardiac Surgery Patients.

AUTHORS Hill AB, Obrand D, Steinmetz OK.

CITATION *J Cardiovasc Surg*. 1999;40(6):829-836.

QUESTION Among patients scheduled for cardiac surgery, does a carotid bruit identify those with carotid stenosis?

DESIGN Prospective, consecutive patients.

SETTING Single site, McGill University, Montreal.

PATIENTS Two hundred consecutive patients who were scheduled for elective cardiac surgery (196 for coronary bypass grafting). Most of the patients were asymptomatic for carotid artery disease ($n = 186$). The data are given so that the results for patients with asymptomatic carotid bruits ($n = 23$) can be extracted.

The distribution of patient characteristics suggests that they were typical of those undergoing coronary bypass grafting. Fifty percent were older than 65 years; half of all patients were smokers, 22% having diabetes mellitus, 31% having hyperlipidemia, and 20% having peripheral vascular disease.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

The patients were examined before the carotid ultrasonography. The ultrasonography was done by vascular technicians who had proved their proficiency compared with angiography.

MAIN OUTCOME MEASURES

Carotid stenosis of 80% or more by duplex ultrasonography. All patients with a positive duplex result also had arteriography.

MAIN RESULTS

See Table 9-12. In a logistic model with many clinical variables, the neurologic history (odds ratio [OR], 14; 95% confidence interval [CI], 2.9-73) and a carotid bruit (OR, 28; 95% CI, 6.6-123) were the only variables that were important.

Table 9-12 Likelihood Ratio for a Carotid Bruit to Predict Stenosis of at Least 80%

Test	Sensitivity	Specificity	LR+ (95% CI)	LR- (95% CI)
Asymptomatic carotid bruit	0.78	0.91	8.6 (4.3-15)	0.24 (0.07-0.60)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

CONCLUSIONS

LEVEL OF EVIDENCE

Level 2.

STRENGTHS Prospective consecutive enrollment of patients, primarily those asymptomatic for carotid artery disease. Although the study patients were all scheduled for cardiac surgery, the population included patients for whom carotid artery stenosis might be considered. It is one of the few studies that contain specificity data for a population of patients who are asymptomatic for cerebrovascular disease. A logistic regression was done to determine whether carotid bruits were important after controlling for other clinical variables.

LIMITATIONS

Small sample size.

Despite the small sample size compared with studies of symptomatic patients, this is an important study. The prevalence of carotid disease (defined as >80%) was 4.8% for individuals who were asymptomatic for neurologic symptoms vs 36% for those with symptoms. The positive predictive value for finding an asymptomatic bruit was 30%.

The prevalence of carotid stenosis in this study is approximately what could be expected for an age-matched population of patients with atherosclerotic disease. Although more studies with specificity data for the bruit in asymptomatic patients are needed, these results may generalize to those with atherosclerotic disease.

Reviewed by David L. Simel, MD, MHS

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

A neurologist evaluated all patients to confirm a bruit. Ultrasonography was performed, although obviously the radiologist was aware of the presence of a bruit.

MAIN OUTCOME MEASURE

The predictive value of a carotid bruit for identifying various levels of carotid stenosis.

MAIN RESULTS

See Table 9-13.

Table 9-13 Predictive Value of a Carotid Bruit for Identifying Various Levels of Carotid Stenosis

Stenosis, No. (Degree of Stenosis, %)	Positive Predictive Value of a Bruit (95% CI)
37 (100)	5 (4-7)
113 (80-99)	16 (13-19)
207 (50-79)	29 (26-32)
113 (16-49)	16 (13-19)
180 (1-15)	25 (22-29)
64 (Normal)	9 (7-11)

Abbreviation: CI, confidence interval.

TITLE Predictive Power of Duplex Ultrasonography in Asymptomatic Carotid Disease.

AUTHORS Lewis R, Abrahamowicz M, Core R, Battista RN.

CITATION *Ann Intern Med.* 1997;127(1):13-20.

QUESTION What is the prevalence of carotid stenosis in a large cohort of asymptomatic patients?

DESIGN Prospective natural history study and randomized trial of aspirin vs placebo, begun in 1988.¹

SETTING Multicenter.

PATIENTS General practitioners and specialists referred patients from community and teaching hospital settings for evaluation of carotid stenosis. Patients were excluded if they had cerebrovascular symptoms, valvular heart disease, recent myocardial infarction, and a variety of other conditions that would have affected outcomes in the randomized trial. Seven hundred fourteen patients were enrolled, with the focus of this review being only the baseline evaluation.

The patient population showed a typical prevalence of patients with atherosclerotic risk: mean age, 65 years; hypertension, 47%; heart disease, 39%; hyperlipidemia, 50%; diabetes, 20%; and current smokers, 35%.

CONCLUSIONS

LEVEL OF EVIDENCE

Positive predictive value study.

STRENGTHS A typical population of patients referred for ultrasonography. However, what makes this study unique is that all of the patients were asymptomatic for cerebrovascular disease. The ultrasonographers validated their proficiency.

LIMITATIONS The results generalize only to populations with the same prevalence of carotid stenosis among patients with carotid bruits. No patient who lacked a carotid bruit was included, so the sensitivity and specificity cannot be determined.

The study population and trial design are similar to those of an earlier study.² Furthermore, the patients in the 2 studies are similar in terms of their risk factors for atherosclerotic disease, which is important because the positive predictive value of a test depends on the prevalence of disease. The 2 studies had almost identical positive predictive values for carotid stenosis (21% in this study for stenosis $\geq 80\%$ vs 23% in the earlier study that used a cut point of 75%).

REFERENCES FOR THE EVIDENCE

- Asymptomatic Cervical Bruit Study Group. Natural history and effectiveness of aspirin in asymptomatic patients with cervical bruits. *Arch Neurol.* 1991;48(7):683-686.
- Chambers BR, Norris JW. Outcome in patients with asymptomatic neck bruits. *N Engl J Med.* 1986;315(14):860-865.

Reviewed by David L. Simel, MD, MHS

Table 9-14 Likelihood Ratios of Bruit for Carotid Stenosis of at Least 70%^a

Test	Sensitivity	Specificity	LR+ (95% CI)	LR- (95% CI)
Bruit	0.56	0.91	6.0 (3.2-10)	0.48 (0.25-0.74)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

^aData are not broken out for symptomatic vs asymptomatic patients.

MAIN RESULTS

See [Table 9-14](#).

CONCLUSIONS

LEVEL OF EVIDENCE Level 3.

LIMITATIONS Relatively small sample size, referred population.

STRENGTHS Includes a mixture of patients with and without cerebrovascular symptoms. Auscultation was done without knowledge of ultrasonographic results.

The study enrolled consecutive referred patients and includes a population of patients with and without symptoms. With patients at various risk levels of cerebrovascular disease, the results ought to overlap with other populations of asymptomatic patients and symptomatic patients—the confidence intervals for the likelihood ratios are similar to those of most other studies.

Reviewed by David L. Simel, MD, MHS

TITLE Carotid Artery Auscultation—Anachronism or Useful Screening Procedure?

AUTHORS Magyar MT, Nam E, Csiba L, Ritter MA, Ringelstein EB, Droste DW.

CITATION *Neurol Res.* 2002;24(7):705-708.

QUESTION Among patients referred for carotid ultrasonographic studies, does the presence of a bruit predict carotid stenosis of 70% to 99%?

DESIGN Prospective, consecutive patients referred for ultrasonography.

SETTING Single site. Inpatients and outpatients of a neurology department at a university hospital (Germany) who were referred for carotid ultrasonography.

PATIENTS A total of 145 patients, of whom 43% had no history of cerebrovascular event (“asymptomatic”).

The sample reflects a referred population of patients at risk for atherosclerotic vascular disease (hypertension, 43%; hyperlipidemia, 35%; smokers, 24%; angina, 19%; previous myocardial infarction, 18%; claudication, 12%; and diabetes, 12%), although other patients were referred for lower-risk conditions (vertigo, dizziness, and psychosomatic symptoms). A total of 273 carotid arteries were evaluated.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

A single physician blinded to the patient’s medical history and the ultrasonographic results conducted the carotid auscultation. A different physician performed the carotid ultrasonography.

MAIN OUTCOME MEASURE

Carotid stenosis of 70% to 99%.

TITLE Can Simple Clinical Features Be Used to Identify Patients With Severe Carotid Stenosis on Doppler Ultrasound?

AUTHORS Mead GE, Warlaw JM, Lewis SC, McDowall M, Dennis MS.

CITATION *J Neurol Neurosurg Psychol.* 1999;66(1):16-19.

QUESTION Do carotid bruits, or a combination of clinical findings, predict the presence of significant carotid stenosis in symptomatic patients?

DESIGN Prospective.

SETTING British hospital and neurovascular clinic.

PATIENTS A total of 726 patients with an acute stroke, transient ischemic attack, or retinal stroke entered into the Lothian Stroke Registry. All patients had ultrasonography, independent of whether or not a bruit was detected.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

All patients were examined by a stroke physician or research registrar. Carotid Doppler ultrasonography was performed by one of 2 neuroradiologists who had excellent agreement with a subset of patients referred to angiography ($\kappa = 0.7-0.8$). The ultrasonographers were blinded to the clinical data.

MAIN OUTCOME MEASURES

Stenosis of 70% to 99% by ultrasonography vs a nonsurgical stenosis (<70% or complete occlusion). Data were evaluated for univariate predictors and in a logistic model to assess for combinations of findings that might predict surgically correctable carotid stenosis.

MAIN RESULTS

See Table 9-15. For the logistic model evaluating the combination of findings, the presence of an ipsilateral bruit (odds ratio, 11; 95% confidence interval [CI], 7.0-19) overwhelms other significant findings, making the likelihood ratio (LR) for 2 or more findings similar to that for an ipsilateral bruit alone.

CONCLUSIONS

LEVEL OF EVIDENCE Level 1.

STRENGTHS Prospective design for a large number of symptomatic patients. All patients underwent ultrasonography and were included in the analysis, even for lower degrees of carotid stenosis.

Table 9-15 Likelihood Ratios of Bruit for Carotid Stenosis of at Least 70%

Test	Sensitivity	Specificity	LR+ (95% CI)	LR- (95% CI)
Ipsilateral bruit	0.56	0.90	5.5 (4.1-7.2)	0.48 (0.38-0.60)
Peripheral vascular disease ^a	0.28	0.84	1.7 (1.2-2.4)	0.86 (0.74-0.96)
Diabetes	0.16	0.91	1.7 (1.0-2.8)	0.93 (0.83-1.0)
Combination of Findings (Ipsilateral Bruit, Diabetes, Previous TIA, Not a Lacunar Event)				
≥2 Findings			4.8 (3.5-6.5)	
0-1 Finding			0.57 (0.46-0.69)	

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio; TIA, transient ischemic attack.

^aDefined as absence of both foot pulses or femoral bruits, history of intermittent claudication, or a history of peripheral vascular surgery.

LIMITATIONS The case definition for a “lacunar” event was not described and is important only when the results of the logistic model are applied.

This is a high-quality study that appears to have less verification bias than the report on bruits from the North American Symptomatic Carotid Endarterectomy Trial.¹ As expected, when there is less verification bias, the specificity is much better and accounts for the higher positive LR. In this study, which showed a 13% prevalence, a clinician would have to screen 3 patients to detect 1 with a bruit indicating a 70% to 99% stenosis (number needed to screen, 95% CI, 2-5 patients). In probability terms, finding a carotid bruit increases the probability of a surgical carotid stenotic lesion from 13% to 46%. However, what if the patient has no bruit? The LR may not be good enough for most clinicians in that the probability of a 70% to 99% lesion only decreases from 13% to 7%. These are the types of data that lead prudent physicians to infer that it is acceptable to listen to every symptomatic patient’s carotid arteries for bruits, but the results ought to be ignored if the patient is a suitable candidate for surgery. In other words, clinicians should not use the absence of a bruit to “rule out” carotid stenosis in symptomatic patients who would otherwise be amenable to endarterectomy.

REFERENCE FOR THE EVIDENCE

1. Sauve JS, Thorpe KE, Sackett DL, et al. Can bruits distinguish high-grade from moderate symptomatic carotid stenosis? *Ann Intern Med.* 1994;120(8):633-637.

Reviewed by David L. Simel, MD, MHS

TITLE Can Bruits Distinguish High-Grade From Moderate Symptomatic Carotid Stenosis?

AUTHORS Sauve JS, Thorpe KE, Sackett DL, et al.

CITATION *Ann Intern Med.* 1994;120(8):633-637.

QUESTION Does the presence of a carotid bruit in a patient with a recent transient ischemic attack or nondisabling stroke predict whether the patient will have high-grade carotid stenosis (70%-99%) vs a less significant stenotic lesion (30%-69%)?

DESIGN Analysis of data collected prospectively as part of the North American Symptomatic Carotid Endarterectomy Trial of carotid endarterectomy.¹

SETTING Multicenter study at hospitals that qualified by proving their excellence in multidisciplinary care of patients with cerebrovascular disease.

PATIENTS Patients had to have a qualifying cerebrovascular event and be appropriate candidates for carotid endarterectomy. The analytic set for this study includes 1268 patients of 4526 patients screened. Of the 1268 patients, 667 (53%) had a carotid stenosis of 70% to 99%.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

The original data were collected prospectively at enrollment. The bruits were described as focal or diffuse and ipsilateral or contralateral. The examiners could have known the angiogram results before their evaluation. The reference standard was applied to all patients included in the final analysis. All patients had duplex ultrasonography of the carotid arteries and had carotid arteriogram, performed by neuroradiologists. The angiograms were reviewed by a data coordinating center.

MAIN OUTCOME MEASURE

Significant carotid stenosis (70%-99%) vs nonsurgical carotid stenosis (30%-69%).

MAIN RESULTS

See Table 9-16. The analysis focuses on focal ipsilateral bruits. Insufficient data were provided to assess the confidence intervals around diffuse ipsilateral bruits or contralateral bruits. A variety of risk factors collected during the history did not distinguish between high- and low-grade stenosis: hypertension, diabetes, hyperlipidemia, smoking, claudication, angina pectoris, myocardial infarction, heart failure, valvular heart disease, and atrial fibrillation, among others, were not useful.

Table 9-16 Likelihood Ratios of Bruit for Carotid Stenosis of at Least 70%

Test	Sensitivity	Specificity	LR+ (95% CI)	LR- (95% CI)
Focal, ipsilateral bruit	0.63	0.61	1.6 (1.4-1.8)	0.64 (0.54-0.68)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

CONCLUSIONS

LEVEL OF EVIDENCE Level 4 for answering the diagnostic accuracy questions.

STRENGTHS The study had a large sample size from a well-designed, randomized, controlled clinical trial.

LIMITATIONS Clinicians may have had access to the results of the ultrasonography or angiography. Verification bias exists in that patients without stenosis were excluded, so users of these data must understand the population before generalizing the results.

The parent study from which these data were obtained was a well-designed clinical trial. However, the trial was not designed to assess the diagnostic power of carotid bruits. Nonetheless, it is appropriate to see what we can learn from such a rich data set. Understanding the study question is critical to understanding the results. The study exhibits verification bias for answering the diagnostic question of whether carotid bruits identify significant carotid stenosis in symptomatic patients. Verification bias typically, but not always, leads to overestimates of sensitivity (ie, a too optimistic negative likelihood ratio [LR-]) and underestimates specificity (ie, a too pessimistic positive likelihood ratio [LR+]). The effect can be dramatic, such that if all patients were included (including those without any carotid stenosis), the LR+ most certainly would have been higher for the presence of a carotid bruit. However, it is certain that the absence of an ipsilateral bruit in a patient with a recent carotid artery distributed cerebrovascular event does little to rule out the presence of a significant carotid stenosis.

This study is also a bit different from other studies in that the comparison group did not consist of all patients, but only those with moderate stenosis. By excluding patients with lesser degrees of stenosis, the presence of a bruit would lose some of its discriminatory power and both the LR+ and LR- would look worse in comparison.

A second form of bias may also be in play in this study—expectation bias. Expectation bias occurs when the examiner has a preset belief about the presence of a finding. For example, if the examiner knows that the patient has a high-grade stenosis, the examiner may expect to hear a bruit (or vice versa). It is difficult to assess the effect of expectation bias on the LRs because they could make the values change in either direction.

REFERENCE FOR THE EVIDENCE

- North American Symptomatic Carotid Endarterectomy Trial (NASCET) Steering Committee. North American Symptomatic Carotid Endarterectomy Trial. *Stroke*. 1991;22(6):711-720.

Reviewed by David L. Simel, MD, MHS

Table 9-17 Predictive Value of a Carotid Bruit for Identifying Various Levels of Carotid Stenosis

Test	Outcome	LR+ (95% CI)	LR- (95% CI)
Bruit	Stroke during follow-up	1.5 (0.95-2.2)	0.96 (0.92-1.0)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

TITLE The Prognostic Significance of Asymptomatic Carotid Bruits in the Elderly.

AUTHORS Shorr RI, Johnson KC, Wan JY, et al.

CITATION *J Gen Intern Med*. 1998;13(2):86-90.

QUESTION Does the presence of a carotid bruit predict subsequent stroke in older patients with hypertension?

DESIGN Prospective, observational study among patients enrolled in the Systolic Hypertension in the Elderly Program (SHEP).¹ The mean follow-up was 4.5 years.

SETTING Multicenter trial in the United States.

PATIENTS Patients were aged 60 years or older, with isolated systolic hypertension, and formed part of a randomized clinical trial. The patients in this trial had no evidence of previous cerebrovascular disease, atrial fibrillation, insulin or warfarin use, coronary disease, or dementia.

In the SHEP trial, 4736 patients were enrolled, with 294 excluded from this analysis because they had a previous stroke, transient ischemic attack (TIA), or myocardial infarction. Thus, the analysis consists of 4442 patients, of whom 284 had an asymptomatic carotid bruit and 4158 had no bruit. Of those patients with bruits, 44% ($n = 124$) had bilateral carotid bruits.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

All patients were evaluated before study entry for carotid bruits. An adjudication committee reviewed medical records for all persons who developed symptoms suggestive of a stroke or TIA.

MAIN OUTCOME MEASURE

Stroke or TIA.

MAIN RESULTS

See Table 9-17. Patients with bruits were slightly older (mean age, 73 vs 71 years; $P < .001$) and less likely to be white (79% of patients with bruit were white vs 84% without bruit; $P =$

.03). The patients with bruits were also more likely to smoke (18% vs 12%; $P = .003$) and had higher blood pressures (mean systolic, 173 vs 170 mm Hg; $P < .001$), higher cholesterol levels (mean, 244 vs 236 mg/dL; $P = .006$), and more frequent electrocardiogram abnormalities (67% vs 60%; $P = .01$).

A proportional hazards model showed that the risk of a stroke or TIA did not change over time.

Although patients with bruits had a higher stroke rate than those without bruits, the difference is not significant. To explain part of the effect, patients with bruits were slightly more likely to be using aspirin (22% vs 16%), but they also were more likely to have been randomized to placebo for hypertension treatment (58% vs 50%). Adjusting for the hypertension treatment assignment vs placebo makes the carotid bruit effect slightly less, whereas adjusting for other risk factors extinguishes the effect of bruits even more (relative risk [RR], 1.3). Even when creating 2 strata of patients (ie, low risk vs high risk according to the number of risk factors present), the RR of stroke for those with carotid bruits is 1.38 vs 1.36, respectively. The risk of bruits might be greater in the subset of patients 60 to 69 years of age (RR, 2.0; 95% confidence interval [CI], 0.92-4.7), but the increased risk is less apparent for older patients (RR, 0.98; 95% CI, 0.55-1.8).

CONCLUSIONS

LEVEL OF EVIDENCE Level 1.

STRENGTHS Given the observed stroke rate in patients without bruits, the study had a post hoc power of 90% to find a difference of 5% strokes in patients without bruits vs 10% for patients with bruits. This is a high power to rule out an important difference, although smaller differences could have gone undetected. The prevalence (6.4%) of bruits in these asymptomatic hypertensive patients is compatible with a general, community population. The study had clear case definitions.

LIMITATIONS It is likely that the committee did have access to the medical records with information about the presence of a carotid bruit. However, given the rigorous case definitions, requirement for hospitalization for all patients with ischemic events, and adjudication by 3 neurologists, it seems unlikely that carotid bruits would have had a large effect on assessing the presence of a TIA or stroke.

Among patients with a single risk factor for stroke (hypertension), followed as part of a clinical trial, it seems clear that

any importance of a carotid bruit in predicting stroke is small. As in the Fremantle Diabetes Study,² the effect of the bruit is likely related to its association with other risk factors for atherosclerosis. Both studies showed that patients with bruits were more likely to have important risk factors for atherosclerosis.

REFERENCES FOR THE EVIDENCE

1. Perry HM, Davis BR, Price TR, et al. Effect of treating isolated systolic hypertension on the risk of developing various types and subtypes of stroke: the Systolic Hypertension in the Elderly Program (SHEP). *JAMA*. 2000;284(4):465-471.
2. Gillett M, Davis WA, Jackson D, Bruce DG, Davis TME. Prospective evaluation of carotid bruit as a predictor of first stroke in type 2 diabetes: The Fremantle Diabetes Study. *Stroke*. 2003;34(9):2145-2151.

Reviewed by David L. Simel, MD, MHS

This page intentionally left blank

CHAPTER 10

Does This Patient Have Carpal Tunnel Syndrome?

Christopher A. D'Arcy, MD

Steven McGee, MD

CLINICAL SCENARIO

A 55-year-old woman has difficulty sleeping because of numbness and tingling in her right hand for 6 months. On a hand diagram, she uses a pencil to locate precisely her numbness and tingling over the dorsal and palmar aspects of all 5 fingers, sparing the palm. On inspection, the patient has no evidence of thenar atrophy, but thumb abduction is weak on the affected side. Sensory examination result using monofilaments and a vibrating tuning fork is normal. Tinel sign is positive and Phalen sign is negative. Which of this patient's symptoms and signs are useful and which are useless for accurately predicting the diagnosis of carpal tunnel syndrome (CTS)?

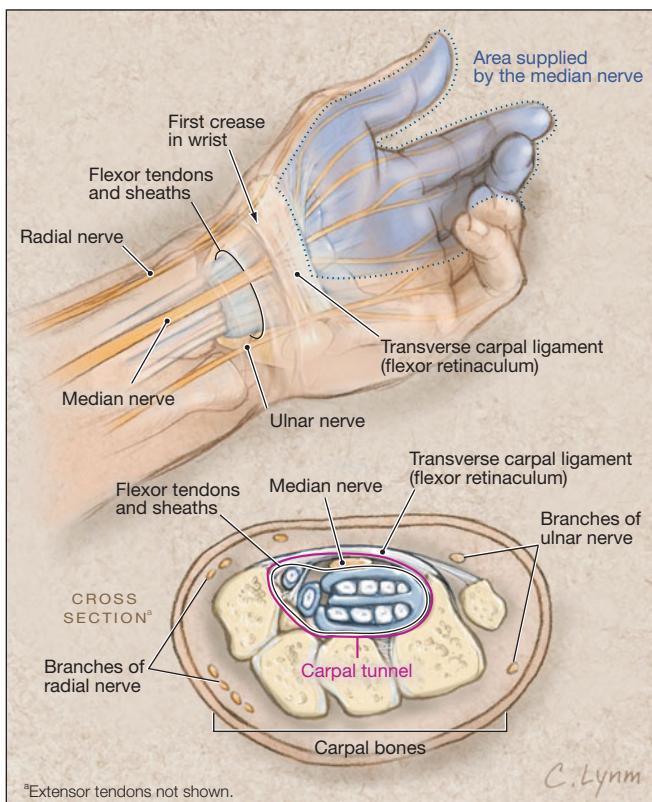
WHY IS THE DIAGNOSIS IMPORTANT?

Carpal tunnel syndrome is an important cause of hand pain and functional impairment, attributable to compression of the median nerve at the wrist (Figure 10-1). Patients are usually between 30 and 50 years old, with women affected 3 times as often as men.^{2,3} About 0.5% of the general population reports being diagnosed with CTS.² It is likely, however, that a minority of affected patients consult clinicians because population-based studies reveal that about 3% of adults have symptomatic electrodiagnostically confirmed CTS.⁴

In many patients, symptoms are self-limited or resolve with conservative measures such as splinting the wrist, using anti-inflammatory medication, and modifying their activities. Corticosteroid injection into or near the carpal tunnel results in improvement in 49%-81% of those affected, although 50%-86% of those experience recurrence.⁵⁻⁹ In patients whose condition fails conservative treatment, surgical division of the transverse carpal ligament promptly improves or relieves sensory complaints (dysesthesias) 75% to 99% of the time.¹⁰⁻¹⁸ Permanent complications from surgery occur in less than 1%,¹⁹ but the subsequent recovery often requires leave from work, lasting days to several weeks.¹⁸

Many conditions, including pregnancy, rheumatoid arthritis, diabetes mellitus, and previous wrist trauma, are associated with CTS,¹⁹ although histologic sections from the carpal tunnel of most affected patients are normal.^{20,21} Many patients have an abnormally high tissue pressure within the carpal tunnel,²² which presumably causes intraneuronal ischemia that leads to dysesthesias and abnormal results of sensory testing.²³⁻²⁵

This article systematically reviews the diagnostic accuracy of bedside findings for CTS. Presentation of this information, however, first requires understanding some of the issues

**Figure 10-1 Normal Anatomy of the Carpal Tunnel**

The carpal tunnel consists of the median nerve and 9 flexor tendons surrounded by the rigid carpal bones and transverse carpal ligament (flexor retinaculum). The distal wrist crease marks the proximal edge of the carpal tunnel. Within the tunnel, the median nerve divides into a motor branch that innervates the thenar muscles (opponens, abductor, and short flexor) and distal sensory branches that supply the thumb, index, and middle fingers and the radial half of the ring finger. Because the sensory branches to the radial palm do not usually pass through the carpal tunnel, palm sensation is preserved in a classic case of carpal tunnel syndrome.¹

surrounding electrodiagnosis, the current CTS diagnostic standard.

THE DIAGNOSTIC STANDARD FOR CARPAL TUNNEL SYNDROME

In his original definition of CTS, Phalen²⁶ required patients to have 1 or more of 3 bedside findings: sensory changes restricted to the median nerve distribution of the hand (Table 10-1), a positive Tinel sign, and a positive Phalen sign. Although electrodiagnosis was not part of Phalen's definition, clinicians now use electrodiagnosis frequently to confirm the diagnosis, and some third-party payers require it before compensating claims.³⁴ Consensus committees from professional societies have endorsed electrodiagnosis as the diagnostic test of choice.^{35,36} Diagnostic standards for nerve conduction studies in CTS have been developed, which report sensitivities of 49% to 84% and specificities of 95% to 99%.³⁷

The sensitivity and specificity of electrodiagnosis in CTS requires explanation. For the sensitivity calculation, the criterion standard was bedside findings alone (eg, compatible

Table 10-1 Definition of Abnormal Physical Findings

Physical Finding	Definition of Abnormal Finding
Motor Examination	
Weak thumb	Weakness of resisted abduction, ie, movement of the thumb at right angles to the palm ^a
Thenar atrophy	A concavity of the thenar muscles when observed from the side
Sensory Examination	
Hypalgesia	Diminished ability to perceive painful stimuli applied along the palmar aspect of the index finger when compared with the ipsilateral little finger ^b
Diminished 2-point discrimination	Diminished ability to identify correctly the number of points using calipers whose points are set 4-6 mm apart, comparing the index with little finger ^c
Abnormal vibratory sensation	Diminished ability to perceive vibratory sensations using a standard vibrating tuning fork (128 or 256 Hz), comparing the distal interphalangeal joint of the index finger to the ipsilateral fifth finger
Abnormal monofilament testing	Using a Semmes-Weinstein monofilament applied to the pulp of the index finger, the patient's threshold is greater than the 2.83 monofilament
Other Tests	
Square wrist sign ²⁷	The anteroposterior dimension of the wrist divided by the mediolateral dimension equals a ratio of greater than 0.70, when measured with calipers at the distal wrist crease
Closed fist sign ²⁸	Paresthesias in the distribution of the median nerve when the patient actively flexes the fingers into a closed fist for 60 s
Flick sign ²⁹	When asking the patient, "What do you actually do with your hand(s) when the symptoms are at their worst?" the patient demonstrates a flicking movement of the wrist and hand, similar to that used in shaking down a thermometer ^d
Tinel sign	Paresthesias in the distribution of the median nerve when the clinician taps on the distal wrist crease over the median nerve
Phalen sign	Paresthesias in the distribution of the median nerve when the patient flexes both wrists 90° for 60 s
Pressure provocation test ³⁰	Paresthesias in the distribution of the median nerve when the examiner presses with his/her thumb on the palmar aspect of the patient's wrist at the level of the carpal tunnel for 60 s
Tourniquet test ³¹	Paresthesias in the distribution of the median nerve when a blood pressure cuff around the patient's arm is inflated above systolic pressure for 60 s

^aMost clinicians define weakness as muscle power less than that of the companion muscle in contralateral hand (which has the disadvantage of assuming that the opposite hand has normal strength) or that of a standard of normal strength based on the experience of examining many normal individuals (Figure 10-2).

^bMost clinicians use an open safety pin or broken applicator stick, which must be discarded after use to prevent transmission of infection.

^cThe studies in this review separated the points of the calipers 4 mm,³² 5 mm,³³ and 6 mm.³¹

^dAny other response is a negative result.

symptoms plus a positive Tinel sign),³⁸⁻⁴⁰ which then raises the question of whether electrodiagnosis or bedside findings are the more accurate standard. False-negative test results probably occur because the condition is intermittent⁴¹ or because the patient's symptoms emanate from small, unmy-

elinated fibers that are invisible to surface electrodes (electrodiagnosis detects only larger myelinated fibers).⁴²

The high specificity figures in these studies are also misleading, being arbitrarily set at 2 SDs above the mean of observations of normal hands. The values of 95% to 99% are based on the assumption that nerve conduction recordings follow a standard gaussian distribution, which has been shown to be inaccurate.^{43,44} False-positive test results are well documented when these test thresholds are applied to other populations.^{10,45-47}

It is well documented that many hand surgeons perform carpal tunnel release successfully in patients with normal electrodiagnostic findings.^{15,34,48-50} Even in patients with positive electrodiagnostic findings who undergo surgery, symptoms usually resolve within days despite nerve conduction abnormalities that persist for months or longer.^{11,17,42,51,52}

Nonetheless, most physicians rely on electrodiagnosis as the best available diagnostic standard. Electrodiagnostic studies may help identify other conditions that also cause hand dysesthesias, such as cervical radiculopathy, polyneuropathy, or other median nerve entrapment syndromes.^{41,53-55} Furthermore, the majority of patients in surgical studies have compatible symptoms and electrodiagnostic studies are positive for CTS.^{10,12,17,56} Electrodiagnosis may not predict recovery after carpal tunnel release, but neither does any other clinical variable with any certainty. The potential use of computed tomography (CT), magnetic resonance imaging (MRI), and ultrasonography is still being determined, and they remain primarily research tools.⁵⁷⁻⁶¹ For these reasons, our review addresses the accuracy of the history and physical examination in diagnosing CTS, as confirmed by electrodiagnostic studies.

METHODS

Using the MEDLINE database for articles from January 1966 to February 2000, both authors independently used the following search strategy, limited to the English language and human subjects, to retrieve all relevant publications on the diagnosis of CTS in adults: “exp carpal tunnel syndrome” and “exp diagnosis.” In addition, text word searches were completed for “Tinel” or “Tinels” or “Hoffman-Tinels,” and “Phalen” or “Phalens.” Based on review of titles and abstracts, relevant publications were retrieved. To complete the search, the authors reviewed the bibliographies of these articles and retrieved all relevant articles.

To be included in this review, a study had to satisfy the following criteria: (1) the patients presented to a clinician for symptoms suggestive of CTS, (2) the physical examination maneuvers were clearly described, (3) there was an independent comparison with 1 or more electrodiagnostic parameters (which had to include at least some measurement of motor or sensory nerve conduction), and (4) the authors could extract from figures or tables in the articles the numbers needed to construct 2 × 2 tables and calculate sensitivity, specificity, and likelihood ratios (LRs).

Twelve articles met these criteria and are included.^{27-33,62-66} Thirty articles were excluded: 14 because the control group

was asymptomatic,⁶⁷⁻⁸⁰ 8 because the data were incomplete,^{15,49,57,81-85} 4 because the participants were identified by population surveys,^{45,86-88} 3 because the criterion standard was unacceptable (ie, electromyography alone,⁸⁹ electrodiagnosis and abnormal monofilament testing,⁹⁰ or criterion standard missing⁹¹), and 1 because the examination maneuvers were not clearly defined.⁹²

Sensitivity, specificity, and LRs and their confidence intervals (CIs) were calculated using conventional definitions.⁹³ When a cell of a 2 × 2 table was 0, 0.5 was added to all cells before summarizing the data for a particular test. Our summary measures pooled all the data using the DerSimonian and Laird⁹⁴ random-effects model, which considers both within-study variance and variability among studies. Our test for homogeneity between studies was the effectiveness score, a test of overall accuracy.⁹⁵

LRs are the odds that a given finding would occur in a patient with CTS as opposed to a patient without CTS. When a positive LR (LR+) or negative LR (LR-) has a value close to 1, the result is unhelpful in clinical diagnosis.

PRECISION AND ACCURACY

How to Elicit Symptoms and Signs of Carpal Tunnel Syndrome

Table 10-1 summarizes how to elicit the physical examination signs of CTS analyzed in this review. When examining thumb strength, the clinician should focus on abduction of the thumb (Figure 10-2), not flexion or opposition, which sometimes can be accomplished by muscles innervated by nerves other than the recurrent motor branch of the median nerve.^{54,59} The Katz hand diagram is a self-administered diagram that depicts both the dorsal and palmar aspect of the patient’s hands and arms (Figure 10-3). Patients use this diagram to mark the specific location of their symptoms, characterizing them as pain, numbness or tingling, or other. Diagrams are then graded as classic, probable, possible, or unlikely to be CTS on the basis of criteria that appear in Figure 10-3.^{32,63}



Figure 10-2 Testing Thumb Abduction

The patient is instructed to raise his or her thumb perpendicular to the palm as the examiner applies downward pressure on the distal phalanx. This maneuver reliably isolates the strength of the abductor pollicis brevis, which is innervated only by the median nerve.

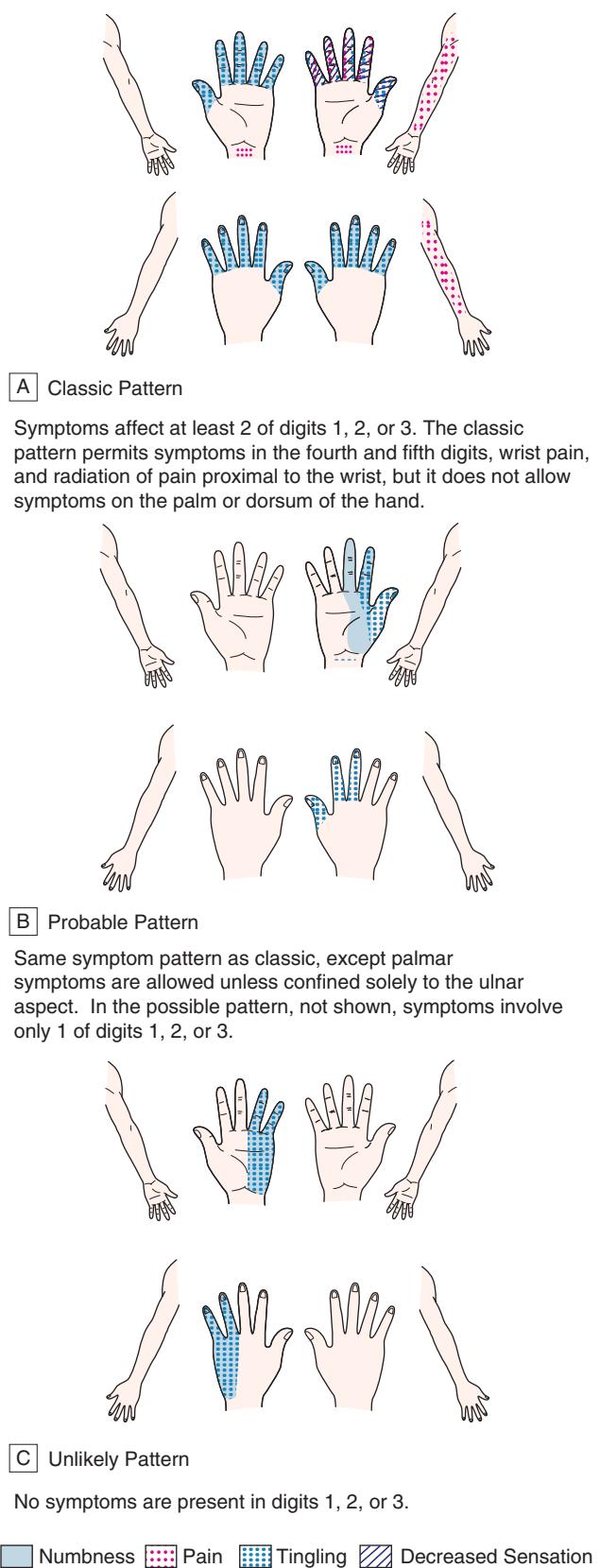


Figure 10-3 Katz Hand Diagram
Adapted with permission from Golding et al.⁶⁴

Precision of the History and Physical Examination for Carpal Tunnel Syndrome

Few studies have addressed the precision of findings for CTS. In one study, simple agreement was 84% for 2 physicians rating 54 of the Katz hand diagrams.⁶³ In another small study, the interobserver agreement was substantial for Tinel sign ($\kappa = 0.77$) and Phalen sign ($\kappa = 0.65$), moderate for vibration ($\kappa = 0.40$), and fair for motor strength ($\kappa = 0.25$).⁶⁶ The Tinel test, however, is probably much less precise than these data suggest because the proportion of healthy, asymptomatic hands with a positive Tinel sign ranges from 0%²⁸ to 45%.⁷¹ Some of this variability with Tinel sign may relate to technique; in one study, a greater percussion force increased sensitivity at the expense of specificity.⁸⁹

Diagnostic Accuracy of Physical Findings

Table 10-2 summarizes the studies addressing the diagnostic accuracy of the history and physical examination for CTS. Based on the CIs of LRs, the following findings favor the electrodiagnosis of CTS when they are present in patients who present with hand dysesthesias: decreased sensitivity to pain (hypalgesia) in the median nerve territory (LR, 3.1; 95% CI, 2.0-5.1), classic or probable Katz hand diagram results (LR, 2.4; 95% CI, 1.6-3.5), and weak thumb abduction strength (LR, 1.8; 95% CI, 1.4-2.3). Using a slightly different system for grading hand diagrams, another study also found that the definite or possible hand diagram argued for CTS (LR, 2.1; 95% CI, 1.5-3.0).⁹² In our analysis, 2 findings argued against the electrodiagnosis of CTS: a Katz hand diagram classified as unlikely (LR, 0.2; 95% CI, 0.0-0.7; not shown in Table 10-2) and normal thumb abduction strength (LR, 0.5; 95% CI, 0.4-0.7).

The following findings had limited or no value in distinguishing patients with CTS from those without it: the patient's age, presence of bilateral or nocturnal symptoms, thenar atrophy, other sensory abnormalities (2-point, vibration, monofilament), Tinel sign, Phalen sign, pressure provocation test, and the tourniquet test.

Several studies addressed the diagnostic accuracy of combined findings,^{32,65,90} but no combination consistently proved significantly more helpful than the individual findings themselves. One study did find that a positive Tinel sign with a classic or probable hand diagram was slightly more discriminating (LR, 3.6; 95% CI, 1.6-8.1) than either finding alone (LR, 1.8 for positive Tinel sign and 2.4 for classic or probable hand diagram),³² although this result requires validation, given the problems with Tinel sign in other studies.

According to our analysis, several unconventional findings—flick sign, closed fist sign, and square wrist sign—show promise in diagnosing CTS. However, these maneuvers are not widely used and have been tested in only one or two studies. Two letters to editors suggest that the sensitivity of the flick sign is much lower (only 25%-36%) than

Table 10-2 Diagnostic Accuracy of History and Physical Examination for Carpal Tunnel Syndrome

Findings by Reference and Year	No. of Hands	Sensitivity	Specificity	LR+ (95% CI)	LR- (95% CI)
Patient Interview					
Classic or Probable Hand Diagram					
Katz et al, ⁶³ 1990	145	0.64	0.73	2.4 (1.6-3.5)	0.5 (0.3-0.7)
Age > 40 y					
Katz et al, ³² 1990	110 ^a	0.80	0.41	1.3 (1.0-1.7)	0.5 (0.3-1.0)
Nocturnal Paresthesia					
Buch-Jaeger and Foucher, ³¹ 1994	112 ^a	0.51	0.68	1.6 (1.0-2.6)	0.7 (0.5-1.0)
Gupta and Benstead, ⁶² 1997	92	0.84	0.33	1.2 (1.0-1.6)	0.5 (0.2-1.1)
Katz et al, ³² 1990	110	0.77	0.27	1.1 (0.9-1.3)	0.8 (0.4-1.6)
Pooled results	1.2 (1.0-1.4)	0.7 (0.5-0.9)
Bilateral Symptoms					
Katz et al, ³² 1990	110 ^a	0.61	0.58	1.4 (1.0-2.1)	0.7 (0.4-1.0)
Motor Examination					
Weak Thumb Abduction					
Gerr et al, ³³ 1995	115	0.63	0.62	1.7 (1.1-2.4)	0.6 (0.4-0.9)
Kuhlman and Hennessey, ³⁰ 1997	228	0.66	0.66	2.0 (1.4-2.7)	0.5 (0.4-0.7)
Pooled results	1.8 (1.4-2.3)	0.5 (0.4-0.7)
Thenar Atrophy					
Gerr et al, ³³ 1995	115	0.28	0.82	1.6 (0.8-3.2)	0.9 (0.7-1.1)
Golding et al, ⁶⁴ 1986	110	0.04	0.99	5.4 (0.2-130)	1.0 (0.9-1.0)
Katz et al, ³² 1990	110 ^a	0.14	0.90	1.5 (0.5-4.1)	0.9 (0.8-1.1)
Pooled results	1.6 (0.9-2.8)	1.0 (0.9-1.0)
Sensory Examination					
Hypalgesia					
Golding et al, ⁶⁴ 1986	110	0.15	0.93	2.2 (0.7-6.7)	0.9 (0.8-1.1)
Kuhlman and Hennessey, ³⁰ 1997	228	0.51	0.85	3.4 (2.0-5.8)	0.6 (0.5-0.7)
Pooled results	3.1 (2.0-5.1)	0.7 (0.5-1.1)
2-Point Discrimination					
Buch-Jaeger and Foucher, ³¹ 1994, 6 mm	167	0.06	0.99	4.5 (0.6-37)	1.0 (0.9-1.0)
Gerr et al, ³³ 1995, 5 mm	115	0.28	0.64	0.8 (0.5-1.3)	1.1 (0.9-1.5)
Katz et al, ³² 1990, 4 mm	110 ^a	0.32	0.80	1.6 (0.8-3.1)	0.8 (0.7-1.1)
Pooled results	1.3 (0.6-2.7)	1.0 (0.9-1.1)
Abnormal Vibration					
Buch-Jaeger and Foucher, ³¹ 1994	172	0.20	0.81	1.1 (0.6-2.0)	1.0 (0.8-1.1)
Gerr et al, ³³ 1995	115	0.61	0.71	2.1 (1.3-3.3)	0.5 (0.4-0.8)
Pooled results	1.6 (0.8-3.0)	0.8 (0.4-1.3)
Abnormal Monofilament Findings					
Buch-Jaeger and Foucher, ³¹ 1994	167	0.59	0.59	1.5 (1.1-2.0)	0.7 (0.5-0.9)
Other Tests					
Square Wrist Sign					
Kuhlman and Hennessey, ³⁰ 1997	228	0.69	0.73	2.6 (1.8-3.7)	0.4 (0.3-0.6)
Radecki, ²⁷ 1994	665	0.47	0.83	2.8 (2.1-3.8)	0.6 (0.6-0.7)
Pooled results	2.7 (2.2-3.4)	0.5 (0.4-0.8)
Closed Fist Sign					
De Smet et al, ²⁸ 1995	35	0.61	0.92	7.3 (1.1-49)	0.4 (0.2-0.7)
Flick Sign					
Pryse-Phillips, ²⁹ 1984	396	0.93	0.96	21 (11-42)	0.1 (0-0.1)

(Continued)

Table 10-2 Diagnostic Accuracy of History and Physical Examination for Carpal Tunnel Syndrome (*Continued*)

Findings by Reference and Year	No. of Hands	Sensitivity	Specificity	LR+ (95% CI)	LR- (95% CI)
Other Tests					
Tinel Sign					
Gerr et al, ³³ 1995	115	0.25	0.67	0.7 (0.4-1.3)	1.1 (0.9-1.4)
Golding et al, ⁶⁴ 1986	110	0.26	0.80	1.3 (0.6-2.6)	0.9 (0.7-1.2)
Heller et al, ⁶⁵ 1986	80	0.60	0.77	2.7 (1.2-5.9)	0.5 (0.3-0.8)
Katz et al, ³² 1990	110 ^a	0.59	0.67	1.8 (1.2-2.7)	0.6 (0.4-0.9)
Kuhlman and Hennessey, ³⁰ 1997	228	0.23	0.87	1.8 (1.0-3.4)	0.9 (0.8-1.0)
Buch-Jaeger and Foucher, ³¹ 1994	172	0.42	0.64	1.1 (0.8-1.7)	0.9 (0.7-1.2)
Pooled results	1.4 (1.0-1.9)	0.8 (0.7-1.0)
Phalen Sign					
Buch-Jaeger and Foucher, ³¹ 1994	166	0.58	0.54	1.3 (0.9-1.7)	0.8 (0.6-1.1)
Gerr et al, ³³ 1995	115	0.75	0.33	1.1 (0.9-1.4)	0.7 (0.4-1.3)
Heller et al, ⁶⁵ 1986	80	0.67	0.59	1.6 (1.0-2.8)	0.6 (0.3-0.9)
Katz et al, ³² 1990	110 ^a	0.75	0.47	1.4 (1.1-1.9)	0.5 (0.3-0.9)
Kuhlman and Hennessey, ³⁰ 1997	228	0.51	0.76	2.1 (1.4-3.2)	0.6 (0.5-0.8)
Golding et al, ⁶⁴ 1986	110	0.10	0.86	0.7 (0.2-2.2)	1.0 (0.9-1.2)
Burke et al, ⁶⁶ 1999	200	0.51	0.54	1.1 (0.7-1.8)	0.9 (0.6-1.3)
De Smet et al, ²⁸ 1995	66	0.91	0.33	1.4 (0.9-2.0)	0.3 (0.1-0.9)
Pooled results	1.3 (1.1-1.6)	0.7 (0.6-0.9)
Pressure Provocation Test					
Kuhlman and Hennessey, ³⁰ 1997	228	0.28	0.74	1.1 (0.7-1.7)	1.0 (0.8-1.1)
Burke et al, ⁶⁶ 1999	205	0.52	0.38	0.8 (0.6-1.2)	1.3 (0.7-2.2)
Buch-Jaeger and Foucher, ³¹ 1994	155	0.49	0.54	1.1 (0.8-1.5)	0.9 (0.7-1.3)
De Smet et al, ²⁸ 1995	66	0.63	0.33	0.9 (0.6-1.5)	1.1 (0.5-2.7)
Pooled results	1.0 (0.8-1.3)	1.0 (0.9-1.1)
Tourniquet Test					
Buch-Jaeger and Foucher, ³¹ 1994	145	0.52	0.36	0.8 (0.6-1.1)	1.3 (0.9-2.0)
Golding et al, ⁶⁴ 1986	110	0.21	0.87	1.6 (0.7-3.9)	0.9 (0.8-1.1)
Pooled results	1.0 (0.5-1.9)	1.0 (0.7-1.5)

Abbreviations: CI, confidence interval; LR, likelihood ratio.

A positive LR (LR+) indicates a positive finding for carpal tunnel syndrome; a negative LR (LR-) indicates either a negative finding or an absent finding.

^aRefers to individual subjects instead of individual hands.

^bEllipses indicate not applicable.

that indicated in Table 10-2.^{84,85} Therefore, before any of these 3 findings can be recommended, further supportive evidence is necessary.

There are several reasons why some findings are not as helpful diagnostically as traditionally thought. Thenar atrophy is probably not useful because it occurs only in long-standing or neglected cases of CTS and can also result from lower cervical radiculopathies or polyneuropathies. Tinel described his sign for following the course of regenerating nerve in patients after blunt traumatic nerve injury.^{30,76,87} The idea that patients with CTS would also have a stub of continually regenerating nerve at the distal wrist crease seems unlikely, limiting the diagnostic utility of this particular test. Our analysis shows that hypalgesia in the median nerve distribution is a more useful diagnostic finding than are abnormalities of other sensory modalities, in part because hypalgesia is a more specific

finding. It is not clear why this should be, although it may indicate that the threshold for abnormal results when testing sensation for vibration, 2-point discrimination, and monofilaments is set too low (eg, in one study, 20% of asymptomatic hands also displayed abnormal monofilament results⁷⁶).

In our analysis, only results for the Tinel sign were heterogeneous. The heterogeneity is not explained by differences in the electrodiagnostic parameters used as criterion standards in the individual studies, variations in examination technique (ie, whether the clinician tapped over the median nerve using the index finger or a reflex hammer), differences in prevalence of CTS in each of the studies (mean prevalence was 57%), differences in the age and sex composition (mean age was 50 years; 77% were women), or by an apparent workup bias. Excluding the 2 studies that account for the heterogeneity^{62,64} does not change the

summary measure in any meaningful way, and therefore, these studies are included in our analysis.

THE BOTTOM LINE

When evaluating patients with hand dysesthesias, the findings most helpful in predicting the electrodiagnosis of CTS are hand symptom diagrams, hypalgesia, and weak thumb abduction strength testing. The square wrist sign, flick sign, and closed fist sign also show promise but require validation by other investigators. Many traditional findings, including Phalen and Tinel signs, have limited ability to predict the electrodiagnosis of CTS.

The main limitation of the existing literature is the lack of an ideal criterion standard, which complicates all clinical research in the field of CTS. It is also important that these data are derived from symptomatic patients presenting to a surgeon, physical therapist, or an electrodiagnostic laboratory. There are no data addressing the value of physical diagnosis in patients presenting to a primary care physician with symptoms suggestive of CTS. Our analysis, therefore, is most applicable to patients with severe enough symptoms to warrant such a referral.

Returning to the case presented at the beginning of the article, the findings of a classic hand diagram and thumb abduction weakness support the diagnosis of CTS. The findings of a normal thenar eminence, a positive Tinel sign, and a negative Phalen sign do not contribute significant diagnostic information. The patient's clinician believed that she probably had CTS and chose to manage her symptoms by splinting her wrists and recommending anti-inflammatory medications. If the patient's symptoms fail to improve, nerve conduction testing, additional empiric therapeutic modalities (eg, corticosteroid injections), or referral for surgical assessment should be considered.

Author Affiliations at the Time of the Original Publication

University of Washington Health Sciences Center, Seattle (Dr D'Arcy); and University of Washington, Seattle–Puget Sound Veterans Affairs Health Care System (Dr McGee), Seattle.

Acknowledgments

We thank Jaya Rao, MD, MHS, and Richard W. Tim, MD, who reviewed this article and provided many helpful comments.

REFERENCES

- Lum PB, Kanaklamedala R. Conduction of the palmar cutaneous branch of the median nerve. *Arch Phys Med Rehabil*. 1986;67(11):805-806.
- Tanaka S, Wild D, Seligman P, et al. The US prevalence of self-reported carpal tunnel syndrome: 1988 national health interview survey data. *Am J Public Health*. 1994;84(11):1846-1848.
- Stevens JC, Sun S, Beard CM, O'Fallon WM, Kurland L. Carpal tunnel syndrome in Rochester, Minnesota, 1961-1980. *Neurology*. 1988;38(1):134-138.
- Atroshi I, Gummesson C, Johnsson R, et al. Prevalence of carpal tunnel syndrome in a general population. *JAMA*. 1999;282(2):153-158.
- Green DP. Diagnostic and therapeutic value of carpal tunnel injection. *J Hand Surg [Am]*. 1984;9(6):850-854.
- Gelberman RH, Aronson D, Weisman MH. Carpal tunnel syndrome: results of a prospective trial of steroid injection and splinting. *J Bone Joint Surg Am*. 1980;62(7):1181-1184.
- Weiss AP, Sachar K, Gendreau M. Conservative management of carpal tunnel syndrome: a reexamination of steroid injection and splinting. *J Hand Surg [Am]*. 1994;19(3):410-415.
- Dammers JWHH, Veering MM, Vermeulen M. Injection with methylprednisolone proximal to the carpal tunnel: randomized double blind trial. *BMJ*. 1999;319(7214):884-886.
- Gainer JV Jr, Nugent GR. Carpal tunnel syndrome: report of 430 operations. *South Med J*. 1977;70(3):325-328.
- Czeuz KA, Thomas JE, Lambert EH, Love JG, Lipscomb PR. Long-term results of operation for carpal tunnel syndrome. *Mayo Clin Proc*. 1966;41(4):232-241.
- Bande S, De Smet L, Fabry G. The results of carpal tunnel release: open versus endoscopic technique. *J Hand Surg [Br]*. 1994;19(1):14-17.
- Tountas CP, Macdonald CJ, Meyerhoff JD, Bahrle DM. Carpal tunnel syndrome: a review of 507 patients. *Minn Med*. 1983;66(8):479-482.
- Mühlau G, Both R, Kunath H. Carpal tunnel syndrome—course and prognosis. *J Neurol*. 1984;231(2):83-86.
- Kendall D. Aetiology, diagnosis, and treatment of paraesthesiae in the hands. *BMJ*. 1960;2(5213):1633-1640.
- Phalen GS. The carpal-tunnel syndrome: seventeen years' experience in diagnosis and treatment of six hundred fifty-four hands. *J Bone Joint Surg Am*. 1966;48(2):211-228.
- Doyle JR, Carroll RE. The carpal tunnel syndrome: a review of 100 patients treated surgically. *Calif Med*. 1968;108(4):263-267.
- Brown RA, Gelberman RH, Seiler JG 3rd, et al. Carpal tunnel release: a prospective, randomized assessment of open and endoscopic methods. *J Bone Joint Surg Am*. 1993;75(9):1265-1275.
- Boeckstyns MEH, Sorensen AI. Does endoscopic carpal tunnel release have a higher rate of complications than open carpal tunnel release? An analysis of published series. *J Hand Surg [Br]*. 1999;24:9-15.
- Stevens JC, Beard CM, O'Fallon WM, Kurland L. Conditions associated with carpal tunnel syndrome. *Mayo Clin Proc*. 1992;67(6):541-548.
- Nakamichi K, Tachibana S. Histology of the transverse carpal ligament and flexor tenosynovium in idiopathic carpal tunnel syndrome. *J Hand Surg [Am]*. 1998;23(6):1015-1024.
- Kerr CD, Sybert DR, Albarracin NS. An analysis of the flexor synovium in idiopathic carpal tunnel syndrome: report of 625 cases. *J Hand Surg [Am]*. 1992;17(6):1028-1030.
- Gelberman RH, Hergenroeder PT, Hargens AR, Lundborg GN, Akeson WH. The carpal tunnel syndrome: a study of carpal canal pressures. *J Bone Joint Surg Am*. 1981;63(3):380-383.
- Gelberman RH, Szabo RM, Williamson RV, Dimick MP. Sensibility testing in peripheral nerve compression syndromes: an experimental study in humans. *J Bone Joint Surg Am*. 1983;65(5):632-638.
- Lundborg G, Gelberman RH, Minteer-Convery M, Lee YF, Hargens AR. Median nerve compression in the carpal tunnel—functional response to experimentally induced controlled pressure. *J Hand Surg [Am]*. 1982;7(3):252-259.
- Gelberman RH, Rydevik BL, Pess GM, Szabo RM, Lundborg G. Carpal tunnel syndrome: a scientific basis for clinical care. *Orthop Clin North Am*. 1988;19(1):115-124.
- Phalen GS. The birth of a syndrome, or carpal tunnel revisited. *J Hand Surg [Am]*. 1981;6(2):109-110.
- Radecki P. A gender specific wrist ratio and the likelihood of a median nerve abnormality at the carpal tunnel. *Am J Phys Med Rehabil*. 1994;73(3):157-162.
- De Smet L, Steenwerckx A, Van Den Bogaert G, Cnudde P, Fabry G. Value of clinical provocative tests in carpal tunnel syndrome. *Acta Orthop Belg*. 1995;61(3):177-182.
- Pryse-Phillips WE. Validation of a diagnostic sign in carpal tunnel syndrome. *J Neurol Neurosurg Psychiatry*. 1984;47(8):870-872.
- Kuhlman KA, Hennessey WJ. Sensitivity and specificity of carpal tunnel syndrome signs. *Am J Phys Med Rehabil*. 1997;76(6):451-457.
- Buch-Jaeger N, Foucher G. Correlation of clinical signs with nerve conduction tests in the diagnosis of carpal tunnel syndrome. *J Hand Surg [Br]*. 1994;19(6):720-724.
- Katz JN, Larson MG, Sabra A, et al. Carpal tunnel syndrome: diagnostic utility of history and physical examination findings. *Ann Intern Med*. 1990;112(5):321-327.

33. Gerr F, Letz R, Harris-Abbott D, Hopkins LC. Sensitivity and specificity of vibrometry for detection of carpal tunnel syndrome. *J Occup Environ Med.* 1995;37(9):1108-1115.
34. Concannon MJ, Gainor B, Petroski GJ, Puckett CL. The predictive value of electrodiagnostic studies in carpal tunnel syndrome. *Plast Reconstr Surg.* 1997;100(6):1452-1458.
35. American Academy of Neurology, American Association of Electrodiagnostic Medicine, American Academy of Physical Medicine and Rehabilitation. Practice parameter for electrodiagnostic studies in carpal tunnel syndrome (summary statement). *Neurology.* 1993;43(11):2404-2405.
36. Quality Standards Subcommittee of the American Academy of Neurology. Practice parameter for carpal tunnel syndrome (summary statement). *Neurology.* 1993;43(11):2406-2409.
37. Jablecki CK, Andary MT, So YT, Wilkins DE, Williams FH. Literature review of the usefulness of nerve conduction studies and electromyography for the evaluation of patients with carpal tunnel syndrome. *Muscle Nerve.* 1993;16(12):1392-1414.
38. Kimura J. The carpal tunnel syndrome: localization of conduction abnormalities within the distal segment of the median nerve. *Brain.* 1979;102(3):619-635.
39. Nathan P, Meadow KD, Doyle LS. Sensory segmental latency values of the median nerve for a population of normal individuals. *Arch Phys Med Rehabil.* 1988;69(7):499-501.
40. Jackson DH, Clifford JC. Electrodiagnosis of mild carpal tunnel syndrome. *Arch Phys Med Rehabil.* 1989;70:199-204.
41. Dawson DM, Hallett M, Wilbourn AJ. Carpal tunnel syndrome. In: Dawson DM, Hallett M, Wilbourn AJ, eds. *Entrapment Neuropathies.* 3rd ed. Philadelphia, PA: Lippincott-Raven Publishers; 1999:20-94.
42. Gilliat RW, Sears TA. Sensory nerve action potentials in patients with peripheral nerve lesions. *J Neurol Neurosurg Psychiatry.* 1958;21(2):109-118.
43. Robinson LR, Temkin NR, Fujimoto WY, Stolov WC. Effect of statistical methodology on normal limits in nerve conduction studies. *Muscle Nerve.* 1991;14(11):1084-1090.
44. Goodgold J. A statistical problem in diagnosis of carpal tunnel disease. *Muscle Nerve.* 1994;17(12):1490-1491.
45. Ferry S, Silman AJ, Pritchard T, Keenan J, Croft P. The association between different patterns of hand symptoms and objective evidence of median nerve compression. *Arthritis Rheum.* 1998;41(4):720-724.
46. Thomas JE, Lambert EH, Cseuz KA. Electrodiagnostic aspects of the carpal tunnel syndrome. *Arch Neurol.* 1967;16(6):635-641.
47. Redmond KD, Rivner MH. False-positive electrodiagnostic tests in carpal tunnel syndrome. *Muscle Nerve.* 1988;11(5):511-518.
48. Grundberg AB. Carpal tunnel decompression in spite of normal electromyography. *J Hand Surg [Am].* 1983;8(3):348-349.
49. Phalen GS. The carpal tunnel syndrome: clinical evaluation of 598 hands. *Clin Orthop Relat Res.* 1972;83:29-40.
50. Mainous III AG, Nelson KR. How often are preoperative electrodiagnostic studies obtained for carpal tunnel syndrome in a Medicaid population? *Muscle Nerve.* 1996;19(2):256-257.
51. Harris CM, Tanner E, Goldstein MN, Pettee DS. The surgical treatment of the carpal-tunnel syndrome correlated with preoperative nerve-conduction studies. *J Bone Joint Surg Am.* 1979;61(1):93-98.
52. Pätiälä H, Rokkanen P, Kruuna O, et al. Carpal tunnel syndrome: anatomical and clinical investigation. *Arch Orthop Trauma Surg.* 1985;104(2):69-73.
53. Kaufman MA. Differential diagnosis and pitfalls in electrodiagnostic studies and special tests for diagnosing compressive neuropathies. *Orthop Clin North Am.* 1996;27(2):245-252.
54. Spinner RJ, Bachman JW, Amadio PC. The many faces of carpal tunnel syndrome. *Mayo Clin Proc.* 1989;64(7):829-836.
55. Haig AJ, Tzeng HM, LeBreck D. The value of electrodiagnostic consultation for patients with upper extremity nerve complaints: a prospective comparison with the history and physical examination. *Arch Phys Med Rehabil.* 1999;80(10):1273-1281.
56. Bessette L, Keller RB, Lew RH, et al. Prognostic value of a hand symptom diagram in surgery for carpal tunnel syndrome. *J Rheumatol.* 1997;24(4):726-734.
57. Rosenbaum RB. The role of imaging in the diagnosis of carpal tunnel syndrome. *Invest Radiol.* 1993;28(11):1059-1062.
58. Winn FJ Jr, Habes DJ. Carpal tunnel area as a risk factor for carpal tunnel syndrome. *Muscle Nerve.* 1990;13(3):254-258.
59. Cantatore FP, Dell'Accio F, Lapadula G. Carpal tunnel syndrome: a review. *Clin Rheumatol.* 1997;16(6):596-603.
60. Seyfert S, Boegner F, Hamm B, Kleindienst A, Klatt C. The value of magnetic resonance imaging in carpal tunnel syndrome. *J Neurol.* 1994;242(1):41-46.
61. Lee D, van Holsbeeck MT, Janevski PK, et al. Diagnosis of carpal tunnel syndrome: ultrasound versus electromyography. *Radiol Clin North Am.* 1999;37(4):859-872.
62. Gupta SK, Benstead TJ. Symptoms experienced by patients with carpal tunnel syndrome. *Can J Neurol Sci.* 1997;24(4):338-342.
63. Katz JN, Stirrat C, Larson MG, et al. A self-administered hand symptom diagram in the diagnosis and epidemiologic study of carpal tunnel syndrome. *J Rheumatol.* 1990;17(11):1495-1498.
64. Golding DN, Rose DM, Selvarajah K. Clinical tests for carpal tunnel syndrome: an evaluation. *Br J Rheumatol.* 1986;25(4):388-390.
65. Heller L, Ring H, Costeff H, Solzi P. Evaluation of Tinel and Phalen signs in the diagnosis of the carpal tunnel syndrome. *Eur Neurol.* 1986;25(1):40-42.
66. Burke DT, Burke MA, Bell R, et al. Subjective swelling: a new sign for carpal tunnel syndrome. *Am J Phys Med Rehabil.* 1999;78(6):504-508.
67. Yii NW, Elliot D. A study of the dynamic relationship of the lumbrical muscles and the carpal tunnel. *J Hand Surg [Br].* 1994;19(4):439-443.
68. González del Pino J, Delgado-Martínez AD, González González I, Lovic A. Value of the carpal compression test in the diagnosis of carpal tunnel syndrome. *J Hand Surg [Br].* 1997;22(1):38-41.
69. Bowles AP Jr, Asher SW, Pickett JD. Use of Tinel's sign in carpal tunnel syndrome. *Ann Neurol.* 1983;13(6):689-690.
70. Durkan JA. A new diagnostic test for carpal tunnel syndrome. *J Bone Joint Surg Am.* 1991;73(4):535-538.
71. Seror P. Tinel's sign in the diagnosis of carpal tunnel syndrome. *J Hand Surg [Br].* 1987;12(3):364-365.
72. Seror P. Phalen's test in the diagnosis of carpal tunnel syndrome. *J Hand Surg [Br].* 1988;13(4):383-385.
73. Tetro AM, Evanoff BA, Hollstien SB, Gelberman RH. A new provocative test for carpal tunnel syndrome: assessment of wrist flexion and nerve compression. *J Bone Joint Surg Br.* 1998;80(3):493-498.
74. Williams TM, Mackinnon SE, Novak CB, McCabe S, Kelly L. Verification of the pressure provocative test in carpal tunnel syndrome. *Ann Plast Surg.* 1992;29(1):8-11.
75. Stewart JD, Eisen E. Tinel's sign and the carpal tunnel syndrome. *BMJ.* 1978;2(6145):1125-1126.
76. Gelmers HJ. The significance of Tinel's sign in the diagnosis of carpal tunnel syndrome. *Acta Neurochir (Wien).* 1979;49(3-4):255-258.
77. Borg K, Lindblom U. Diagnostic value of quantitative sensory testing (QST) in carpal tunnel syndrome. *Acta Neurol Scand.* 1988;78(6):537-541.
78. Fertl E, Wöber C, Zeithofer J. The serial use of two provocative tests in the clinical diagnosis of carpal tunnel syndrome. *Acta Neurol Scand.* 1998;98(5):328-332.
79. Ghavanini MR, Haghight M. Carpal tunnel syndrome: reappraisal of five clinical tests. *Electromyogr Clin Neurophysiol.* 1998;38(7):437-441.
80. Koris M, Gelberman RH, Duncan K, Boubllick M, Smith B. Carpal tunnel syndrome: evaluation of a quantitative provocative diagnostic test. *Clin Orthop Relat Res.* 1990;(251):157-161.
81. Gilliatt RW, Wilson TG. A pneumatic-tourniquet test in the carpal tunnel syndrome. *Lancet.* 1953;265(6786):595-597.
82. Spindler H, Dellon A. Nerve conduction studies and sensibility testing in carpal tunnel syndrome. *J Hand Surg [Am].* 1982;7(3):260-263.
83. Szabo RM, Slater RR, Farver TB, Stanton DB, Sharman WK. The value of diagnostic testing in carpal tunnel syndrome. *J Hand Surg [Am].* 1999;24(4):704-714.
84. Roquer J, Herranz J. Validity of flick sign in CTS diagnosis. *Acta Neurol Scand.* 1988;78(4):351.
85. Krendell DA, Jöbsis M, Gaskell PC Jr, Sanders DB. The flick sign in carpal tunnel syndrome. *J Neurol Neurosurg Psychiatry.* 1986;49(2):220-221.
86. de Krom MC, Knipschild PG, Kester ADM, Spaans F. Efficacy of provocative tests for diagnosis of carpal tunnel syndrome. *Lancet.* 1990;335(8686):393-395.
87. Kuschner SH, Ebramzadeh E, Johnson D, Brien WW, Sherman R. Tinel's sign and Phalen's test in carpal tunnel syndrome. *Orthopedics.* 1992;15(11):1297-1302.
88. Homan MM, Franzblau A, Werner RA, et al. Agreement between symptom surveys, physical examination procedures and electrodiagnostic findings for the carpal tunnel syndrome. *Scand J Work Environ Health.* 1999;25(2):115-124.

89. Mossman SS, Blau JN. Tinel's sign and the carpal tunnel syndrome. *Br Med J (Clin Res Ed)*. 1987;294(6573):680.
90. Gellman H, Gelberman RH, Tan AM, Botte MJ. Carpal tunnel syndrome. *J Bone Joint Surg Am*. 1986;68(5):735-737.
91. Novak CB, Mackinnon SE, Brownlee R, Kelly L. Provocative sensory testing in carpal tunnel syndrome. *J Hand Surg [Br]*. 1994;19(5):817-820.
92. Stevens JC, Smith BE, Weaver AL, et al. Symptoms of 100 patients with electromyographically verified carpal tunnel syndrome. *Muscle Nerve*. 1999;22(10):1448-1456.
93. Simel DL, Samsa GP, Matchar DB. Likelihood ratios with confidence: sample size estimation for diagnostic test studies. *J Clin Epidemiol*. 1991; 44(8):763-770.
94. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Control Clin Trials*. 1986;7(3):177-188.
95. Hasselblad V, Hedges LV. Meta-analysis of screening and diagnostic tests. *Psychol Bull*. 1995;117(1):167-178.
96. Marx RG, Hudak PL, Bombardier C, et al. The reliability of physical examination for carpal tunnel syndrome. *J Hand Surg [Br]*. 1998;23(4): 499-502.

This page intentionally left blank

Prepared by David L. Simel, MD, MHS

Reviewed by Richard Bedlack, MD

CLINICAL SCENARIO

Your 50-year-old secretary complains to you that she cannot complete your clinic notes on the computer without her hands tingling, especially her thumb and second and third fingers. Her symptoms are there even when she is not typing. In fact, she says that she has more problems at home because discomfort in her hands awakens her at night. She has had diabetes for 6 years. You purchase a variety of office supply products that might help her type and then wait to see whether her symptoms resolve.

A week later, she still has problems, although a cushion she ordered for her wrists has not arrived. You check for Tinel sign (which she has), and when you flex her wrists, it reproduces her symptoms. You suggest that she consult her primary care physician, and she asks you what to expect. You suggest that her physician assess her diabetes to see whether she might have a neuropathy, check neck radiographs to ensure there is no evidence of cervical degenerative changes, and review thyroid function tests, nerve conduction tests, and a magnetic resonance image (MRI) of the wrists. Have you requested all the necessary tests, or did you suggest too many?

Clinical Examination article, we were interested only in studies that assessed clinical findings in a population of patients with hand symptoms, that were an independent comparison with electrodiagnosis, and from which we could extract the data. The abstracts were reviewed to identify studies that might allow us to assess the sensitivity and specificity either of the findings judged helpful in the original review (eg, hand symptom diagram, hypalgesia, and thumb abduction strength testing) or for less commonly used maneuvers that required additional data (eg, square-wrist sign, flick sign, and closed-fist sign). We found 12 original articles for further review. A review of the reference lists identified 6 other articles that were obtained. For original articles, we retained those that studied at least 100 hands.

We excluded articles that used normal persons without symptoms as a control population or that were retrospective studies, which is necessary because the usefulness of tests can be overstated when a population of patients for whom CTS would not be considered is included.¹ Including "normal" control patients tends to overstate the specificity and makes it appear that a finding helps identify those with the disorder. For example, a Phalen sign has a positive likelihood ratio (LR+) of 2.9 when normal, asymptomatic patients are included. However, when only symptomatic patients for whom CTS would be considered are studied, the finding appeared useless in the same study, with an LR+ of 0.91.¹

No systematic review of the clinical examination findings used the inclusion criteria we required. A systematic review of surgery for CTS evaluated the role of electrodiagnostic testing as a suitable reference standard for predicting a successful outcome.

UPDATED SUMMARY ON CARPAL TUNNEL SYNDROME

Original Review

D'Arcy C, McGee S. Does this patient have carpal tunnel syndrome? *JAMA*. 2000;283(23):3110-3117.

UPDATED LITERATURE SEARCH

Our literature search used the parent search for The Rational Clinical Examination series, which combined the subject heading carpal tunnel syndrome (CTS) with meta-analysis or receiver operating characteristic curve. The results were crossed with the text words "Phalen," "Tinel," "square wrist," "thumb abduction," "hypalgesia," "closed fist," "flick," or "hand diagram" appearing in studies published in English from 1999 to 2004. The results yielded 141 titles and abstracts for review. As in the original Rational

NEW FINDINGS

- People flick their hands when they have hand symptoms, whether or not they have CTS.²
- Clinical maneuvers designed to induce or exacerbate the patient's symptoms cause them discomfort, but do nothing to alter the likelihood for or against CTS.³⁻⁵
- Additional evidence confirms the uselessness of Tinel or Phalen signs.^{2,6}

- Combining symptoms² and signs⁷ does not appear to improve accuracy.
- Clinicians should focus further diagnostic efforts on patients with symptoms in the median nerve distribution. These symptomatic patients are the only patients who will meet the reference standard criteria of combined hand diagram results and electrodiagnosis.

IMPROVEMENTS IN THE DATA PRESENTED IN THE ORIGINAL PUBLICATION

Additional data confirm the lack of utility for Tinel or Phalen signs and provocation tests. New summary estimates are provided for these findings. No studies were found that were missed in the original publication.

New data help us come up with prior probability estimates for CTS. When screened by a questionnaire, about 10% of patients in the community claim numbness or tingling in the radial fingers (median nerve distribution) in at least 1 of their hands.^{8,9} About 70% of patients with numbness or tingling in a median nerve distribution will complete hand diagrams that suggest “classic or probable” CTS.² Thus, among all adults, the prior probability of hand symptoms compatible with CTS is 7% (ie, 0.10×0.70). Because the diagnosis of CTS is considered only when the patient has hand symptoms, we can use the value of 7% as a starting point for our prior probability of CTS. This makes sense because the classic/probable distribution on the hand diagram is part of our pragmatic reference standard for CTS. These estimates from a population sample are supported by a large clinical sample of patients referred for electrodiagnosis; among 8223 electrodiagnostic studies in patients evaluated for CTS,⁷ the distribution of positive electrodiagnostic studies is the following:

- First, second, and third finger symptoms: 26% positive
- All fingers (1-5): 17%

CHANGES IN THE REFERENCE STANDARD

The original publication in The Rational Clinical Examination series focused on patients with CTS symptoms who had their disease status confirmed by electrical studies. A letter to the editor highlighted the dilemma in making this diagnosis,

with the author’s suggestion that we should have titled the article “Does This Patient Have Abnormal Median Conduction?”¹⁰ Some researchers have advocated MRI to identify affected patients. A systematic review of MRI revealed that much-higher-quality evidence must be generated before MRI can be accepted as a screening test, but it seems unlikely that it will ever suffice as a reference standard.¹¹ The use of electrodiagnosis for CTS is not perfect. The explanations for the fallibility of electrodiagnosis as “the” reference standard are as follows: some patients have clinically significant nerve compression with normal electrodiagnosis study results, the use of population means and standard deviations to define normality ensures that 2.5% of the population will have CTS (ie, the area beyond 2 SDs of 1 tail in the normal distribution curve for median nerve conduction velocity), and studies use various cut points for normality on median nerve testing.¹²

A group of experts in carpal tunnel epidemiology, clinical care, and outcome assessment used a nominal group process method to develop case definitions suitable for epidemiologic research.¹³ Although the authors state that their criteria were not meant for actual clinical practice, we used these criteria in the original review in The Rational Clinical Examination article, and they reflect the combination of symptoms and electrodiagnosis that most clinicians use to establish the diagnosis (Table 10-3). The symptoms refer to the Katz hand diagram as shown in Figure 10-3 of the original Rational Clinical Examination article.

A systematic review by a panel of neurology experts identified 497 articles published from 1990 to 2000 on CTS diagnosis.¹⁴ According to formal criteria that included (among others) prospective study design and that all patients must have had a clinical diagnosis of CTS performed independently of electrodiagnosis, they retained 25 articles for review. Their meta-analysis found a pooled sensitivity of 0.85 and a specificity of 0.98 for sensory or mixed median nerve conduction to confirm the clinical diagnosis. At face value, this seems reassuring. However, the group noted the problems with selection bias and observer bias in extant studies of CTS and electrodiagnosis. They proposed clinical diagnostic criteria for future CTS research that give important insight into the symptoms that primary care providers should evaluate. As in the Rempel et al¹³ report, no particular physical examination findings are required to establish the clinical diagnosis¹⁴ (Table 10-4). The combination of clinical diagnosis and electrodiagnosis serves as both a suitable epidemiologic standard and pragmatic clinical reference standard for primary care clinicians. However, it is clear that some patients with classic symptoms but normal electrodiagnosis can improve with treatment of CTS.

Jordan et al¹⁵ performed a systematic review of surgical therapy for CTS, specifically for assessing whether the results of electrodiagnostic testing predicted treatment response. They found the results not only of generally poor quality but also showing no differences in surgical outcomes for patients with symptoms and positive electrodiag-

Table 10-3 Carpal Tunnel Syndrome (CTS) Using the Paired Hand Diagram and Electrodiagnostic Results as the Reference Standard

Symptom	Electrodiagnosis	Ordinal Rank in Terms of Likelihood of CTS
Classic/probable	Abnormal	1 (Most likely)
Possible	Abnormal	2
Classic/probable	Negative	3
Possible	Negative	4
Unlikely	Abnormal	5
Unlikely	Negative	6 (Least likely)

nosis vs symptoms and normal electrodiagnostic study results. Of the 4 studies they included with relative risk data, the confidence interval included 1 for the relative risk, favoring good outcomes for those with a positive electrodiagnosis vs those with a normal electrodiagnosis. Three recent Cochrane reviews of CTS treatment found a few studies that did not require electrodiagnosis, but none included an analysis of whether patients diagnosed with symptoms alone had a different response compared with those with symptoms plus abnormal electrodiagnostic test results.¹⁶⁻¹⁸

RESULTS OF LITERATURE REVIEW

See [Table 10-5](#).

EVIDENCE FROM GUIDELINES

No US or Canadian guidelines exist for routine screening for CTS.

CLINICAL SCENARIO—RESOLUTION

The diagnosis of CTS seems reasonably certain, given that your secretary has the appropriate symptoms in the appropriate distribution (median nerve). You did not need to do the Tinel sign or make her fingers tingle with a provocation test. The suggestion that she be evaluated for diabetic neuropathy is important. Neck radiographs do not seem indicated unless there are some other symptoms to suggest a cervical problem. A systematic review of routine testing for diabetes, thyroid disease, or rheumatoid arthritis in patients with CTS showed that this practice infrequently picks up new diagnoses and is not necessary.²⁰ An electrodiagnostic test result, if positive, would mean that she meets the research criteria for CTS. MRI does not have an established role in diagnostic assessment for CTS.

The remaining question is, should you have suggested a nerve conduction study? A nerve conduction study might be indicated as part of an assessment for a systemic neuropathy. Her carpal tunnel symptoms, together with a positive electrodiagnostic test, would fulfill the accepted reference standard for research studies. However, some patients with positive symptoms have normal nerve conduction study results. It might be appropriate to wait and see whether she responds to simple ergonomic measures, wrist splinting, and, perhaps, steroid injection for short-term relief before considering the nerve conduction test.

Table 10-4 Carpal Tunnel Syndrome (CTS) Diagnosis Using the Paired-Hand Diagram, Additional Symptoms, and Electrodiagnostic Results as the Reference Standard

Inclusion Criteria for CTS for Research Studies on Electrodiagnosis
1. Symptom distribution as noted above (but the fourth finger is also allowed)
2. Symptoms must be present for 1 month, and there must be periods when the symptoms are intermittent
3. Symptoms must be aggravated by sleep, sustained hand or arm positioning, or repetitive motion of the hand
4. Symptoms must be relieved by change in hand position, shaking the hand, or use of a wrist splint
5. When pain is present, the pain in the wrist, hand, or finger must be worse than any pain in the elbow, shoulder, or neck
Exclusion Criteria for CTS for Research Studies on Electrodiagnosis
1. Symptoms primarily in the fifth finger
2. Neck or shoulder pain preceding digital paresthesias
3. Numbness or paresthesias in the feet that preceded hand symptoms
4. Another disorder that explains symptoms that is more likely than CTS

Table 10-5 Likelihood Ratios for a Variety of Signs and Combinations of Findings for Carpal Tunnel Syndrome

Finding (n = No. of Combined Studies)	Sensitivity	Specificity	LR+ (95% CI)	LR- (95% CI)
Tinel (n = 8) ^a			1.5 (1.2-2.1)	0.82 (0.72-0.93)
Phalen (n = 10) ^a			1.3 (1.2-1.5)	0.74 (0.62-0.87)
Provocation tests (n = 8) ^a			1.1 (0.96-1.3)	0.89 (0.79-1.0)
Multivariate model with 11 clinical variables (n = 1) ^b	0.79	0.54	1.7 (1.6-1.8)	0.39 (0.35-0.43)
Flick or Tinel (n = 1) ^c	0.46	0.68	1.5 (0.94-2.4)	0.79 (0.60-1.0)
Phalen or Tinel (n = 1) ^c	0.41	0.72	1.5 (0.89-2.5)	0.81 (0.63-1.0)
Flick (n = 1) ^c	0.37	0.74	1.4 (0.80-2.4)	0.85 (0.68-1.1)
Flick or Phalen (n = 1) ^c	0.49	0.62	1.3 (0.86-2.0)	0.82 (0.61-1.1)
Abnormal monofilament in digits 1, 2, or 3 (n = 1) ^c	0.98	0.15	1.2 (1.0-1.3)	0.11 (0.02-0.64)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

^aUpdated summary adds data from Hansen et al² and O'Gradaigh and Merry⁶ to data from the original Rational Clinical Examination article.

^bA multivariate model⁷ using 4 symptoms (nocturnal symptoms, morning symptoms, worsens on driving, and relieved by "waking and shaking"), symptom distribution, side of worst symptoms, handedness, duration of symptoms, response to splinting, and patient age was studied with a large "training" set and "test" set. The model had an accuracy of only 66% (area under the receiver operating characteristic curve).

^cThe sensitivity from this study¹⁹ requires confirmation in additional studies.

CARPAL TUNNEL SYNDROME—MAKE THE DIAGNOSIS

PRIOR PROBABILITY

Among all adults, the prior probability of hand symptoms compatible with CTS is 7%. See **Table 10-6** for the likelihood ratios for Tinel and Phalen signs.

Table 10-6 Likelihood Ratios for Tinel and Phalen Signs

Finding	LR
The presence of Tinel or Phalen signs in a patient with symptoms	≈1
The absence of Tinel or Phalen signs in a patient with symptoms	≈1

Abbreviation: LR, likelihood ratio.

POPULATION FOR WHOM CARPAL TUNNEL SYNDROME SHOULD BE CONSIDERED

- Patients with tingling or numbness in the hands or arms—always assess for median nerve involvement.
- Special populations include those with occupational exposure of repetitive motion or pregnancy in the third trimester.
- The rates of CTS might be slightly higher in those with diabetes mellitus, rheumatoid arthritis, or hypothyroidism. However, the data are not convincing, and routine screening for these diseases will infrequently lead to new diagnoses.

DETECTING THE LIKELIHOOD OF CARPAL TUNNEL SYNDROME

The examination should focus on the distribution of symptoms in a hand diagram, rather than provocative maneuvers to elicit symptoms.

REFERENCE STANDARD TESTS

The distribution of hand symptoms (from a hand diagram) plus abnormal nerve conduction studies is the reference standard for epidemiologic studies.

For clinical care, patients can have CTS despite a normal nerve conduction result. Data are inconclusive about whether treatment outcomes differ according to the nerve conduction results.

REFERENCES FOR THE UPDATE

- Gerr F, Letz R. The sensitivity and specificity of tests for carpal tunnel syndrome vary with the comparison subjects. *J Hand Surg*. 1998;23(2):151-155.
- Hansen PA, Mickelsen P, Robinson LR. Clinical utility of the flick maneuver in diagnosing carpal tunnel syndrome. *Am J Phys Med Rehabil*. 2004;83(5):363-367.^a
- Karl AI, Carney ML, Kaul MP. The lumbrical provocation test in subjects with median inclusive paresthesia. *Arch Phys Med Rehabil*. 2001;82(7):935-937.^a
- Kaul MP, Pagel KJ, Wheatley MJ, Dryden JD. Carpal compression test and pressure provocative test in veterans with median-distribution paresthesias. *Muscle Nerve*. 2001;24(1):107-111.^a
- Kaul MP, Pagel KJ, Dryden JD. Lack of predictive power of the “tethered” median stress test in suspected carpal tunnel syndrome. *Arch Phys Med Rehabil*. 2000;81(7):348-350.^a
- O’Gradaigh D, Merry P. A diagnostic algorithm for carpal tunnel syndrome based on Bayes’ theorem. *Rheumatology*. 2000;39(9):1040-1041.^a
- Bland JDP. The value of the history in the diagnosis of carpal tunnel syndrome. *J Hand Surgery [Br]*. 2000;25(5):445-450.^a
- Atroshti I, Gummesson C, Johnsson R, Ornstein E. Diagnostic properties of nerve conduction tests in population-based carpal tunnel syndrome. *BMC Musculoskelet Disord*. 2003;4:9.
- Reading J, Walker-Bone K, Palmer KT, Cooper C, Coggon D. Anatomic distribution of sensory symptoms in the hand and their relation to neck pain, psychosocial variables, and occupational activities. *Am J Epidemiol*. 2003;157(6):524-530.
- LeBlond RF. Clinical diagnosis of carpal tunnel syndrome [letter]. *JAMA*. 2000;284(15):1925; author reply, 1925-1926.
- Pasternack II, Malmivaara A, Tervahartiala P, Forsberg H, Vehmas T. Magnetic resonance imaging findings in respect to carpal tunnel syndrome. *Scand J Work Environ Health*. 2003;29(3):189-196.
- Graham B, Regehr G, Wright JG. Delphi as a method to establish consensus for diagnostic criteria. *J Clin Epidemiol*. 2003;56(12):1150-1156.
- Rempel D, Evanoff B, Amadio PC, et al. Consensus criteria for the classification of carpal tunnel syndrome in epidemiologic studies. *Am J Public Health*. 1998;88(10):1447-1451.
- Jablecki CK, Andary MT, Gloster MK, et al. Practice parameter: electrodiagnostic studies in carpal tunnel syndrome. *Neurology*. 2002;58(11):1589-1592.
- Jordan R, Carter T, Cummins C. A systematic review of the utility of electrodiagnostic testing in carpal tunnel syndrome. *Br J Gen Pract*. 2002;52(481):670-672.
- Verdugo RJ, Salinas RS, Castillo J, Cea JG. Surgical versus non-surgical treatment for carpal tunnel syndrome: Cochrane Neuromuscular Disease Group. *Cochrane Database Syst Rev*. 2002;(2):CD001552. doi: 10.1002/14651858.CD001552.
- O’Connor D, Marshall S, Massy-Westropp N. Non-surgical treatment (other than steroid injection) for carpal tunnel syndrome. *Cochrane Database Syst Rev*. 2003;(1):CD003219. doi: 10.1002/14651858.CD003219.
- Scholten RJPM, Mink van der Molen A, Uitdehaag BMJ, Bouter LM, de Vet HCW. Surgical treatment options for carpal tunnel syndrome. *Cochrane Database Syst Rev*. 2002;(4):CD003905. doi: 10.1002/14651858.CD003905.pub3.
- Pagel KJ, Kaul MP, Dryden JD. Lack of utility of Semmes-Weinstein monofilament testing in suspected carpal tunnel syndrome. *Am J Phys Med Rehabil*. 2002;81(8):597-600.^a
- van Dijk MAJ, Reitsma JB, Fischer JC, Sanders GTB. Indications for requesting laboratory tests for concurrent diseases in patients with carpal tunnel syndrome: a systematic review. *Clin Chem*. 2003;49(9):1437-1444.

^aFor the Evidence to Support the Update for this topic, see <http://www.JAMAevidence.com>.

10

EVIDENCE TO SUPPORT THE UPDATE:

Carpal Tunnel

TITLE The Value of the History in the Diagnosis of Carpal Tunnel Syndrome.

AUTHOR Bland JDP.

CITATION *J Hand Surgery [Br]*. 2000;25(5):445-450.

QUESTION Do any patient symptoms predict abnormality on electrodiagnostic studies?

DESIGN Data collected prospectively during an 8-year period.

SETTING Single center in the United Kingdom that performs all the electrodiagnostic studies for the local area.

PATIENTS Referred ($n = 8223$) for electrodiagnosis among a broad array of patients being considered for carpal tunnel surgery or for diagnostic evaluation.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

A questionnaire was given to all patients before the electrodiagnostic study. A single examiner performed all studies. It is likely that the examiner reviewed the questionnaire before the nerve conduction studies. See [Table 10-7](#).

Table 10-7 Likelihood Ratios of the Tinel, Flick, and Phalen Signs for Carpal Tunnel Syndrome

Test	Sensitivity	Specificity	LR+ (95% CI)	LR- (95% CI)
Tinel	0.27	0.91	3.2 (1.2-8.6)	0.79 (0.68-0.92)
Flick	0.37	0.74	1.4 (0.80-2.4)	0.85 (0.68-1.1)
Phalen	0.34	0.74	1.3 (0.74-2.3)	0.89 (0.71-1.1)
Phalen or Tinel	0.41	0.72	1.5 (0.89-2.5)	0.81 (0.63-1.0)
Flick or Tinel	0.46	0.68	1.5 (0.94-2.4)	0.79 (0.60-1.0)
Flick or Phalen	0.49	0.62	1.3 (0.86-2.0)	0.82 (0.61-1.1)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

MAIN OUTCOME MEASURE

A multivariate model using electrodiagnosis as the reference standard.

MAIN RESULTS

The data were split into a training set ($n = 5000$) and a test set ($n = 3223$). A logistic model for patient symptoms was created using the data for 5000 patients. The model contained 4 symptoms (nocturnal symptoms, morning symptoms, worse on driving, and relieved by “waking and shaking”), symptom distribution, side of worst symptoms, handedness, duration of symptoms, response to splinting, and patient age as continuous variables.

The only variables with an odds ratio (OR) greater than 2 were the presence of symptoms in the thumb and the second and third fingers (OR, 2.5; 95% confidence interval [CI], 2.1-3.0) or symptoms in the third and fourth fingers (OR, 2.4; 95% CI, 1.9-3.1). The only variable that had an OR less than 0.5 was the presence of symptoms in the fourth and fifth fingers (OR, 0.42; 95% CI, 0.29-0.62). As a continuous variable, age also had an important impact on the probability of carpal tunnel syndrome (CTS). For example, with a typical symptom pattern, without regard to any other symptom, a right-handed patient with right-handed symptoms has a predicted probability of 29% at age 30 years vs 66% at age 50 years.

CONCLUSIONS

LEVEL OF EVIDENCE Level 1.

STRENGTHS Very large patient population that captured all patients referred for electrodiagnostic studies. It is likely that these patients reflect the array of patients who are referred in other community studies for the evaluation of CTS.

LIMITATIONS The examiner would have known the results of the questionnaire (although the examiner would not have known the variables that would ultimately go in the logistic model).

The results of the logistic model would be difficult to apply in general practice. However, understanding the role of the dis-

tribution of symptoms in the digits is important and is integral to the current accepted reference standard of hand diagrams plus electrodiagnosis. Unfortunately, despite including 11 seemingly relevant clinical variables, the multivariate logistic model had a sensitivity of only 79% and a specificity of 54%. The area under the receiver operating characteristic (ROC) curve was only 0.66 (standard error of 0.01), reflecting an accuracy that seems too low for clinical use.

Reviewed by David L. Simel, MD, MHS

TITLE Clinical Utility of the Flick Maneuver in Diagnosing Carpal Tunnel Syndrome.

AUTHORS Hansen PA, Mickelsen P, Robinson LR.

CITATION *Am J Phys Med Rehabil.* 2004;83(5):363-367.

QUESTION Is the flick sign better than the Phalen or Tinel sign in identifying patients with hand symptoms who will have abnormal electrodiagnostic tests?

DESIGN Prospective, consecutive enrollment.

SETTING Electrodiagnostic clinic.

PATIENTS All patients ($n = 142$) had upper limb symptoms and were referred by their physicians for electrodiagnostic testing to establish the diagnosis. For all patients, carpal tunnel syndrome (CTS) was part of the differential diagnosis. When patients had bilateral symptoms, only the more severely affected hand was evaluated for the study.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

Standard assessment of the Phalen and Tinel signs. The flick sign was obtained by asking the patients how they relieved the discomfort in their hands and wrists when they were experiencing severe symptoms. Patients who demonstrated that they flick their hands (like shaking down a mercury thermometer) were considered "positive." The criterion standard was standard electrodiagnostic testing, performed after the clinical evaluation. It is not clear whether the same examiner did the clinical examination and the electrodiagnostic testing. However, the electrodiagnostic testing was based on the quantitative output nerve latency.

MAIN OUTCOME MEASURE

Electrodiagnosis of CTS.

MAIN RESULTS

One hundred forty-two patients were studied, of whom 95 had electrodiagnostic testing of CTS.

CONCLUSIONS

LEVEL OF EVIDENCE Level 1.

STRENGTHS Prospective, consecutive enrollment among a group of referred patients for whom CTS was part of their differential diagnosis. The examination was done before the electrodiagnostic test.

LIMITATIONS Electrodiagnosis may not have been blinded to the clinical findings, but the reporting of nerve conduction studies based on quantitative time rather than subjective time may make this less of a problem.

The authors sum up the results best: "people ... [with hand symptoms] flick their hands" whether or not they have CTS. These data confirm the uselessness of the Phalen sign. Unfortunately, the combination of the flick or Tinel sign does not improve the diagnostic efficiency. The positive likelihood ratio for the Tinel was the highest, but the confidence interval is broad (see Table 10-7).

Reviewed by David L. Simel, MD, MHS

TITLE The Lumbrical Provocation Test in Subjects With Median Inclusive Paresthesia.

AUTHORS Kaul AI, Carney ML, Kaul MP.

CITATION *Arch Phys Med Rehabil.* 2001;82(7):935-937.

TITLE Carpal Compression Test and Pressure Provocative Test in Veterans With Median-distribution Paresthesias.

AUTHORS Kaul MP, Pagel KJ, Wheatley MJ, Dryden JD.

CITATION *Muscle Nerve.* 2001;24(1):107-111.

TITLE Lack of Predictive Power of the "Tethered" Median Stress Test in Suspected Carpal Tunnel Syndrome.

AUTHORS Kaul MP, Pagel KJ, Dryden JD.

CITATION *Arch Phys Med Rehabil.* 2000;81(7):348-350.

QUESTION Does a physical examination maneuver meant to provoke symptoms predict patients who will have abnormal electrodiagnostic testing? Each study in this summary reports a different maneuver.

DESIGN Prospective, consecutive.

SETTING Electrodiagnostic laboratory of a Veterans Affairs medical center, Portland, Oregon.

PATIENTS In each study, patients had median nerve symptoms, no previous surgery for carpal tunnel syndrome, and no proximal neuropathy on the affected side.

Table 10-8 Likelihood Ratios of Provocation Tests for Carpal Tunnel Syndrome

Test (n)	Abnormal Electrodiagnostic Study Result	Sensitivity	Specificity	LR+ (95% CI)	LR- (95% CI)
Pressure provocation (134)	77	0.55	0.68	1.7 (1.1-2.7)	0.66 (0.49-0.90)
Carpal compression (135)	80	0.52	0.56	1.4 (0.94-2.1)	0.77 (0.56-1.0)
Lumbrical (fist) provocation (96)	51	0.37	0.71	1.3 (0.73-2.3)	0.88 (0.66-1.2)
"Tethered" median nerve stretch (112)	58	0.50	0.59	1.2 (0.8-1.9)	0.85 (0.59-1.2)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

Positive test results induce or exacerbate the median nerve symptoms.

The lumbrical provocation test is performed by having the patient hold a fist for 1 minute. (The lumbricales are the 4 small muscles of the palm of the hand that flex the proximal phalanx and extend the 2 distal phalanges of each finger.)

The "tethered" median nerve test creates a stretch of the median nerve by the examiner's passively hyperextending the wrist and distal interphalangeal joint of the index finger.

The carpal compression test is performed by applying moderate pressure with both thumbs over the transverse carpal ligament.

The pressure provocation test uses a 2.5-cm-wide pressure cuff applied to the patient's wrist. The cuff is inflated to 50 mm Hg, and then direct pressure is applied to bring the sphygmomanometer reading to 150 mm Hg.

The electrodiagnostic studies were performed immediately after the provocation tests. When the provocation test result was positive, the patient was allowed to have the symptoms return to baseline before the electrodiagnostic studies.

were based on a quantitative assessment. The results apply only to patients with median nerve symptoms.

Even with the possibility that the provocation test affected the electrodiagnostic studies, this maneuver did not work to identify the patients with median nerve symptoms who would have an abnormal electrodiagnosis. As in all clinical diagnosis studies, it is important to recognize that the clinicians included only patients with median nerve syndromes, something that can be evaluated at the bedside and is part of the recommended hand diagram. The provocation tests seem relatively useless as both the summary positive and negative likelihood ratios approach 1. Clinicians should stop trying to reproduce a patient's median nerve symptoms because the response should not affect clinical decisions.

Reviewed by David L. Simel, MD, MHS

MAIN OUTCOME MEASURE

Electrodiagnostic studies.

MAIN RESULTS

See Table 10-8.

CONCLUSIONS

LEVEL OF EVIDENCE Level 2.

STRENGTHS All patients had median nerve symptoms. The provocation tests were applied before the electrodiagnostic tests. An additional strength is that patients with neck pain were also included, as long as they also had median nerve symptoms.

LIMITATIONS The electrodiagnostic testing was performed blinded to the "tethered" median nerve test. It is not clear whether the electrodiagnostic tests were performed independently in the other 2 studies. However, the protocol for the electrodiagnostic procedure is described well and the results

TITLE A Diagnostic Algorithm for Carpal Tunnel Syndrome Based on Bayes' Theorem.

AUTHORS O'Gradaigh D, Merry P.

CITATION *Rheumatology*. 2000;39(9):1040-1041.

QUESTION Can the results of a hand diagram, Phalen test, and Tinel test be applied sequentially?

DESIGN Two-phase study. An initial study to determine the sensitivity and specificity of the findings may have been a convenience sample ($n = 105$ patients). The second phase assessed the sensitivity and specificity prospectively, but it is not stated whether these were consecutive patients ($n = 42$).

SETTING Rheumatology clinic in the United Kingdom.

PATIENTS Patients were referred because of a suspicion of carpal tunnel syndrome (CTS).

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

Patients completed a hand diagram. Patients with classic or probable patterns were considered to have a positive test result. Phalen and Tinel tests were done by a single examiner.

MAIN OUTCOME MEASURE

Electrodiagnosis.

MAIN RESULTS

In the first set of 105 patients, 75 had abnormal electrodiagnostic testing results. See Table 10-9.

For patients with a positive hand diagram result, the probability of an abnormal electrodiagnostic test increased from 79% to 92% when both the Tinel and Phalen test results were positive. Only 6 patients with a negative hand diagram result had an abnormal electrodiagnostic test result. Because the prevalence of an abnormal electrodiagnosis test result was so high, the posterior probability with a negative hand diagram result was still 33%. The second prospective phase of the study obtained posterior probabilities for an abnormal elec-

Table 10-9 Likelihood Ratios of Tinel and Phalen Signs and the Hand Diagram for Carpal Tunnel Syndrome

Test	Sensitivity	Specificity	LR+ (95% CI)	LR- (95% CI)
Tinel test	0.55	0.72	2.1 (1.2-4.0)	0.60 (0.44-0.88)
Hand diagram	0.92	0.40	1.5 (1.2-2.2)	0.20 (0.08-0.50)
Phalen test	0.72	0.53	1.5 (1.1-2.4)	0.52 (0.32-0.88)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

trodiagnostic test result similar to those obtained from the initial phase of the study.

CONCLUSIONS

LEVEL OF EVIDENCE Level 3.

STRENGTHS Prospective assessment of sequentially conducting the Tinel and Phalen tests for patients after a hand diagram test.

LIMITATIONS We infer that these patients were referred to the rheumatologist for therapeutic injections, accounting for the high prevalence of disease. The enrollment was not consecutive patients. It is not clear whether the electrodiagnosis was done by the same person who performed the clinical examination. The prevalence of disease was much higher in this study than in many other studies.

In a high-prevalence setting, the Phalen and Tinel tests will not demonstrate clinically important differences in the probability of disease. We infer that these patients are not representative of all patients with CTS symptoms. However, the data support the concept that the Phalen or Tinel test will not alter the information from a hand diagram in a clinically important fashion. The authors suggest that patients with a high probability of CTS could be offered treatment (injection therapy) without nerve conduction tests.

Reviewed by David L. Simel, MD, MHS

TITLE Lack of Utility of Semmes-Weinstein Monofilament Testing in Suspected Carpal Tunnel Syndrome.

AUTHORS Pagel KJ, Kaul MP, Dryden JD.

CITATION *Am J Phys Med Rehabil*. 2002;81(8):597-600.

QUESTION Do 2 types of testing with a monofilament among patients who have median nerve symptoms identify those who will have abnormal electrodiagnostic test results?

DESIGN Prospective, consecutive enrollment.

SETTING Electrodiagnostic laboratory of a Veterans Affairs hospital, Portland, Oregon.

PATIENTS All patients ($n = 113$) had paresthesias of the median nerve. Patients with a previous carpal tunnel release operation, stroke, paresthesias in the fourth and fifth fingers only, or neurologic disease were excluded.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

Two types of monofilament testing were done on the pad of each digit so that the filament bowed for 1.5 seconds. If the

monofilament was felt on at least 1 of 3 trials in each digital pad, the test result was considered normal. In the first protocol, the patient had an abnormal response if there was no sensation or a sensation only with an increased stimulus (>2.83 monofilament) in any of the radial 3 digits. In the second protocol, patients were considered to have an abnormal response only if abnormal findings in the third finger were associated with normal findings in the fifth finger. The examiners used a monofilament testing kit with various sizes of filaments. The reference test was a standard electrodiagnostic study, blinded to the monofilament results.

MAIN OUTCOME MEASURE

Abnormal electrodiagnosis studies.

MAIN RESULTS

Of 113 patients, 60 (53%) had abnormal electrodiagnostic testing results. See [Table 10-10](#).

CONCLUSIONS

LEVEL OF EVIDENCE Level 1.

STRENGTHS Evidence that the test (monofilament) and reference standard (electrodiagnosis) were applied independently. Clear guidelines on how to do the monofilament testing.

LIMITATIONS There was some selection bias in that not only were the patients all referred to the electrodiagnostic laboratory, but they were also evaluated to confirm that they had symptoms in the median nerve distribution. However, this is the appropriate population for whom carpal tunnel syndrome [CTS] ought to be correctly considered.

The authors conclude that the tests are worthless. Certainly, this appears true for the second method of monofilament testing (comparing the median nerve findings to the fifth finger). However, the ability of a normal response to monofilament testing in each of the first three digits decreases the likelihood of abnormal electrodiagnostic testing results in this population

Table 10-10 Likelihood Ratio of Monofilament Testing for Carpal Tunnel Syndrome

Test	Sensitivity	Specificity	LR+ (95% CI)	LR- (95% CI)
Decreased threshold or absent sensation in terminal digit pads 1, 2, or 3	0.98	0.15	1.2 (1.0-1.3)	0.11 (0.02-0.64)
Decreased threshold in terminal digit pad 3 with normal terminal digit pad 5	0.13	0.88	1.2 (0.45-3.1)	0.98 (0.84-1.1)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

of patients. Why might the results be different (ie, better) than what was reported in the original Rational Clinical Examination article? The study we initially used assessed only the response in the index finger rather than all 3 digits of the median nerve and found a sensitivity of only 59%.¹ Thus, requiring a normal response in all 3 digits would automatically improve the sensitivity. If the utility of a normal monofilament response can be validated, then this might be a useful test for identifying patients much less likely to have abnormal electrodiagnostic testing results. We would like to see this study repeated in a large population of patients with upper arm symptoms for whom CTS is considered.

REFERENCE FOR THE EVIDENCE

- Buch-Jaeger N, Foucher G. Correlation of clinical signs with nerve conduction tests in the diagnosis of carpal tunnel syndrome. *J Hand Surg [Br]*. 1994;19(6):720-724.

Reviewed by David L. Simel, MD, MHS

TITLE The Relationship Among Five Common Carpal Tunnel Syndrome Tests and the Severity of Carpal Tunnel Syndrome.

AUTHORS Priganc VW, Henry SM.

CITATION *J Hand Ther*. 2003;16(3):225-236.

QUESTION Among patients with carpal tunnel syndrome, do the diagnostic tests separate patients with mild, moderate, or severe electrodiagnostic results? Are the test results reliable during a 2- to 7-day period?

DESIGN Prospective. All tests were done before nerve conduction studies. The order of tests was randomized, except that the provocation tests were always done after the other maneuvers. The examiner waited 2 to 3 minutes between provocation tests for all the patients to return to baseline. Patients ($n = 27$) returned to the laboratory 2 to 7 days after the first test to assess reliability.

SETTING Patients referred from 3 neurology clinics in one community (Burlington, Vermont) for nerve conduction studies.

PATIENTS Patients scheduled for nerve conduction studies ($n = 206$) were contacted and invited to participate. Patients were excluded if they had systemic peripheral neuropathy, previous carpal tunnel release, proximal median nerve compression, or foot numbness not attributable to an orthopedic problem. Sixty-six patients (95 hands) were ultimately qualified for the study because the study reported only those with abnormal electrodiagnostic results.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

Phalen, Tinel, and carpal compression tests (examiners apply both of their thumbs to the patient's transverse carpal ligament), and Katz hand diagram. All patients had a nerve conduction test, along with a carpal tunnel outcomes assessment test that had scales for symptom severity and functional status. The tests were applied without knowledge of the electrodiagnostic results.

MAIN OUTCOME MEASURES

According to preestablished criteria, the nerve conduction quantitative results were classified into mild (55 hands), moderate (23 hands), or severe (17 hands) outcomes. Reliability was assessed during a 2- to 7-day follow-up period.

MAIN RESULTS

The Katz hand diagram was the most reliable finding (Table 10-11). The authors reported that only the Phalen test showed an association with the nerve conduction severity ($P < .05$). Our reanalysis of the data shows minimal significance ($P = .05$). In a logistic model, the odds ratio is 2.6 (95% confidence interval, 0.98-6.9) and the accuracy of the model as displayed by the area under the receiver operating characteristic curve is only 0.50 (a measure of accuracy).

Table 10-11 Reliability of Various Tests for Carpal Tunnel Syndrome

Test	κ (95% CI)
Katz hand diagram	0.95 (0.84-1.0)
Carpal compression	0.63 (0.33-0.92)
Phalen	0.58 (0.22-0.94)

Abbreviation: CI, confidence interval.

CONCLUSIONS

LEVEL OF EVIDENCE Level 2.

STRENGTHS A different type of study design to see whether the tests correlate with the degree of abnormality, rather than just the presence of carpal tunnel syndrome.

LIMITATIONS The results can be applied only to patients with known carpal tunnel syndrome. Thus, they are of limited value in the primary care clinic.

The goal of identifying patients who will have abnormal nerve conduction results differs from the goal of using the physical examination results to identify those who will have mild, moderate, or severely abnormal electrodiagnostic results. These results suggest that the physical examination findings did not help much with categorizing the severity. The intrarater reliability for these findings is reassuring in that the results are similar during a 2- to 7-day period.

Reviewed by David L. Simel, MD, MHS

CHAPTER 11

Does This Patient Have Abnormal Central Venous Pressure?

Deborah J. Cook, MD, FRCPC, MSc (Epid)

David L. Simel, MD, MHS

CLINICAL SCENARIO

A 65-year-old woman has had dyspnea for 2 months. She has had to give up her hobby of hiking and is now short of breath after climbing even 1 flight of stairs. Her dyspnea is sometimes worse at night. She has no chest pain, cough, or sputum, and the result of systems review is otherwise negative. On physical examination, her blood pressure is 135/90 mm Hg, and she has a regular cardiac rhythm at 72/min. You turn your attention to the jugular veins and next ask yourself, “Does this patient have abnormal central venous pressure (CVP)?”

WHY IS THIS QUESTION IMPORTANT?

Evaluation of the jugular venous pulse provides important information about pressure and other hemodynamic events in the right atrium.¹⁻³ The jugular venous pulse provides a useful estimate of CVP and thus the patient’s intravascular volume status. Inspection of the waveforms can assist the diagnosis of several tricuspid and pulmonic valvular abnormalities. Moreover, accurate assessment of CVP by physical examination may obviate the necessity for invasive hemodynamic monitoring.

Accordingly, the clinical evaluation of jugular venous pressure (JVP) and waveforms is useful whenever intravascular volume status, ventricular function, valvular disease, or pericardial constriction is in question. Proficiency in this examination is especially important, given that it may be difficult, if not impossible, to identify venous pulsation in patients with low CVP,⁴ in patients receiving mechanical ventilation,^{4,5} in patients with short or fat necks, and in some patients who have conditions causing wide swings in CVP during the respiratory cycle (eg, during acute asthma).

ANATOMIC AND PHYSIOLOGIC ORIGINS OF THE JUGULAR VENOUS PRESSURE

Because the jugular veins act as manometer tubes for the right atrium, they display changes in blood flow and pressure caused by right atrial filling, contraction, and emptying. In general, the jugular vein with the most distinct, undamped waveform is likely to most accurately reflect right atrial pressure. Because the right internal jugular vein is directly in line with the right atrium, thereby favoring an unimpeded transmission of atrial pulsations and pressure, it is the preferred site for examining the jugular venous pulse.

Direct measurements of CVP according to the left jugular veins tend to be higher than those on the right, but the correlation between the 2 is high.⁶ The discrepancy may reflect the fact that both the innominate vein and the left internal jugular vein can be compressed by a variety of normal or abnormal structures.

Although the internal jugular vein lies deep to the sternocleidomastoid muscle and may not always be visible as a discrete structure, its pulsation usually is transmitted to the overlying skin. Normally, the CVP pulsation moves toward the heart during inspiration because of a sudden increase in venous return to the right side of the heart.

The external jugular veins, although sometimes easier to see, may be constricted as they pass through the fascial planes of the neck and thus may not accurately reflect right atrial pressures. However, in one study, venous pressures measured in the external jugular vein accurately reflected right atrial pressures during anesthesia and with controlled or spontaneous ventilation.⁷ Positive-pressure ventilation caused regular, periodic changes to occur in venous return, which resulted in similar phasic changes in right atrial and external jugular pressures. The only significant difference was the greater right atrial pressure variation during mechanical ventilation, although the maximal venous pressures at the 2 sites were nearly identical.⁷

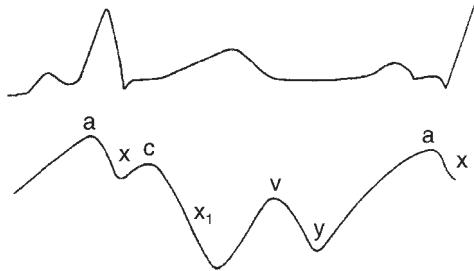


Figure 11-1 Venous Pulsation in the Neck Corresponds With the Electrocardiogram

Simultaneous recording of an electrocardiogram (top tracing) and jugular venous pressure waves (lower tracing). The *a* wave reflects right atrial contraction just before the first heart sound and carotid pulse; atrial relaxation is reflected by the *x* descent; *c* wave reflects the bulging of the tricuspid valve into the right atrium during ventricular isovolumetric contraction; *x₁* descent reflects subsequent atrial relaxation; *v* wave reflects the closure of tricuspid valve and subsequent distention of the right atrium; and *y* descent reflects the right atrium emptying after the opening of the tricuspid valve.

Table 11-1 Abnormalities of the Venous Waveforms

Waveform	Cardiac Condition
Absent <i>a</i> wave	Atrial fibrillation, sinus tachycardia
Flutter waves	Atrial flutter
Prominent <i>a</i> waves	First-degree atrioventricular block
Large <i>a</i> waves	Tricuspid stenosis, right atrial myxoma, pulmonary hypertension, pulmonic stenosis
Cannon <i>a</i> waves	Atrioventricular dissociation, ventricular tachycardia
Absent <i>x</i> descent	Tricuspid regurgitation
Prominent <i>x</i> descent	Conditions causing enlarged <i>a</i> waves
Large <i>cv</i> waves	Tricuspid regurgitation, constrictive pericarditis
Slow <i>y</i> descent	Tricuspid stenosis, right atrial myxoma
Rapid <i>y</i> descent	Constrictive pericarditis, severe right heart failure, tricuspid regurgitation, atrial septal defect
Absent <i>y</i> descent	Cardiac tamponade

Among critically ill patients, one group of investigators found jugular venous pulsations sufficiently obvious for examination only 20% of the time,⁸ whereas another group was able to estimate CVP in 84% of critically ill patients.⁴ In the former study, although external jugular pulsations were visible in all patients, clinicians' estimates of venous pressure according to physical examination were within 2 cm of CVP determined by central venous catheter only 47% of the time.

The evaluation of individual components of the venous pulse in health and disease lies outside the focus of this overview but can be summarized as follows.

ANALYSIS OF THE VENOUS WAVEFORM

The normal JVP reflects phasic pressure changes in the right atrium and consists of 3 positive waves and 3 negative troughs (Figure 11-1). Although these pressure changes can be recorded with pressure monitors, they are not always appreciable on clinical examination of the jugular pulse. Auscultation of the heart or simultaneous palpation of the left carotid artery may aid the examiner in relating the pattern of venous pulsations to the cardiac cycle.

Taken in sequence, right atrial contraction is reflected by the dominant positive *a* wave and occurs just before the first heart sound and carotid pulse. Atrial relaxation is reflected by the first negative trough, the *x* descent. The second positive wave is produced by the bulging of the tricuspid valve into the right atrium during ventricular isovolumetric contraction; this is called the *c* wave. Subsequent atrial relaxation creates the most dominant descent, the *x₁* descent. When the tricuspid valve closes, subsequent distention of the right atrium creates the *v* wave, which occurs just after the arterial pulse. Finally, after the opening of the tricuspid valve, the right atrium empties, resulting in the *y* descent.

Various cardiac conditions are associated with waveform abnormalities. A few of the most common include the absence of *a* waves in atrial fibrillation, large *cv* waves in tricuspid regurgitation, the slow *y* descent of tricuspid stenosis, and the brisk *y* descent seen in constrictive pericarditis. Table 11-1 shows a summary of abnormal venous waveforms and the conditions in which they occur. Remember, it is not always possible to see each of these waves and descents.

HOW TO EXAMINE THE NECK VEINS

The right internal jugular vein should be used to assess CVP for several reasons. It is in direct line with the right atrium, thereby favoring unimpeded transmission of atrial pulsations and pressure. Clinical assessment of CVP on the left may be marginally higher than that on the right. Finally, constricted or tortuous external jugular veins may introduce inaccuracy.

Positioning

Proper positioning is crucial for examination of the neck veins. The patient's head is supported to relax the neck muscles, and the trunk is inclined at an angle that brings the top of the column of

blood in the internal jugular vein to a level above the clavicle but below the angle of the jaw; in normal subjects, this positioning is accomplished at 30 to 45 degrees above the horizontal. In patients with elevated venous pressure, it often is necessary to elevate the trunk beyond 45 degrees, and patients with severe venous congestion may have to stand up and inspire deeply to bring the meniscus down into view. In some cases, the level of venous pulsation will be seen behind the angle of the jaw or will appear to move the earlobes. If the pressure in the internal jugular vein is high, venous pulsations will be lost in the completely full vein, and the high venous pressure may be overlooked.

Conversely, patients with low CVP may have to be positioned at 0 to 30 degrees. When CVP is low, the neck veins will be empty, and pulsations may not be visible even when the patient is horizontal.

Tangential light often improves the detection of the venous pulse. When ambient light is insufficient for this purpose, a pen-light, directed away from the examiner's eyes, may be useful.

Distinguishing Arterial (Carotid) From Venous (Jugular) Pulsation

Difficulty in distinguishing between the carotid arterial pulse and jugular venous pulse may be overcome by noting several differentiating features (Table 11-2).⁹ First, the venous pulsation is diffuse, usually has 2 waves, and the upward deflection is slow. In contrast, the carotid pulse is a fast, well-localized, single, outward deflection. Second, venous pulsations (unless the venous pressure is extremely high) diminish toward the clavicle or disappear beneath it as the patient sits up or stands and advance toward the angle of the jaw as the patient reclines; carotid pulses generally do not vary with position. Third, in the absence of intrathoracic disease, the top of the venous wave descends during inspiration (because of increasingly negative intrathoracic pressure). However, the visible carotid pulse does not vary with the respiratory cycle, except during pulsus paradoxus. Fourth, the JVP is non-palpable, and gentle pressure applied by the examiner's finger to the root of the neck above the clavicle will obstruct the vein, fill its distal segment, and obliterate the venous pulse. However, the carotid pulse is almost always palpable, usually striking the examining finger with considerable force. Finally, sustained pressure on the abdomen (the abdominojugular reflux test, to be described later) usually will cause even a normal venous pulse to increase briefly but will have no effect on the carotid pulse.

Estimation of Central Venous Pressure

The level of venous pressure is estimated by identifying the highest point of oscillation of the internal jugular vein (which usually occurs during the expiratory phase of respiration). This level must then be related to the middle of the right atrium, where venous pressure is, by convention, zero. Because the latter site is inaccessible on clinical examination, an accessible, reliable landmark is substituted: the sternal angle of Louis. This easily palpated landmark, found at the junction of the manubrium with the body of the sternum, lies 5 cm above the middle of the right atrium (for all practical purposes) in reclining patients of normal size and shape, regardless of the angle at which they are reclining.

Table 11-2 Distinguishing the Carotid Arterial From Jugular Venous Pulsation

Characteristic	Venous Pulse	Carotid Pulse
Waveform	Diffuse biphasic	Single sharp
Positional change	Varies with position	No variation
Respiratory variation	Height falls on inspiration	No variation
Effect of palpation	Wave nonpalpable, pressure obliterates pulse, vein fills	Pulse palpable, not compressible
Abdominal pressure	Displaces pulse upward	Pulse unchanged

Using the sternal angle as the reference point, the vertical distance (in centimeters) to the top of the jugular venous wave can be determined (Figure 11-2) and reported as the JVP; thus, JVP is 5 cm less than CVP.

When the patient is positioned at 45 degrees above the horizontal, the clavicle lies a vertical distance of about 2 cm above the sternal angle, and only CVPs of at least 7 cm will be observed.¹⁰ Because the normal CVP in adults is 5 cm, the top of their venous pressure column lies at their sternal angle, 2 cm below their lowest visible point in a patient at 45 degrees, and will only appear as the patient reclines toward the horizontal. The upper limit of normal for CVP is 9 cm H₂O, which produces a JVP extending 4 cm above the sternal angle.¹ (Note: The Update that follows this section revealed that physicians underestimate the value of the central venous pressure from the jugular vein meniscus. Part of the underestimate may result from variability in the depth measured from the sternal notch to the mid-right atrium. This can be partially corrected by accepting a JVP of 3 cm or more as elevated.)

Estimating CVP may be done as follows: Identify the highest point of pulsation in the internal jugular vein; find the sternal angle of Louis; from the sternal angle, measure the vertical distance to the top of the pulsation in centimeters; and report as "the JVP is xx cm."

Alternative methods of assessing CVP exist but have not been validated. For example, with a reclining patient, the clinician can inspect the veins of the back of the hand as the arm is slowly, passively raised; the level at which the veins collapse can then be related to the angle of Louis. This method may give false high readings with local obstruction and peripheral venous constriction, so it is not recommended.

Abnormal Central Venous Pressure

Elevated JVP reflects an increase in CVP. This increase can be due to increased right ventricular diastolic pressure (eg, right ventricular failure or infarction, pulmonary hypertension, or pulmonic stenosis), obstruction to right ventricular inflow (eg, tricuspid stenosis, right atrial myxoma, or constrictive pericarditis), hypervolemia, or superior vena cava obstruction.

Decreased JVP reflects a decreased or a low CVP. Low CVP may be due to intravascular volume depletion from gastrointestinal losses (vomiting or diarrhea), urinary losses (diuretics, uncontrolled diabetes mellitus, or diabetes insipidus), third-space fluid losses, and hypovolemic shock.

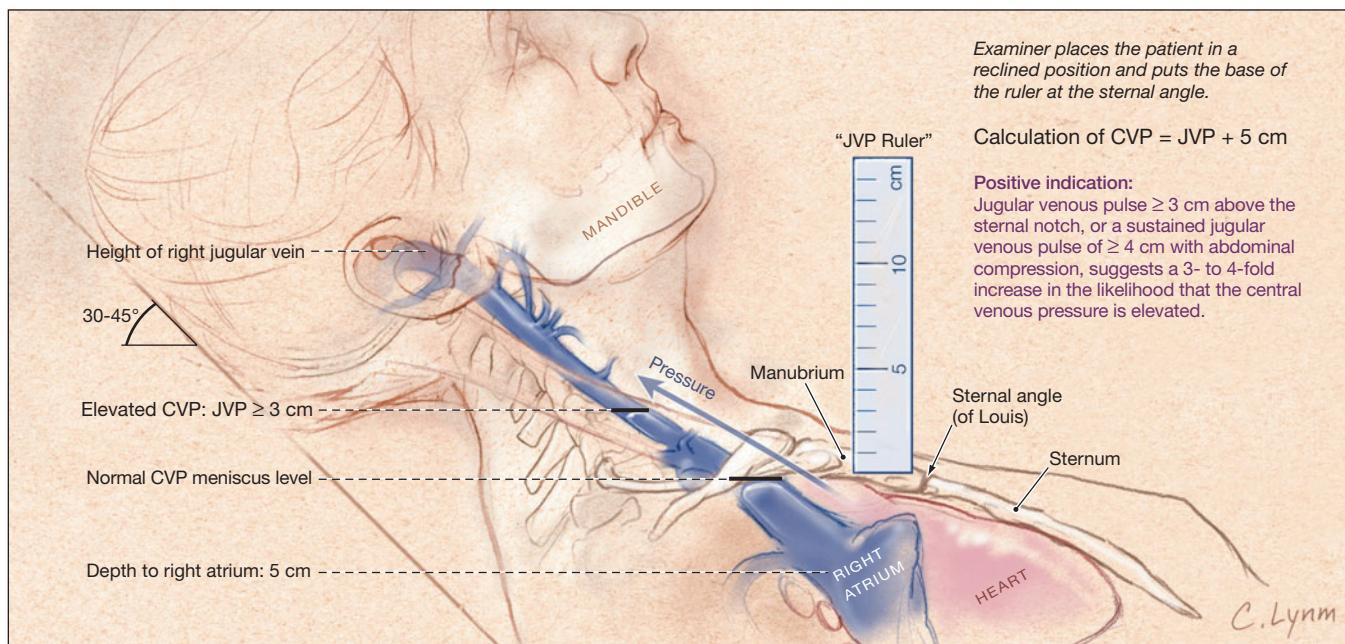


Figure 11-2 Estimation of Central Venous Pressure From the Jugular Venous Pulse

At any patient position, the top of the jugular vein meniscus is identified. The jugular venous pulse measurement is sighted from the height read from a ruler placed vertically over the sternal notch. The traditional assumption has been that the CVP is the JVP + 5 cm. However, the Update for this article showed that physicians tend to underestimate the CVP and the assumption of a 5-cm depth from the sternal notch to the right atrium is probably not valid. Thus, this figure has been updated to reflect current recommendations that a JVP ≥ 3 cm suggests an elevated CVP. Abbreviations: CVP, central venous pressure; JVP, jugular venous pressure.

Abdominojugular Reflux Test (Hepatojugular Reflux)

The abdominojugular reflux test consists of observing JVP before, during, and after abdominal compression. The increase in jugular pressure that follows abdominal compression is believed to be a consequence of blood shifting from abdominal veins into the right atrium. Pasteur first described the hepatojugular reflux in 1885.¹¹ Now, this bedside test is used to confirm the presence of right ventricular failure or reduced right ventricular compliance. Like all clinical tests, it is most reliable when performed in a standardized fashion.

The patient is instructed to relax and breathe normally through an open mouth (to avoid the false-positive increase in jugular pressure that accompanies the Valsalva maneuver). Firm pressure is then applied with the palm of the hand to the midabdomen for 15 to 30 seconds (abdominal compression for 1 minute, as has previously been described, is not required).^{10,12,13} This pressure should approximate 20 to 35 mm Hg when an unrolled bladder of a standard adult blood pressure cuff, partially inflated with 6 full bulb compressions, is placed between the examiner's hand and the patient's abdomen.^{10,13} Pressure directly over the liver, as was originally described,^{1,2,12,14} appears to be unnecessary.^{13,15} Therefore, designation of the test as abdominojugular reflux, rather than hepatojugular reflux is more appropriate. If pain is produced by this maneuver, or if the patient strains (Valsalva), the test becomes falsely positive. Either instruct the patient to open his or her mouth and breathe slowly or try a trial run, which is sometimes useful to demonstrate to the patient the force that will be applied over the abdomen.

Healthy individuals may exhibit one of 3 responses to abdominal compression: no change in JVP; a transient (few seconds) increase of more than 4 cm that returns to its former level or near the baseline before 10 seconds, with little or no decrease when abdominal pressure is released; or an increase of more than 3 cm sustained throughout compression.^{10,13} A positive abdominojugular test result occurs when abdominal compression causes a sustained increase in JVP of greater than or equal to 4 cm.

Kussmaul Sign

The JVP normally decreases during inspiration. The Kussmaul sign is the paradoxical increase in the height of JVP that occurs during inspiration. It can be explained by a heart that is unable to accommodate the increased venous return that accompanies the inspiratory decrease in intrathoracic pressure. Although classically described in constrictive pericarditis, the most common contemporary cause of the Kussmaul sign is severe right-sided heart failure, regardless of etiology. Other causes include myocardial restrictive disease such as amyloidosis, tricuspid stenosis, and superior vena cava syndrome.

PRECISION OF THE CLINICAL ASSESSMENT OF CENTRAL VENOUS PRESSURE

When 2 clinicians examine the same patient once (interobserver variation), and even when 1 clinician examines the same patient twice (intraobserver variation), estimates of CVP commonly vary by up to 7 cm.⁴ Final-year medical stu-

dents, first- and second-year medical residents, and attending physicians examined the same 50 intensive care unit patients (but were blinded to simultaneous CVP manometry) and estimated these patients' CVPs as low (<5 cm), normal (5–10 cm), or high (>10 cm).⁴ Agreement between students and residents was substantial (κ , a measure of chance-corrected agreement, was 0.65), agreement between students and attending physicians was moderate (κ = 0.56), and agreement between residents and attending physicians was modest (κ = 0.30).

Suggested causes for disagreement include variations in the positioning of patients, poor ambient lighting, difficulty in distinguishing carotid from venous pulsations, biological variation in CVP with the phases of respiration, and the effects of vasoactive medication and diuretics.

The precision of the abdominojugular reflux test has not been reported, but its results will vary with the force of abdominal compression. Different investigators suggest different forces: Ducas et al¹⁰ compressed a semi-inflated blood pressure cuff placed in the middle of the abdomen to 35 mm Hg (equivalent to a weight of approximately 8 kg), whereas Ewy¹³ applied a pressure of approximately 20 mm Hg.

Although no validated methods for improving precision in determining JVP have been reported, it seems prudent to standardize the procedure as described herein, encourage normal breathing, rehearse abdominal compression until the Valsalva maneuver is avoided, and gradually increase abdominal compression during a few seconds.¹⁶ Even when the Valsalva maneuver is avoided, there is still a small variation in JVP with the phases of breathing.¹⁷

ACCURACY OF THE CLINICAL ASSESSMENT OF CENTRAL VENOUS PRESSURE

We describe 3 studies that have reported the relation between clinical assessments of CVP and the gold standard of simultaneous pressure measurements through an indwelling central venous catheter.^{4,5,18} When the clinical assessment was reported as low, normal, or high, the pooled overall accuracy was 56%. In one study,⁴ venous pressure was assessed in each of 50 intensive care unit patients by one of 3 intensive care unit attending physicians, one of 6 medical residents, and one of 6 medical students. Although all groups tended to underestimate venous pressure, only the residents did so to a statistically significant degree. The correlation coefficient between clinical assessment and central line measured CVP was highest for medical students (0.74), slightly lower for residents (0.71), and lowest for staff physicians (0.65), and these correlations improved slightly when patients receiving mechanical ventilation were excluded. The students' data from this study⁴ (Table 11-3) display the results for 2 clinical questions: "Is the patient's true CVP low?" and "Is the patient's true CVP high?"² Despite small numbers of participants, it is apparent that a clinically assessed low CVP increases the likelihood by about 3-fold that the measured CVP will be low;

Table 11-3 Measured Central Venous Pressure^a

Is the CVP Low?			
Clinical Assessment	Low, CVP <5 cm	Normal or High, CVP >5 cm	LR That CVP Is Low (95% CI)
CVP low	3	5	3.4 (1.0-11)
CVP normal	4	22	1.0 (0.5-2.1)
CVP high	0	13	0 (0-1.5)
Is the CVP High?			
Clinical Assessment	High, CVP >10 cm	Normal or Low, CVP <5 cm	LR That CVP Is High (95% CI)
CVP high	10	3	4.1 (1.3-13)
CVP normal	10	16	0.8 (0.5-1.3)
CVP low	1	7	0.2 (0.02-1.3)

Abbreviations: CVP, central venous pressure; CI, confidence interval; LR, likelihood ratio.

^aAdapted from Cook.⁴

no patient clinically assessed as having a high CVP had a low measured CVP. Similar results hold when the clinician considers whether the patient has increased CVP. Clinical assessments of a high CVP increase the likelihood by about 4-fold that the measured CVP will be high; conversely, clinical assessments of a low CVP make the probability of finding a high measured CVP extremely unlikely (likelihood ratio [LR], 0.2). The data demonstrate that clinical assessments of a normal CVP are truly indeterminate, with LRs approaching 1; such estimates provide no information because they neither increase nor decrease the probability of an abnormal CVP.¹⁹ Aside from less observer variation, the data suggest that CVP estimates achieve greater accuracy among patients breathing spontaneously. However, the relatively small patient population creates an opportunity for further studies on how mechanical ventilatory assistance affects clinical assessment of CVP.

In a study of 62 patients undergoing right-sided heart catheterization,⁵ an attending physician, a critical care fellow, a medical resident, an intern, and a student each predicted whether 4 hemodynamic variables, including CVP, were low, normal, high, or very high. The sensitivity of the clinical examination for identifying low (<0 mm Hg), normal (0–7 mm Hg), or high (>7 mm Hg) CVP was 0.33, 0.33, and 0.49, respectively (10 cm of H₂O is equivalent to 7.5 mm Hg). The specificity of the clinical examination for identifying low, normal, or high CVP was 0.73, 0.62, and 0.76, respectively. Predictions of right atrial pressure (CVP) were more accurate in patients with low cardiac indexes (<2.2 L/min) and high pulmonary artery wedge pressures (>18 mm Hg) and less accurate among patients in coma or receiving mechanical ventilation. Accuracy was not improved in cases in which precision (interobserver agreement) among the examiners was high.

In a third study, Eisenberg et al¹⁸ compared clinical assessments with pulmonary artery catheter readings in 97 critically ill patients. The physicians caring for these patients were asked to predict whether CVP was less than

2, 2 through 6, or greater than 6 mm Hg; whether cardiac output was less than 4.5, 4.5 through 7.5, or greater than 7.5 L/min; whether systemic vascular resistance was 1100, 1100 through 1300, or greater than 1300 ($\text{dyn} \times \text{s}$)/ cm^5 ; and whether pulmonary artery wedge pressure was less than 10, 10 through 14, 15 through 19, or greater than or equal to 20 mm Hg. Physicians correctly predicted the patients' CVP only 55% of the time and cardiac index, systemic vascular resistance, and pulmonary artery wedge pressure only 51%, 44%, and 30% of the time, respectively. CVP was more frequently underestimated (27%) than overestimated (17%).

Although the abdominojugular reflux test is an insensitive way to diagnose congestive heart failure, the specificity of this test is high.^{20,21} Moreover, the positive LRs (6.4 when the strict criteria are used and 6.0 when emergency physician judgment is used) indicate that this is a useful bedside test (Table 11-4).

IMPROVING CLINICAL EXAMINATION OF THE JUGULAR VEINS

Examining patients with indwelling central venous catheters provides the clinician with an opportunity for calibrating and periodically testing clinical skills for evaluating CVP. Of course, the examination should be performed blind to the catheter reading. If the examination is also conducted blind to other patient data, interpretation of waveforms can be compared to electrocardiograms and other data from cardiac investigations. Learning aids such as pocket cards displaying the normal jugular pulsations may also be helpful. Assessment of JVP in patients with tachycardia, irregular cardiac rhythms, and rapid and deep respirations and those requiring mechanical ventilation provide a challenge for even seasoned clinicians.²²

Table 11-4 Sensitivity and Specificity of the Abdominojugular Reflux in Diagnosing Congestive Heart Failure^a

Abdominojugular Reflux	CHF	No CHF	Total
By Explicit Criteria for the Abdominojugular Reflux Response			
Present	5	1	6
Absent	16	26	42
Total	21	27	48
By Emergency Physician's Judgment of the Abdominojugular Reflux Response			
Present	4	2	6
Absent	8	34	42
Total	12	36	48

Abbreviations: CHF, congestive heart failure; CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

^aAdapted from Marantz et al.²⁰ For diagnosing by criteria, sensitivity = 0.24; specificity = 0.96; LR+ = 6.4 (95% CI, 0.8-51); and LR- = 0.8 (95% CI, 0.6-1.0). For diagnosing by emergency physicians, sensitivity = 0.33; specificity = 0.94; LR+ = 6.0 (95% CI, 1.3-29); and LR- = 0.7 (95% CI, 0.5-1.1).

THE BOTTOM LINE

According to the results of this overview, the following recommendations apply to the clinical assessment of JVP. First, in a well-lit room, position the patient at an angle such that the meniscus of blood in the right jugular vein is brought into vision (usually an angle of 30 to 45 degrees from the horizontal). To identify the top of the meniscus, it may be necessary to raise or lower this angle. Second, distinguish the jugular venous waveform from the carotid pulsation by remembering the following: The venous waveform is diffuse and biphasic, varies with position and respiration, is nonpalpable, and may be displaced upward during abdominal pressure. In contrast, the carotid pulsation is single, sharp, and palpable; does not vary with position or respiration; and is unchanged with abdominal pressure. Third, measure the vertical distance in centimeters from the sternal angle of Louis to the top of the column of blood in the jugular vein. The upper limit of normal is approximately 4 cm (Note: The Update to this article recommends that the clinician consider a value of 3 cm or more as elevated). Armed with evidence about how to examine and interpret the clinical assessment of CVP, you can now answer the question of whether the patient presented at the beginning of this article, and subsequent patients you care for, have abnormal CVP.

Author Affiliations at the Time of the Original Publication

Departments of Medicine and Clinical Epidemiology and Biostatistics, McMaster University Faculty of Health Sciences, Hamilton, Ontario, Canada, and the Ontario Ministry of Health, Toronto, Ontario, Canada (Dr Cook); The Center for Health Services Research in Primary Care, Durham Veterans Affairs Medical Center, Duke University Medical Center, Durham, North Carolina (Dr Simel).

Acknowledgments

The authors thank Ernest Fallen, MD, FRCPC, and David Sackett, MD, FRCPC, for their thoughtful review and helpful comments on the manuscript.

REFERENCES

- Braunwald E. The physical examination. In: Braunwald E, ed. *Heart Disease: A Textbook of Cardiovascular Medicine*. Philadelphia, PA: WB Saunders; 1988:13-40.
- O'Rourke RA. The measurement of systemic blood pressure, normal and abnormal pulsations of the arteries and veins. In: Hurst JW, Schlant RC, eds. *The Heart, Arteries and Veins*. New York, NY: McGraw Hill; 1989: 157-162.
- Davison R, Cannon R. Estimation of central venous pressure by examination of the jugular veins. *Am Heart J*. 1974;87(3):279-282.
- Cook DJ. The clinical assessment of central venous pressure. *Am J Med Sci*. 1990;299(3):175-178.
- Connors AF, McCaffree DR, Gray BA. Evaluation of right heart catheterization in the critically ill patient without acute myocardial infarction. *N Engl J Med*. 1983;308(5):263-267.
- Briscoe CE. A comparison of jugular and central venous pressure measurements during anesthesia. *Br J Anaesth*. 1973;45(2):173-178.
- Stoelting RK. Evaluation of external jugular venous pressure as a reflection of right atrial pressure. *Anesthesiology*. 1973;38(3):291-294.
- Davison R, Cannon R. Estimation of central venous pressure by examination of jugular veins. *Am Heart J*. 1974;87(3):279-282.

9. Matthews MB. The cardiovascular system. In: MacLeod J, ed. *Clinical Examination*. Edinburgh, Scotland: Churchill Livingstone; 1982:99-146.
10. Ducas J, Magder S, McGregor M. Validity of the hepatojugular reflux as a clinical test for congestive heart failure. *Am J Cardiol*. 1983;52(10):1299-1303.
11. Pasteur W. Note on a new physical sign of tricuspid regurgitation. *Lancet*. 1885;2:524.
12. Cohn JN. Central venous pressure as a guide to volume expansion. *Ann Intern Med*. 1967;66(6):1283-1287.
13. Ewy GA. The abdominojugular test: technique and hemodynamic correlates. *Ann Intern Med*. 1988;109(6):456-460.
14. Constant J, Lippischutz EJ. The one minute abdominal compression test or "the hepatojugular reflux," a useful bedside test. *Am Heart J*. 1964;67:701-707.
15. Singh SP, Haider R. The lack of value of hepato-jugular reflux as a sign of heart failure. *Postgrad Med J*. 1973;49(567):10-13.
16. Burch GE, Ray CT. Mechanism of the hepatojugular reflux test in congestive heart failure. *Am Heart J*. 1954;48(3):373-382.
17. Matthews MB, Hampson J. Hepatojugular reflux. *Lancet*. 1958;1(7026):873-876.
18. Eisenberg PR, Jaffe AS, Schuster DP. Clinical evaluation compared to pulmonary artery catheterization in the hemodynamic assessment of critically ill patients. *Crit Care Med*. 1984;12(7):549-553.
19. Simel DL, DeLong ER, Feussner JR, Matchar DB. Intermediate, indeterminate, and uninterpretable diagnostic test results. *Med Decis Making*. 1987;7(2):107-114.
20. Marantz PR, Kaplan MC, Alderman MH. Clinical diagnosis of congestive heart failure in patients with acute dyspnea. *Chest*. 1990;97(4):776-781.
21. Maisel AS, Atwood JE, Goldberger AL. Hepatojugular reflux: useful in the bedside diagnosis of tricuspid regurgitation. *Ann Intern Med*. 1984;101(6):781-782.
22. Sapira JD. Veins. In: Sapira JD, ed. *The Art and Science of Bedside Diagnosis*. Baltimore, MD: Williams & Wilkins; 1990:355-370.

This page intentionally left blank

Prepared by David L. Simel, MD, MHS

Reviewed by Deborah Cook, MD, MSc (Epid)

CLINICAL SCENARIO

A 48-year-old man who has had 2 myocardial infarctions is having trouble sleeping. He claims shortness of breath while supine but does not notice any ankle edema. The lungs are clear, whereas the cardiac evaluation reveals an S₄ but no S₃ heart sound. There is a short systolic murmur along the left sternal border. He has no peripheral edema. You look at his large, thick neck and have no confidence that you will be able to assess the neck veins.

UPDATED SUMMARY ON ABNORMAL CENTRAL VENOUS PRESSURE

Original Review

Cook DJ, Simel DL. Does this patient have abnormal central venous pressure? *JAMA*. 1996;275(8):630-634.

UPDATED LITERATURE SEARCH

Our literature search used the parent search strategy for The Rational Clinical Examination series articles in MEDLINE, combined with the search terms “central venous pressure,” “exp jugular veins,” “exp venous pressure,” and “abdomino-jugular reflux,” limited to human and English-language articles published from 1995 to August 2004. We excluded case reports, leaving 189 titles for review. Of these citations, 13 were applicable and were retrieved to determine whether they had sensitivity, specificity, or likelihood ratio (LR) data for the use of the clinical estimation of the central venous pressure (CVP) or jugular venous pressure (JVP) for identifying patients with high or low CVP measured by a reference standard. Only 1 study provided new data. Through review of references in the 13 articles, we found 1 article published before 1995 that we had not included in our original review.

NEW FINDINGS

- A JVP 3 cm above the sternal angle, in any patient position, suggests an elevated CVP.¹
- Clinicians systematically underestimate the CVP when using the JVP.² The distance from the sternal notch to the

right atrium may be closer to 8 cm rather than the traditionally assumed value of 5 cm.¹

Details of the Update

Two analyses from the same randomized treatment trial for heart failure demonstrate the usefulness of assessing for an elevated JVP.^{3,4} These studies analyzed prospectively collected data by study investigators (cardiologists) from a few thousand patients. The cardiologists answered a simple question: Is the JVP elevated? The assessment was not confirmed with direct measurement of the CVP, but the association with important outcomes suggests that assessing the CVP as elevated or not elevated is useful.

A nonsystematic review of the venous pressure provides additional information for those who believe that the assessment of the JVP is either too difficult or lacks correlation with the CVP. McGee⁵ describes many features that explain the discrepancy between the clinical estimation of CVP and the actual CVP measurement. What is striking about McGee's³ findings and those from empirical studies (see reviews of individual studies) is that the discrepancies are not random, but systematic. There is a distinct and reproducible bias that leads clinicians to underestimate the true CVP. The editorial accompanying McGee's³ review asserts that the “... major limitation to current use [of JVP assessment] is lack of practice,”⁶ a statement also emphasized by others.⁷ McGee⁵ suggests, despite the factors leading to disagreements between the clinical assessment of JVP and the measurement of CVP, that finding the JVP more than 3 cm above the sternal angle indicates an abnormally high CVP. The empirical data support the suggestion.

The information about the sensitivity and specificity of the JVP assessment to identify patients with low CVP is scant. We found no additional studies. The original data in The Rational Clinical Examination article suggest an LR of 3.4 (95% confidence interval, 1-9.9) when the clinical question is whether the patient has a low CVP and the clinician finds that the JVP is not observed (CVP < 5 cm). The correlation between low clinically assessed CVP with the invasive assessment is better than it is for a population of patients with high CVP. Although this makes sense, our confidence around this is low, given the small numbers of patients studied.

The abdominojugular reflux test might be an alternative or complementary test to the JVP assessment in patients with significantly impaired left ventricular failure. A 2000 systematic review⁸ identified no additional studies evaluating abdominojugular reflux. We found no other studies from 1995 to 2004, although we did identify a study that we had not included in the original review. In patients with impaired left ventricular function, the reproducibility of the abdominojugular reflux appears to be excellent ($\kappa = 0.92$). We caution examiners that they must use same techniques as those used in these studies to achieve similar results.⁹

IMPROVEMENTS IN THE DATA PRESENTED IN THE ORIGINAL PUBLICATION

We used the data from the original manuscript and now provide summary estimates for high CVP; we can also provide summary estimates for the abdominojugular reflux (Table 11-5). We created a new figure to demonstrate the assessment of JVP, indicating the newer recommended threshold of ≥ 3 cm for identifying patients with an elevated central venous pressure.

CHANGES IN THE REFERENCE STANDARD

There have been no changes in the reference standard.

RESULTS OF LITERATURE REVIEW

Three studies allow us to combine measures for the assessment of a high CVP. Two of the 3 studies evaluated

Table 11-5 Likelihood Ratios for the Abdominojugular Reflux Test and Clinical Assessments of the Central Venous Pressure

Finding (No. of Combined Studies)	Question	LR+ (95% CI)	LR- (95% CI)
Abdominojugular reflux (2) ^{9,10a}	Would the measured CVP be high?	4.4 (1.8-10)	0.48 (0.22-1.1)
Clinically assessed high CVP (3) ^{9,11,12b}	Would the measured CVP be high?	3.1 (1.6-6.0)	0.50 (0.37-0.68)
Clinically assessed low CVP (1) ¹¹	Would the measured CVP be low?	3.4 (1-9.9)	0.65 (0.28-1.2)

Abbreviations: CI, confidence interval; CVP, central venous pressure; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

^aData are homogenous, with $P = .62$ for LR+, but heterogeneous for LR- ($P < .01$). The populations of patients were different. One study was of patients undergoing an evaluation for cardiac transplantation for left ventricular systolic dysfunction. The other study assessed patients with acute dyspnea.

^bData are homogenous, with $P = .22$ for LR+ and .31 for LR-.

patients with advanced heart failure (ie, at least New York Heart Association class III), whereas 1 included critically ill patients in the intensive care unit. These data capture an important population for whom the finding would be of interest. We do not know how well the results apply to less severely ill patients treated in a primary care clinic. We agree with advocates who suggest that the clinical assessment of the JVP is useful, but we also agree with those who suggest that clinicians need more practice to become proficient.

From these few studies, it is difficult to know whether clinicians, on balance, are better at identifying patients with a low vs a high CVP. However, our confidence in the accuracy of assessing a low CVP is only modest, given the broad confidence intervals.

EVIDENCE FROM GUIDELINES

The Scottish Intercollegiate Guidelines Network recommends using the JVP to help diagnose left ventricular systolic function and to identify patients who need diuretics.¹³ The US Department of Veteran Affairs recommends assessing for jugular venous distention in hypertensive patients.¹⁴

CLINICAL SCENARIO—RESOLUTION

For a variety of reasons such as self-confidence in CVP assessment or patient-specific anatomy such as large, thick necks, primary care providers often assume they will be unable to identify the JVP. The sense that you will not be able to visualize the veins may be accurate, although reinforced from using poor examining technique. We suggest that clinicians reassess their performance by making sure that they are using the proper examining technique. One study of heart failure patients suggests that the abdominojugular reflux evaluation has excellent reproducibility. It is possible that it is easier to see sustained inducible jugular venous distention than the normal venous pulse wave, especially in patients with large, thick necks. At the least, using abdominal pressure to help identify the course of the internal jugular vein might improve technique and ability to identify normal venous pulse waves.

A JVP more than 3 cm above the sternal notch, or a sustained JVP of 4 cm or more with abdominal compression, suggests a 3- to 4-fold increase in the likelihood that the CVP is elevated.

CENTRAL VENOUS PRESSURE—MAKE THE DIAGNOSIS

PRIOR PROBABILITY

Estimating the prior probability for an elevated CVP among patients with a low ejection fraction depends on the patient's underlying condition and the effectiveness of treatment. Current treatment regimens that now include β -blockers and angiotensin-converting enzyme inhibitors may decrease the prevalence of volume overload low in patients with a reduced ejection fraction. In the Studies of Left Ventricular Dysfunction, investigators determined clinically that approximately 10% of patients with a left ventricular ejection fraction 35% or less at baseline had an elevated CVP.^{3,4} Although the CVP was not invasively measured, we know that clinical assessments of a high CVP typically underestimated the true value. As a starting point, the range 10% to 20% seems like a reasonable estimate for elevated CVP among patients previously diagnosed as having a low ejection fraction.

The close relationship between low CVP and underlying disease makes it impossible to come up with a generally useful starting point for a pretest probability about CVP, so clinicians must use their own judgment individualized for the patient.

POPULATION FOR WHOM AN ABNORMAL CENTRAL VENOUS PRESSURE SHOULD BE CONSIDERED

- Patients with a low left ventricular ejection fraction are at risk for a high CVP or a low CVP (eg, overdiuresis).
- Patients with underlying acute clinical conditions that lead to volume loss may have a low CVP.

DETECTING THE LIKELIHOOD OF AN ABNORMAL CENTRAL VENOUS PRESSURE

See Table 11-6.

Table 11-6 Likelihood of Abnormal Central Venous Pressure

Finding	LR+ (95% CI)	LR- (95% CI)
Would the Measured CVP Be High?		
Abdominojugular reflux (n = 2) ^a	4.4 (1.8-10)	0.48 (0.22-1.1)
Would the Measured CVP Be High?		
Clinically assessed high CVP from the JVP ^b	3.1 (1.6-6.0)	0.50 (0.37-0.68)
Would the Measured CVP Be Low?		
Clinically assessed low CVP from the JVP ^c	3.4 (1-9.9)	0.65 (0.28-1.2)

Abbreviations: CI, confidence interval; CVP, central venous pressure; JVP, jugular venous pulse; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

^aA positive test result is a sustained increase in the JVP of 4 cm or more with 10 seconds of abdominal compression, followed by an abrupt decrease with the release of pressure.

^bUse these values when the clinical question is, "Would my patient have a high CVP on invasive measurement?" A positive result suggesting a high CVP is a JVP more than 3 cm above the sternal angle.

^cUse these values when the clinical question is, "Would my patient have a low CVP on invasive measurement?" When the meniscus of the JVP is not observed, the result is positive and suggests a low CVP. However, the CI around these estimates is broad.

REFERENCE STANDARD TESTS

Invasive measurement of the CVP with an internal monitor. To minimize variation, zero the manometer to a line representing the intersection of a cross-sectional plane through the fourth intercostal space and a coronal plane between the back and xiphoid process.⁵

REFERENCES FOR THE UPDATE

- Seth R, Magner P, Matziiner F, van Walraven C. How far is the sternal angle from the mid-right atrium? *J Gen Intern Med*. 2002;17(11):861-865.^a
- Stein JH, Neumann A, Marcus RH. Comparison of estimates of right atrial pressure by physical examination and echocardiography in patients with congestive heart failure and reasons for discrepancies. *Am J Cardiol*. 1997;80(12):1615-1618.^a
- Drazner MH, Rame JE, Stevenson LW, Dries DL. Prognostic importance of elevated jugular venous pressure and a third heart sound in patients with heart failure. *N Engl J Med*. 2001;345(8):574-581.
- Drazner MH, Rame JE, Dries DL. Third heart sounds and elevated jugular venous pressure as markers of the subsequent development of heart failure in patients with asymptomatic left ventricular dysfunction. *Am J Med*. 2003;114(6):431-437.
- McGee SR. Physical examination of venous pressure: a critical review. *Am Heart J*. 1998;136(1):10-18.
- Economides E, Stevenson LW. The jugular veins: knowing enough to look. *Am Heart J*. 1998;136(1):6-9.
- Van't Laar A. Why is the measurement of jugular venous pressure discredited? *Neth J Med*. 2003;61(7):268-272.
- Wiese J. The abdominojugular reflux sign. *Am J Med*. 2000;109(1):59-61.
- Butman SM, Ewy GA, Standen JR, Kern KB, Hahn E. Bedside cardiovascular examination in patients with severe chronic heart failure: importance of rest or inducible jugular venous distention. *J Am Coll Cardiol*. 1993;22(4):968-974.^a
- Marantz PR, Kaplan MC, Alderman MH. Clinical diagnosis of congestive heart failure in patients with acute dyspnea. *Chest*. 1990;97(4):776-781.
- Cook DJ. The clinical assessment of central venous pressure. *Am J Med Sci*. 1990;299(3):175-178.
- Stein JH, Neumann A, Marcus RH. Comparison of estimates of right atrial pressure by physical examination and echocardiography in patients with congestive heart failure and reasons for discrepancies. *Am J Cardiol*. 1997;80(12):1615-1618.
- Diagnosis and Treatment of Heart Failure Due to Left Ventricular Systolic Dysfunction*. Edinburgh, Scotland: Scottish Intercollegiate Guidelines Network; 1999. SIGN Publication No. 35.
- Veterans Administration, Department of Defense. *VA/DoD Clinical Practice Guideline for Diagnosis and Management of Hypertension in the Primary Care Setting*. Washington, DC: Veterans Administration, Department of Defense; 2004. <http://www.oqp.med.va.gov/cpg/HTN04/G/HTN%20Sum508.pdf>. Accessed May 23, 2008.

^aFor the Evidence to Support the Update for this topic, see <http://www.JAMAevidence.com>.

This page intentionally left blank

EVIDENCE TO SUPPORT THE UPDATE:

11

Central Venous Pressure

TITLE Bedside Cardiovascular Examination in Patients With Severe Chronic Heart Failure: Importance of Rest or Inducible Jugular Venous Distention.

AUTHORS Butman SM, Ewy GA, Standen JR, Kern KB, Hahn E.

CITATION *J Am Coll Cardiol.* 1993;22(4):968-974.

QUESTION Do a variety of clinical findings predict cardiac hemodynamics in a group of patients with advanced chronic congestive heart failure?

DESIGN Prospective, convenience sample. Some of the patients (52%) had an examination by a second observer to assess precision.

SETTING Cardiac catheterization laboratory, Tucson, Arizona.

PATIENTS Fifty-two patients under evaluation for possible heart transplantation and who were undergoing right-sided heart catheterization within 24 hours of the physical examination.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

Abdominojugular reflux—a positive test result was defined as 4 cm or more sustained elevation of the jugular venous pulse (JVP) with 10 seconds of abdominal compression that disappeared abruptly with the release of abdominal pressure.

JVP was considered abnormal and elevated if pulsations were seen while the patient was elevated at 45 degrees from horizontal, or if the estimated pressure was greater than 7 cm. The reference standard was the pulmonary capillary wedge pressure (>18 mm Hg) was considered abnormal, indicating volume overload).

MAIN OUTCOME MEASURES

Sensitivity, specificity, and κ for the physical examination findings.

Table 11-7 Likelihood Ratio for Jugular Venous Pressure and Abdominojugular Reflux for an Elevated Central Venous Pressure

Test	Sensitivity	Specificity	LR+ (95% CI)	LR- (95% CI)
Jugular venous pressure	0.57	0.93	8.5 (1.8-49)	0.46 (0.60-0.69)
Abdomino-jugular reflux	0.81	0.80	4.0 (1.8-12)	0.24 (0.11-0.47)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

MAIN RESULTS

See Table 11-7. The agreement for the presence of JVP elevation was good ($\kappa = 0.69$) but even better for abdominojugular reflux ($\kappa = 0.92$). These patients were mostly men and had a low ejection fraction (mean ejection fraction, 18%; range, 6%-39%).

CONCLUSIONS

LEVEL OF EVIDENCE Level 3.

STRENGTHS Precision was determined. The clinicians judged the JVP as abnormal or not. With the patient at 45 degrees from horizontal, the clinicians recorded an elevated central venous pressure (CVP) when they could visualize the jugular vein contours.

LIMITATIONS The pulmonary capillary wedge pressure served as the reference standard rather than the CVP (the wedge pressure is a better indicator of volume status). Small sample size in a select group of patients.

There must have been expectation bias in that most clinicians would have expected these severely affected patients to have volume overload and abnormal physical findings. This should have led to an overestimate of sensitivity and an underestimate of specificity (because more patients would have been expected to be in the first row of the 2×2 table). The results are consistent with those found for estimated CVP greater than 10 cm by Stein et al¹ in a similar population of patients with advanced heart failure. An intriguing finding is that every patient judged to have an elevated JVP also had

an abnormal abdominojugular reflux. The gain in sensitivity from the abdominojugular reflux assessment vs the JVP was offset by the loss of specificity; these tests performed similarly in this population.

REFERENCE FOR THE EVIDENCE

- Stein JR, Neumann A, Marcus RH. Comparison of estimates of right atrial pressure by physical examination and echocardiography in patients with congestive heart failure and reasons for discrepancies. *Am J Cardiol*. 1997;80(12):1615-1618.

Reviewed by David L. Simel, MD, MHS

range, 4.6-6.1 cm). However, when the patient was at 90 degrees, the median distance was 8.3 cm (interquartile range, 7-9.6 cm). Between 30- and 60-degree elevation (the elevation typically used in clinical assessments), the median calculated distance was approximately 8 cm.

CONCLUSIONS

LEVEL OF EVIDENCE Not a diagnostic test study.

STRENGTHS Large sample, asking an important question about the assumptions necessary for the clinical examination.

LIMITATIONS The authors had to make their own assumption about the position of the mid-right atrium. The CT scans were done in a population of patients primarily with lung or thoracic disease (eg, carcinoma).

This is a clever and basic study to test an assumption underlying the clinical examination. The decision to add 5 cm to the estimation of the jugular venous pressure (JVP) makes sense when the patient is supine. However, clinicians almost never assess the JVP in the supine patient. The authors found a median distance of 8 cm in positions typically used during the clinical examination. Thus, clinicians using the JVP would underestimate the central venous pressure (CVP) by 3 cm. The implication of this is that any patient with a JVP 3 cm above the horizontal should be considered as having a high CVP because the likely CVP will be more than 10 cm. This recommendation is consistent with data from the Stein et al¹ study that found the JVP leads to underestimates of around 5 cm for patients with elevated CVP.

REFERENCE FOR THE EVIDENCE

- Stein JR, Neumann A, Marcus RH. Comparison of estimates of right atrial pressure by physical examination and echocardiography in patients with congestive heart failure and reasons for discrepancies. *Am J Cardiol*. 1997;80(12):1615-1618.

Reviewed by David L. Simel, MD, MHS

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

High-speed chest computed tomography (CT) scans were performed on patients while they were in the lateral supine and 90-degree positions during an end-inspiratory breath hold. The authors assumed that the mid-right atrium was 2 cm below the superior vena cava and right atrial junction.

MAIN OUTCOME MEASURE

Measured sternal angle distance for supine and 90-degree positions. The investigators used geometric calculations to determine the distance between the sternal angle at 30 degrees, 45 degrees, and 60 degrees.

MAIN RESULTS

With the patient supine, the median distance between the sternal angle and right atrium was 5.4 cm (interquartile

TITLE Comparison of Estimates of Right Atrial Pressure by Physical Examination and Echocardiography in Patients With Congestive Heart Failure and Reasons for Discrepancies.

AUTHORS Stein JH, Neumann A, Marcus RH.

CITATION *Am J Cardiol.* 1997;80(12):1615-1618.

QUESTION Among patients with severe heart congestive heart failure, how closely do cardiologists predict the central venous pressure?

DESIGN Consecutive, prospective.

SETTING Cardiac catheterization laboratory, Chicago, Illinois.

PATIENTS Twenty-two patients with an average left ventricular ejection fraction of 19% (range, 12%-29%).

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

The central venous pressure (CVP) was measured from the jugular venous pressure (JVP) by identifying the peak JVP. A centimeter ruler placed vertically on the sternal angle was used to measure the distance at an intersection with a horizontal straight edge placed at the JVP. To estimate the CVP, 5 cm was added to the vertical distance. A right-sided heart catheterization was performed immediately thereafter.

MAIN OUTCOME MEASURE

Correlation between clinical estimate of the CVP and the invasive measurement. The data are displayed in a scatterplot, so that lines can be drawn to extract the raw results.

MAIN RESULTS

The correlation between the raw clinical estimate and the invasive measure was 0.92. The clinical estimates systematically underestimated the actual value. The bias was least for those with clinical estimates of less than 8 (correlation was near perfect), but the underestimation became more pronounced as the clinician estimated a higher CVP from the JVP. With estimates of 9 to 14 cm, the clinicians underestimated the true CVP by 5.0 cm. Dichotomizing the data for a clinical estimate of 8 cm or more and extracting the results from the scatterplot reveals the likelihood ratios (LRs) in Table 11-8.

Table 11-8 Likelihood Ratio of Clinically Estimated CVP

CVP	Invasive CVP	LR+ (95% CI)	LR- (95% CI)
Clinical estimate of CVP > 10 cm	CVP ≥ 10 cm	11 (0.73-157)	0.25 (0.10-0.64)
Clinical estimate CVP > 8 cm	CVP ≥ 8 cm	1.6 (0.98-3.7)	0.18 (0.03-1.1)

Abbreviations: CI, confidence interval; CVP, central venous pressure; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

CONCLUSIONS

LEVEL OF EVIDENCE Level 3.

STRENGTHS Objective reference standard done immediately after the CVS was assessed.

LIMITATIONS Small population of patients in a narrow spectrum of disease. The examiners were specialists.

These data are most useful for validating the concept that clinical assessment of the CVP, by measuring the vertical distance to the JVP, and then adding 5 cm, will systematically underestimate the true pressure. Because the population of patients was small, the confidence intervals around the estimates using a cut point of 10 are huge for the positive LR. Every patient with a clinical estimate of more than 10 cm CVP had the result confirmed by the invasive test. However, it seems likely that cardiologists estimating a high pressure (>10 cm) in a population of patients with low ejection fractions are usually going to be correct. It becomes much more difficult to identify patients with volume overload when lower clinical and invasive thresholds are used. Reviews of clinical assessments of CVP recommend that clinicians use a clinical estimate of 8 cm as their threshold for assessing a high pressure.

Reviewed by David L. Simel, MD, MHS

This page intentionally left blank

CHAPTER 12

Does This Patient Have Acute Cholecystitis?

Robert L. Trowbridge, MD

Nicole K. Rutkowski, MD

Kaveh G. Shojania, MD

CLINICAL SCENARIO

A 72-year-old woman with poorly controlled diabetes, coronary artery disease, and hypertension presents to the emergency department complaining of nausea and vomiting. As an emergency department resident, you elicit the history that the patient felt well until 24 hours ago, when she developed anorexia, followed rapidly by bilious emesis. She describes mild upper abdominal discomfort but is unable to further localize the pain. There have been no abnormal bowel movements, gastrointestinal bleeding, or chest pain.

The patient is febrile (39°C) and appears uncomfortable. Her lungs are clear, and cardiac examination reveals only a fourth heart sound. There is moderate epigastric tenderness and guarding throughout the abdomen, but no rigidity. Pelvic and rectal examination results are unremarkable. Electrocardiography shows no changes suggestive of ischemia. Laboratory testing shows a leukocytosis level of $17500 \times 10^3/\mu\text{L}$, serum transaminase levels twice the upper limit of normal, and a total bilirubin level of 3.2 mg/dL. In considering the differential diagnosis for the patient's presenting complaint and laboratory results, you wonder whether the suspicion of acute cholecystitis is high enough to warrant further testing.

WHY IS THIS QUESTION IMPORTANT?

Acute cholecystitis accounts for 3% to 9% of hospital admissions for acute abdominal pain.¹⁻⁴ Most patients presenting with upper abdominal complaints are subsequently found to have a relatively benign cause of pain (eg, dyspepsia or gastroenteritis),^{2,5} but the possibility of acute cholecystitis mandates the completion of a comprehensive and at times laborious diagnostic evaluation. The importance of this clinical dilemma is only magnified by the frequency with which abdominal pain is encountered in clinical practice.⁶⁻⁸

Traditionally, the diagnosis of acute cholecystitis was followed by a several-week "cooling off" period before proceeding to surgery. Most clinicians now advocate early cholecystectomy (ie, within several days of the onset of symptoms),⁹ because it leads to lower complication rates, reduced costs, and shortened recovery periods.¹⁰⁻¹⁴

DEFINITION OF CHOLECYSTITIS

Defining cholecystitis as "inflammation of the gallbladder" implies a pathologic state. What clinicians usually mean by acute cholecystitis, however, is the presence of this pathologic state (seen macroscopically at laparotomy or microscopically by the pathologist) in the setting of a plausibly

related clinical presentation. Practically speaking, cholecystitis is a syndrome encompassing a continuum of clinico-pathologic states. At one end of this continuum is symptomatic cholelithiasis, with acute attacks of pain (biliary colic) that resolve in 4 to 6 hours. At the other end, that which is typically associated with the term *acute cholecystitis*, is a clinical picture in which biliary colic is longer lasting and accompanied by fever, laboratory markers of inflammation, or cholestasis.^{15,16} Gallbladder inflammation without gallstones (ie, acalculous cholecystitis) typically occurs in critically ill patients and is consequently associated with a high mortality rate.^{17,18}

HOW TO ELICIT THE RELEVANT SIGNS AND SYMPTOMS

*Cope's Early Diagnosis of the Acute Abdomen*¹⁵ points out that "biliary colic" is a misnomer because biliary obstruction produces pain of a steady, nonparoxysmal nature. A majority of studies have explicitly defined biliary colic in similar terms (eg, a steady right upper quadrant pain lasting for at least 30 minutes), but others have used the term without definition.¹⁹ *Cope's*¹⁵ also stresses that biliary colic localizes to the midepigastrium as often as to the right upper quadrant. A recent systematic review¹⁹ supports this observation because "upper abdominal pain" exhibited test characteristics comparable to right upper quadrant pain. Thus, the clinician should inquire about both pain in the upper quadrant and more generally pain in the upper abdomen. The clinician should also ask the patient about fat intolerance because abdominal discomfort after fatty meals may have a predictive value similar to that of biliary colic.¹⁹

Physical findings most famously associated with the gallbladder are the Courvoisier and Murphy signs. The Courvoisier sign has evolved in meaning,²⁰ but standard definitions describe the sign as referring to a palpable, nontender gallbladder in a patient with jaundice.^{21,22} Courvoisier observed that dilation of the gallbladder occurred more commonly when obstruction resulted from malignancy, rather than from benign conditions such as gallstones. Although this association is real, the sign should not be elevated to the status of a "law,"²⁰⁻²² because recent reports confirm the occurrence of the Courvoisier sign in biliary conditions other than obstructive malignancies.²³

The Murphy sign refers to pain and arrested inspiration occurring when the patient inspires deeply while the examiner's fingers are hooked underneath the right costal margin.^{21,22,24} Data addressing the usefulness of the Murphy sign in evaluating patients suspected of having acute cholecystitis are discussed along with other findings from the systematic review presented below. The only other physical sign we identified as specifically associated with acute cholecystitis was the Boas sign. Originally, this sign referred to point tenderness in the region to the right of the 10th to 12th thoracic vertebrae,²⁵⁻²⁷ but contemporary sources describe hyperesthesia to light touch in the right upper quadrant or infrascapular area.²² One study²⁸ reported that 7% of patients undergoing cholecystectomy exhibited hyperesthesia

in this region, but no patient exhibited the Boas sign in the original sense. None of the other studies reviewed below assessed the Boas sign in either form.

ACCURACY OF DIAGNOSTIC IMAGING

Ultrasonography of the right upper quadrant has emerged as the most commonly used imaging modality for suspected cholecystitis. Meta-analysis of the diagnostic performance of ultrasonography in detecting acute cholecystitis indicated an unadjusted sensitivity and specificity of 94% and 78%, respectively.²⁹ The investigators included in their analysis adjustments for verification bias³⁰⁻³² (also called workup bias³³), which refers to the distorted diagnostic test characteristics observed when the decision to proceed with a gold standard test (eg, cholecystectomy) is affected by the results of preliminary tests such as right upper quadrant ultrasonography. Patients with a negative ultrasonography result will undergo cholecystectomy only in the setting of extremely typical clinical findings. The consequent loss of patients with atypical clinical presentations reduces the opportunity for false-negative ultrasonography results, thus inflating the apparent sensitivity of ultrasonography and its associated "rule-out" power. Conversely, specificity and the associated "rule in" ability of ultrasonography are underestimated.

Adjustments for the effects of verification bias in the above-mentioned meta-analysis²⁹ indicated that ultrasonography detects acute cholecystitis with sensitivity of 88% (95% confidence interval [CI], 74%-100%) and specificity of 80% (95% CI, 62%-98%). Sensitivity for the detection of cholelithiasis was comparable, but specificity was higher, at approximately 99%. Radionuclide scanning has slightly better test characteristics for the diagnosis of acute cholecystitis but offers no evaluation of alternative abdominal diagnoses and has the disadvantages of greater inconvenience and patient exposure to radionuclides.²⁹ Computed tomography of the abdomen, although useful for the evaluation of suspected complications and concurrent intra-abdominal conditions, is inferior to ultrasonography in the assessment of acute biliary disease.^{34,35}

METHODS

The initial electronic search queried the MEDLINE database for January 1966 through November 2000 (limited to English-language articles) using the Medical Subject Headings (MeSH) "acute abdomen," "abdominal pain," "cholecystitis," "cholelithiasis," "gallbladder," and "gallbladder diseases." These terms were then combined with various combinations of MeSH terms, title words, and text words: "physical examination," "medical history taking," "professional competence," "sensitivity and specificity," "reproducibility of results," "observer variation," "diagnostic tests," "decision support techniques," "Bayes theorem," "predictive value of tests," "palpation," "percussion," "differential diagnosis," and "diagnostic errors." The Science Citation Index and Cochrane Library were also searched, and a hand search of Index Medicus was conducted for 1950 through 1965, using the terms

“cholecystitis,” “acute abdomen,” and “gallbladder.” Bibliographies of identified articles were searched for additional pertinent articles, as were the bibliographies of prominent textbooks of physical examination, surgery, and gastroenterology. An electronic search of MEDLINE was repeated in July 2002 to look for any relevant articles appearing since completion of the more comprehensive search.

Two authors (RT and NR) independently abstracted data from the identified studies, and all 3 authors reviewed these data for inclusion. Included studies evaluated the role of a clinical test (including medical history, physical examination, and basic laboratory tests) in adult patients with abdominal pain or suspected acute cholecystitis. Included studies were also required to report data from a control group of patients subsequently found not to have acute cholecystitis, with sufficient detail to allow construction of 2×2 tables. Finally, studies were required to define cholecystitis according to an adequate gold standard, including surgery, pathologic examination, radiographic imaging (hepatic iminodiacetic acid [HIDA] scan or

right upper quadrant ultrasonography), or clinical follow-up documenting a course consistent with acute cholecystitis and without evidence for an alternate diagnosis.

Summary measures for the sensitivity of the evaluated components of the clinical examination and basic laboratory tests for cholecystitis were derived from published raw data from the reported studies meeting our inclusion criteria. A random-effects model was used to generate conservative summary measures and CIs for the sensitivity and likelihood ratios (LRs).³⁶⁻³⁸ For LRs, a summary measure is reported only when more than 2 studies were identified; otherwise, a range was reported.

RESULTS

Of 195 studies identified by our search, 17 evaluated the role of the clinical examination or basic laboratory test in patients with acute abdominal pain and possible acute cholecystitis and also met our inclusion criteria ([Table 12-1](#)).³⁹⁻⁵⁵

Table 12-1 Studies of the Diagnostic Performance of Clinical and Laboratory Findings in Detecting Acute Cholecystitis

Source	Study Period	Selection Criteria	Design	Sample Size	Consecutive Patients	Basis for Diagnosis
Adedeji and McAdam, ³⁹ 1996	1985-1990	Acute abdominal pain and age > 70 y	Retrospective	431	Yes	Clinical follow-up
Bednarz et al, ⁴⁰ 1986	1983-1984	Suspected acute cholecystitis and referred for HIDA scan	Prospective	70	Yes	Surgery (43%) Clinical impression (57%)
Brewer et al, ⁴¹ 1976	1971-1972	Abdominal pain	Retrospective	570	Yes	Multiple
Dunlop et al, ⁴² 1989	1982-1986	Acute abdominal pain and suspected acute cholecystitis	Prospective	270	Yes	Pathology (71%) Clinical impression (29%)
Eikman et al, ⁴³ 1975	Not stated	Suspected acute cholecystitis and referred for radiology testing	Prospective	38	Yes	Surgical (38%) Clinical impression (62%)
Gruber et al, ⁴⁴ 1996	1990-1993	Positive HIDA scan results and underwent surgery for suspected acute cholecystitis	Retrospective	198	Yes	Pathology
Halasz, ⁴⁵ 1975	1969-1974	Suspected acute cholecystitis	Retrospective	238	Yes	Surgery (65%) Other (35%) ^a
Johnson and Cooper, ⁴⁶ 1995	Not stated	Positive HIDA scan results and underwent surgery for suspected acute cholecystitis	Retrospective	69	No	Pathology
Juvonen et al, ⁴⁷ 1992	1988-1989	Suspected acute cholecystitis referred for ultrasonography	Prospective	129	Yes	Pathology (95%) Ultrasonography (5%)
Liddington and Thomson, ⁴⁸ 1991	Not stated	Abdominal pain	Prospective	142	No	Clinical impression
Lindenauer and Child, ⁴⁹ 1966	1959-1964	Underwent cholecystectomy	Retrospective	200	No	Pathology
Potts and Vukov, ⁵⁰ 1999	1992-1995	Abdominal pain requiring operation and age > 80 y	Retrospective	117	Yes	Pathology
Prevot et al, ⁵¹ 1999	1997-1999	ICU patients with suspected acute acalculous cholecystitis	Prospective	32	Yes	Pathology (50%) Clinical impression (50%)
Raine and Gunn, ⁵² 1975	1965-1973	Suspected acute cholecystitis and underwent surgery	Prospective	156	Yes	Pathology
Schofield et al, ⁵³ 1986	Not stated	Abdominal pain and suspected acute cholecystitis	Prospective	100	Yes	Gallstones at laparotomy
Singer et al, ⁵⁴ 1996	1993	Suspected acute cholecystitis and radiology testing completed	Retrospective	100	Yes	Pathology (44%) HIDA scintigraphy (56%)
Staniland et al, ⁵⁵ 1972	Not stated	Admission for abdominal pain of < 1 wk	Retrospective	600	No	Surgery

Abbreviations: HIDA, hepatic iminodiacetic acid; ICU, intensive care unit.

^aRadiology testing and clinical follow-up; exact proportions not specified.

Twelve of these studies^{40,42-47,49,51-54} enrolled patients specifically suspected of having acute cholecystitis, with inclusion of many of these studies based on patient referral for radiology testing (ie, HIDA scan or right upper quadrant ultrasonography) for the confirmation of a clinical diagnosis. The remaining 5 studies^{39,41,48,50,55} enrolled patients presenting with abdominal pain and did not require a specific suspicion of acute cholecystitis for patient inclusion. Each of the 17 studies evaluated a variable number of clinical and laboratory findings included in the evaluation of suspected cholecystitis, ranging from 1 to 9 characteristics per study (Table 12-2).

Precision of Signs and Symptoms

Measurements of laboratory characteristics and objective clinical signs such as temperature are assumed to have

high precision, but the reproducibility of other aspects of the clinical examination for cholecystitis remains largely unknown. In fact, the only study identified as assessing the precision of some aspect of the clinical examination for biliary disease was an evaluation of the diagnostic value of iridology⁵⁶ (iridologists believe that intricate neural connections between major organs and the iris permit diagnosis of general medical conditions through inspection of iris pigmentation patterns^{57,58}). In this relatively well-designed study, the accuracy and precision of iridologic signs for the diagnosis of cholecystitis were barely distinguishable from values expected by chance alone ($\kappa = -0.06$ to 0.28 for the 10 possible observer pairs).

Unfortunately, analogous studies have not been carried out with conventional clinical maneuvers related to the

Table 12-2 Summary Test Characteristics for Clinical and Laboratory Findings in Included Studies

Finding (No. of Studies)	No. of Patients ^a	Sensitivity (95% CI)	Specificity (95% CI)	Summary LR ^b	
				LR+ (95% CI)	LR- (95% CI)
Clinical					
Anorexia (2) ^{41,55}	1135	0.65 (0.57-0.73)	0.50 (0.49-0.51)	1.1-1.7	0.5-0.9
Emesis (4) ^{41,46,53,55}	1338	0.71 (0.65-0.76)	0.53 (0.52-0.55)	1.5 (1.1-2.1)	0.6 (0.3-0.9)
Fever (>35°C) (8) ^{40,41,44,46,50-53}	1292	0.35 (0.31-0.38)	0.80 (0.78-0.82)	1.5 (1.0-2.3)	0.9 (0.8-1.0)
Guarding (2) ^{41,55}	1170	0.45 (0.37-0.54)	0.70 (0.69-0.71)	1.1-2.8	0.5-1.0
Murphy sign (3) ^{39,46,54}	565	0.65 (0.58-0.71)	0.87 (0.85-0.89)	2.8 (0.8-8.6)	0.5 (0.2-1.0)
Nausea (2) ^{46,54}	669	0.77 (0.69-0.83)	0.36 (0.34-0.38)	1.0-1.2	0.6-1.0
Rebound (4) ^{40,41,48,55}	1381	0.30 (0.23-0.37)	0.68 (0.67-0.69)	1.0 (0.6-1.7)	1.0 (0.8-1.4)
Rectal tenderness (2) ^{41,55}	1170	0.08 (0.04-0.14)	0.82 (0.81-0.83)	0.3-0.7	1.0-1.3
Rigidity (2) ^{41,55}	1140	0.11 (0.06-0.18)	0.87 (0.86-0.87)	0.50-2.32	1.0-1.2
Right upper abdominal quadrant					
Mass (4) ^{40,45,53,54}	408	0.21 (0.18-0.23)	0.80 (0.75-0.85)	0.8 (0.5-1.2)	1.0 (0.9-1.1)
Pain (5) ^{40,45,46,54,55}	949	0.81 (0.78-0.85)	0.67 (0.65-0.69)	1.5 (0.9-2.5)	0.7 (0.3-1.6)
Tenderness (4) ^{40,45,54,55}	1001	0.77 (0.73-0.81)	0.54 (0.52-0.56)	1.6 (1.0-2.5)	0.4 (0.2-1.1)
Laboratory					
Alkaline phosphatase > 120 U/L (4) ^{42,46,49,51}	556	0.45 (0.41-0.49)	0.52 (0.47-0.57)	0.8 (0.4-1.6)	1.1 (0.6-2.0)
Elevated ALT or AST level ^c (5) ^{42,46,49,51,53}	592	0.38 (0.35-0.42)	0.62 (0.57-0.67)	1.0 (0.5-2.0)	1.0 (0.8-1.4)
Total bilirubin > 2 mg/dL (6) ^{40,42,43,46,49,51}	674	0.45 (0.41-0.49)	0.63 (0.59-0.66)	1.3 (0.7-2.3)	0.9 (0.7-1.2)
Total bilirubin, AST, or alkaline phosphatase (1) ⁵²	270				
All 3 elevated		0.34 (0.30-0.36)	0.80 (0.69-0.88)	1.6 (1.0-2.8)	0.8 (0.8-0.9)
Any 1 elevated		0.70 (0.67-0.73)	0.42 (0.31-0.53)	1.2 (1.0-1.5)	0.7 (0.6-0.9)
Leukocytosis ^d (7) ^{41,44,46,50-53}	1197	0.63 (0.60-0.67)	0.57 (0.54-0.59)	1.5 (1.2-1.9)	0.6 (0.5-1.8)
Leukocytosis ^d and fever (2) ^{44,52}	351	0.24 (0.21-0.26)	0.85 (0.76-0.91)	1.6 (0.9-2.8)	0.9 (0.8-1.0)

Abbreviations: ALT, alanine aminotransferase; AST, aspartate aminotransferase; CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

^aMay not equal sums of N in Table 12-1 because not all studies applied all tests to all patients.

^bSummary measures provided only for findings discussed by more than 2 studies.

^cGreater than upper limit of normal (ALT, 40 U/L; AST, 48 U/L).

^dWhite blood cell count of more than 10/ μ L.

diagnosis of cholecystitis. In fact, as observed in a previous article in this series,⁵⁹ the precision of even the most basic components of the abdominal examination (eg, guarding, rigidity, and rebound tenderness) remains uncharacterized. Poor reproducibility for abdominal examination would erode the assessments of sensitivity and specificity provided by different investigators. Presumably, then, one can infer a certain degree of interrater reliability from the fact that multiple studies demonstrate modest sensitivity for these signs in diagnosing important abdominal conditions.⁵⁹ Nonetheless, further assessments of core components of the abdominal examination would be a welcome addition to the literature.

Accuracy of Signs and Symptoms

No single clinical or laboratory finding had an LR⁻ sufficiently low to rule out the diagnosis of acute cholecystitis (Table 12-2). Even the absence of right upper quadrant tenderness does not rule out acute cholecystitis with its LR of 0.4. Elderly patients may be particularly prone to present without signs or symptoms referable to the right upper quadrant.⁶⁰

Similarly, individual symptoms, signs, and laboratory results did not have LR⁺s sufficiently high to rule in the diagnosis of acute cholecystitis. In fact, none of the LR⁺s were more than 2.0, with the exception of the Murphy sign, which was associated with a ratio of 2.8. The 95% CI for this summary estimate included 1.0, but the use of the Murphy sign was especially prone to verification bias. Thus, the true LR⁺ might exceed the estimated value.

Limitations of the Literature

The problem of verification (or workup) bias³⁰⁻³³ was discussed in the section on diagnostic imaging but likely affected all of the clinical and laboratory findings assessed in this review. Patients with upper abdominal tenderness, fever, abnormal liver function results, or other “typical” findings more commonly undergo further evaluation (eg, diagnostic imaging) for acute cholecystitis than do patients presenting without these findings. The lack of patients with atypical presentations in studies leads to overestimates of sensitivity and underestimates of specificity. Supplementing the diagnosis of cholecystitis with clinical follow-up would mitigate the effects of verification bias, but only 1 study³⁹ incorporated clinical follow-up in the diagnostic protocol.

Spectrum bias⁶¹ (or, more recently, spectrum effect⁶²) distorts test characteristics since there is inadequate representation of the relevant disease and disease-free states in the patient samples used to evaluate the test of interest. The prevalence of cholecystitis in the study populations was as high as 80% and averaged 41%, in contrast to the prevalence of 3% to 5% among patients presenting with abdominal pain of less than 1 week’s duration.^{1,2,41}

Subgroup analysis can generate values for sensitivity and specificity in patient populations with substantially differ-

ent previous likelihoods of disease from the average value.⁶² Because available data often do not permit such analysis, one has to make qualitative inferences about the difference between the prior probability of disease in a particular patient and the prevalence in the population used to evaluate the test. For instance, a high prevalence of cholecystitis in clinical reports reduces the opportunity to detect both false-positive and true-negative results compared to the findings in patients with a lower prevalence of disease. Thus, clinical findings and laboratory tests used to evaluate cholecystitis may have different sensitivity and specificity than suggested in the available literature.

Other limitations to the existing literature include the retrospective design of most studies, modest sample sizes, unblinded assessment of key outcomes and test results, and the variability in criteria for establishing a diagnosis of cholecystitis. The included studies varied between accepting clinicians’ diagnostic impressions (usually incorporating imaging results), findings at laparotomy, and pathologic findings as the means of diagnosis. Unfortunately, the correlation between clinical and pathologic diagnoses of cholecystitis is poor.⁶³ Gallstones occur commonly enough that their presence, even in the context of inflammatory cells, may be “true but unrelated” with respect to the patient’s acute presentation. Overdiagnosis from this and other available gold standards likely resulted in an overestimation of the prevalence of acute cholecystitis, with consequent distortion of the usefulness of clinical and basic laboratory findings. Finally, studies assessing both calculous and acalculous cholecystitis were included in the review. Although these entities share many clinical traits, the nonspecific presentation of acalculous cholecystitis likely eroded the value of several clinical findings.

Combinations of Findings and the Clinical “Gestalt”

Even with the above limitations, it seems unlikely that individual clinical or laboratory findings have LR⁺ or LR⁻ of sufficient magnitude to play a decisive role in the diagnosis of acute cholecystitis. Thus, one might look to combinations of clinical signs and symptoms to facilitate, confirm, or exclude the diagnosis of cholecystitis. Unfortunately, only 3 included studies^{42,44,52} specifically evaluated the value of such combinations. Two studies evaluated the combination of fever and leukocytosis; the third reviewed various combinations of liver function tests. Assessments of the LRs of the above combinations demonstrated no benefit over their individual components, suggesting that these tests did not function independently of one another. Indeed, fever and leukocytosis may be seen as different manifestations of the same underlying process of nonspecific inflammation, so it is not surprising that combining them provided no synergistic diagnostic value. Similarly, right upper quadrant pain and the Murphy sign likely reflect the same underlying pathophysiologic process (ie, local inflammation and peritoneal irritation), so that these findings would not be expected to function independently of one another.

Although the existing literature does not identify specific clinically useful combinations of findings, the effect of such combinations can be estimated with available data. In 2 randomized trials of early vs delayed cholecystectomy,^{13,14} laparotomy failed to confirm the preoperative diagnosis of acute cholecystitis in 5 of 99 patients (95% CI, 1.9–12)¹⁴ and in 0 of 104 patients (95% CI, 0–4.4).¹³ Given a likely bias toward confirming the preoperative diagnosis, let us assume that the actual false-positive rate for the clinical diagnosis of cholecystitis is higher (eg, 15%) than suggested by these values.

A 15% false-positive rate would imply an 85% posttest probability for all clinical, laboratory, and radiologic tests. We know that ultrasonography of the right upper quadrant has a sensitivity and specificity of 88% and 80%, respectively.²⁹ Working backward, we can infer that the composite clinical evaluation generates a pretest probability of approximately 60% before the results of ultrasonography are obtained. This posttest probability of 60% for the clinical suspicion of cholecystitis reflects the diagnostic power of the clinical evaluation before ultrasonography, as well as the pretest probability. At this stage in the diagnostic process, the pretest probability reflects the prevalence of the diagnosis, which is approximately 5% among patients presenting to the emergency department with abdominal pain.^{1,2,41} Thus, the clinical diagnosis of acute cholecystitis formulated according to medical history, physical examination, and basic laboratory testing must increase the pretest probability from 5% to 60%.

Achieving this increase in pretest probability requires that the gestalt comprising certain clinical and laboratory findings have an LR+ on the order of 25 to 30. To put this range in perspective, “typical angina” has an LR+ of 115 for the diagnosis of coronary artery stenosis greater than 75% in adult men. Nonsloping depression of the ST segment of at least 2.5 mm during exercise electrocardiography has an LR+ of 39 for the same diagnosis.⁶⁴ Thus, our estimate for the diagnostic usefulness of the clinical gestalt in diagnosing acute cholecystitis, approximate and speculative as it is, confirms the impression of many clinicians that the overall clinical assessment plays a crucial role in arriving at a diagnosis.

It is tempting to supplement the existing literature by asking experts for their opinion on which specific findings drive the clinical impression for or against acute cholecystitis. Unfortunately, discerning the key elements of the clinical assessment can prove deceptive, even for experienced clinicians. For instance, a recent clinical model for the prediction of pulmonary embolism omits hypoxemia and pleurisy from the algorithm for determining pretest probability.^{65,66} Similarly, many of the classic descriptors of angina have surprisingly little influence on the assessment of chest pain.⁶⁷ This dissociation between commonly accepted harbingers of disease and evidence-based determinants of disease probability undermines the role of expert opinion in identifying key clinical findings even for common conditions. Consequently, tempting as it is to open the “black box” of the clinical gestalt for cholecystitis, doing so will require further study of specific clinical findings or, more likely, combinations of findings.

CLINICAL SCENARIO—RESOLUTION

Your differential diagnosis for the patient’s presentation includes viral hepatitis, cholecystitis, and gallstone pancreatitis. To validate your impression and help establish the relative likelihood of each, you ask the emergency department attending physician to evaluate the patient. She regards the likelihood of cholecystitis as high enough to warrant diagnostic imaging. In fact, her clinical impression is that cholecystitis is the leading diagnosis, so she recommends urgent right upper quadrant ultrasonography. The ultrasonography subsequently reveals the presence of gallstones, gallbladder wall thickening, and a sonographic Murphy sign. These findings, in the context of the patient’s presentation, virtually confirm the diagnosis of acute cholecystitis.⁶⁸

THE BOTTOM LINE

The existing literature identifies no single finding with sufficient diagnostic power to establish or exclude acute cholecystitis without further testing (eg, right upper quadrant ultrasonography). Combinations of certain symptoms, signs, and laboratory results likely have more useful LRs and presumably inform the diagnostic impressions of experienced clinicians. Future research may allow the development of prediction rules that combine basic demographics with clinical findings to distinguish patients who require no further testing from those who require continued diagnostic evaluation, as is currently possible with the evaluation of suspected pulmonary embolism.^{66,69} Until then, the clinical evaluation of patients with abdominal pain suggestive of cholecystitis will continue to rely heavily on the clinical gestalt and diagnostic imaging.

Author Affiliations at the Time of the Original Publication

Maine Hospitalist Service, Maine Medical Center, Portland (Dr Trowbridge); Department of Medicine, California Pacific Medical Center, San Francisco (Dr Rutkowski); Department of Medicine, University of California, San Francisco (Dr Shojania).

Acknowledgments

Dr Trowbridge’s work was supported in part by a grant from the Josiah Macy Jr Foundation.

We thank Theodore N. Pappas, MD, David Edelman, MD, Robert Badgett, MD, and James Wagner, MD, for their helpful comments on drafts of the manuscript.

REFERENCES

- Powers RD, Guertler AT. Abdominal pain in the ED. *Am J Emerg Med*. 1995;13(3):301-303.
- Irvin TT. Abdominal pain. *Br J Surg*. 1989;76(11):1121-1125.
- Kizer KW, Vassar MJ. Emergency department diagnosis of abdominal disorders in the elderly. *Am J Emerg Med*. 1998;16(4):357-362.
- Miettinen P, Pasanen P, Lahtinen J, Alhava E. Acute abdominal pain in adults. *Ann Chir Gynaecol*. 1996;85(1):5-9.

5. Wasson JH, Sox HC Jr, Sox CH. The diagnosis of abdominal pain in ambulatory male patients. *Med Decis Making*. 1981;1(3):215-224.
6. Popovic JR. 1999 National Hospital Discharge Survey: annual summary with detailed diagnosis and procedure data. *Vital Health Stat* 13. 2001;(151):i-v, 1-206.
7. Cherry DK, Burt CW, Woodwell DA. *National Ambulatory Medical Care Survey: 1999 Summary Advance Data from Vital and Health Statistics*. Hyattsville, MD: National Center for Health Statistics; 2001. No. 322.
8. McCaig L, Burt CW. *National Hospital Ambulatory Medical Care Survey: 1999 Emergency Department Summary: Advance Data from Vital and Health Statistics*. Hyattsville, MD: National Center for Health Statistics; 2001. No. 320.
9. Lillemoe KD. Surgical treatment of biliary tract infections. *Am Surg*. 2000;66(2):138-144.
10. Hashizume M, Sugimachi K, MacFadyen BV. The clinical management and results of surgery for acute cholecystitis. *Semin Laparosc Surg*. 1998;5(2):69-80.
11. Chandler CF, Lane JS, Ferguson P, Thompson JE, Ashley SW. Prospective evaluation of early versus delayed laparoscopic cholecystectomy for treatment of acute cholecystitis. *Am Surg*. 2000;66(9):896-900.
12. Eldar S, Eitan A, Bickel A, et al. The impact of patient delay and physician delay on the outcome of laparoscopic cholecystectomy for acute cholecystitis. *Am J Surg*. 1999;178(4):303-307.
13. Lai PB, Kwong KH, Leung KL, et al. Randomized trial of early versus delayed laparoscopic cholecystectomy for acute cholecystitis. *Br J Surg*. 1998;85(6):764-767.
14. Lo CM, Liu CL, Fan ST, Lai EC, Wong J. Prospective randomized study of early versus delayed laparoscopic cholecystectomy for acute cholecystitis. *Ann Surg*. 1998;227(4):461-467.
15. Silen W. *Cope's Early Diagnosis of the Acute Abdomen*. 20th ed. New York, NY: Oxford University Press; 2000:128-137, 145-147.
16. Ahrendt SA, Pitt HA. Biliary tract. In: Townsend CM, Beauchamp RD, Evers BM, Mattox KL, eds. *Sabiston Textbook of Surgery*. 16th ed. Philadelphia, PA: WB Saunders; 2001:1076-1111.
17. Kallafas S, Ziegler DW, Flanchbaum L, Choban PS. Acute acalculous cholecystitis: incidence, risk factors, diagnosis, and outcome. *Am Surg*. 1998;64(5):471-475.
18. Warren BL, Carstens CA, Falck VG. Acute acalculous cholecystitis—a clinical-pathological disease spectrum. *S Afr J Surg*. 1999;37(4):99-104.
19. Berger MY, van der Velden JJ, Lijmer JG, de Kort H, Prins A, Bohnen AM. Abdominal symptoms: do they predict gallstones? a systematic review. *Scand J Gastroenterol*. 2000;35(1):70-76.
20. Vergheese A, Dison C, Berk SL. Courvoisier's "law"—an eponym in evolution. *Am J Gastroenterol*. 1987;82(3):248-250.
21. McGee S. *Evidence-Based Physical Diagnosis*. Philadelphia, PA: WB Saunders; 2001:594-639.
22. Sapira JD. *The Art and Science of Bedside Diagnosis*. Baltimore, MD: Urban & Schwartzenberg Inc; 1990:371-390.
23. Fournier AM, Michel J. Courvoisier's sign revisited: two patients with palpable gallbladder. *South Med J*. 1992;85(5):548-550.
24. Aldea PA, Meehan JP, Sternbach G. The acute abdomen and Murphy's signs. *J Emerg Med*. 1986;4(1):57-63.
25. Joachim H. *Practical Bedside Diagnosis and Treatment*. Baltimore, MD: CC Thomas; 1940.
26. Pullen RLR. *Medical Diagnosis: Applied Physical Diagnosis*. London, England: WB Saunders; 1944.
27. Zatuchni J. *Notes on Physical Diagnosis*. Philadelphia, PA: FA Davis Co; 1964.
28. Gunn A, Keddie N. Some clinical observations on patients with gallstones. *Lancet*. 1972;2(7771):239-241.
29. Shea JA, Berlin JA, Escarce JJ, et al. Revised estimates of diagnostic test sensitivity and specificity in suspected biliary tract disease. *Arch Intern Med*. 1994;154(22):2573-2581.
30. Begg CB, Greenes RA. Assessment of diagnostic tests when disease verification is subject to selection bias. *Biometrics*. 1983;39(1):207-215.
31. Begg CB, McNeil BJ. Assessment of radiologic tests. *Radiology*. 1988;167(2):565-569.
32. Zhou XH, Higgs RE. Assessing the relative accuracies of two screening tests in the presence of verification bias. *Stat Med*. 2000;19(11-12):1697-1705.
33. Guyatt G, Rennie D, for the Evidence-Based Medicine Working Group, American Medical Association. *Users' Guides to the Medical Literature: A Manual for Evidence-Based Clinical Practice*. Chicago, IL: AMA Press; 2002.
34. Harvey RT, Miller WT Jr. Acute biliary disease. *Radiology*. 1999;213(3): 831-836.
35. Fidler J, Paulson EK, Layfield L. CT evaluation of acute cholecystitis: findings and usefulness in diagnosis. *AJR Am J Roentgenol*. 1996;166(5): 1085-1088.
36. Eddy D, Hasselblad V, Shachter R. *Meta-Analysis by the Confidence Profile Method: The Statistical Synthesis of Evidence*. San Diego, CA: Academic Press; 1992:93-98.
37. Eddy D, Hasselblad V. *Fast*Pro v1.8: Software for Meta-analysis by the Confidence Profile Method*. San Diego, CA: Academic Press; 1992.
38. Borenstein M, Rothstein H. *Comprehensive Meta-Analysis: A Computer Program for Research Synthesis*. Englewood, NJ: Biostat Inc; 1999.
39. Adejedi OA, McAdam WA. Murphy's sign, acute cholecystitis and elderly people. *J R Coll Surg Edinb*. 1996;41(2):88-89.
40. Bednarz GM, Kalff V, Kelly MJ. Hepatobilary scintigraphy. *Med J Aust*. 1986;145(7):316-318.
41. Brewer BJ, Golden GT, Hitch DC, Rudolf LE, Wangenstein SL. Abdominal pain. *Am J Surg*. 1976;131(2):219-223.
42. Dunlop MG, King PM, Gunn AA. Acute abdominal pain. *J R Coll Surg Edinb*. 1989;34(3):124-127.
43. Eikman EA, Cameron JL, Colman M, et al. A test for patency of the cystic duct in acute cholecystitis. *Ann Intern Med*. 1975;82(3):318-322.
44. Gruber PJ, Silverman RA, Gottesfeld S, Flaster E. Presence of fever and leukocytosis in acute cholecystitis. *Ann Emerg Med*. 1996;28(3):273-277.
45. Halasz NA. Counterfeit cholecystitis, a common diagnostic dilemma. *Am J Surg*. 1975;130(2):189-193.
46. Johnson H Jr, Cooper B. The value of HIDA scans in the initial evaluation of patients for cholecystitis. *J Natl Med Assoc*. 1995;87(1):27-32.
47. Juvonen T, Kiviniemi H, Niemela O, Kairaluoma MI. Diagnostic accuracy of ultrasonography and C reactive protein concentration in acute cholecystitis: a prospective clinical study. *Eur J Surg*. 1992;158(6-7):365-369.
48. Lidington MI, Thomson WH. Rebound tenderness test. *Br J Surg*. 1991;78(7):795-796.
49. Lindenauer SM, Child CG III. Disturbances of liver function in biliary tract disease. *Surg Gynecol Obstet*. 1966;123(6):1205-1211.
50. Potts FE IV, Vukov LF. Utility of fever and leukocytosis in acute surgical abdomens in octogenarians and beyond. *J Gerontol A Biol Sci Med Sci*. 1999;54(2):M55-M58.
51. Prevot N, Mariat G, Mahul P, et al. Contribution of cholescintigraphy to the early diagnosis of acute acalculous cholecystitis in intensive-care-unit patients. *Eur J Nucl Med*. 1999;26(10):1317-1325.
52. Raine PA, Gunn AA. Acute cholecystitis. *Br J Surg*. 1975;62:697-700.
53. Schofield PF, Hulton NR, Baird AD. Is it acute cholecystitis? *Ann R Coll Surg Engl*. 1986;68(1):14-16.
54. Singer AJ, McCracken G, Henry MC, Thode HC Jr, Cabahug CJ. Correlation among clinical, laboratory, and hepatobilary scanning findings in patients with suspected acute cholecystitis. *Ann Emerg Med*. 1996;28(3):267-272.
55. Staniland JR, Ditchburn J, De Dombal FT. Clinical presentation of acute abdomen: study of 600 patients. *BMJ*. 1972;3(5823):393-398.
56. Knipschild P. Looking for gall bladder disease in the patient's iris. *BMJ*. 1988;297(6663):1578-1581.
57. Ernst E. Iridology: not useful and potentially harmful. *Arch Ophthalmol*. 2000;118(1):120-121.
58. Simon A, Worthen DM, Mitas II JA. An evaluation of iridology. *JAMA*. 1979;242(13):1385-1389.
59. Wagner JM, McKinney WP, Carpenter JL. Does this patient have appendicitis? *JAMA*. 1996;276(19):1589-1594.
60. Parker LJ, Vukov LF, Wollan PC. Emergency department evaluation of geriatric patients with acute cholecystitis. *Acad Emerg Med*. 1997;4(1):51-55.
61. Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *N Engl J Med*. 1978;299(17):926-930.
62. Mulherin SA, Miller WC. Spectrum bias or spectrum effect? *Ann Intern Med*. 2002;137(7):598-602.
63. Fitzgibbons RJ Jr, Tseng A, Wang H, Ryberg A, Nguyen N, Sims KL. Acute cholecystitis. *Surg Endosc*. 1996;10(12):1180-1184.
64. Diamond GA, Forrester JS. Analysis of probability as an aid in the clinical diagnosis of coronary-artery disease. *N Engl J Med*. 1979;300(24): 1350-1358.
65. Rodger MA, Carrier M, Jones GN, et al. Diagnostic value of arterial blood gas measurement in suspected pulmonary embolism. *Am J Respir Crit Care Med*. 2000;162(6):2105-2108.

66. Wells PS, Anderson DR, Rodger M, et al. Excluding pulmonary embolism at the bedside without diagnostic imaging. *Ann Intern Med.* 2001;135(2):98-107.
67. Panju AA, Hemmelgarn BR, Guyatt GH, Simel DL. Is this patient having a myocardial infarction? *JAMA.* 1998;280(14):1256-1263.
68. Ralls PW, Colletti PM, Lapin SA, et al. Real-time sonography in suspected acute cholecystitis. *Radiology.* 1985;155(3):767-771.
69. Wicki J, Perneger TV, Junod AF, Bounameaux H, Perrier A. Assessing clinical probability of pulmonary embolism in the emergency ward: a simple score. *Arch Intern Med.* 2001;161(1):92-97.

Prepared by Robert L. Trowbridge, MD, and Kaveh G. Shojania, MD

Reviewed by Amy Rosenthal, MD

CLINICAL SCENARIO

A 52-year-old man presents to the emergency department with a 6-hour history of dull, epigastric discomfort. The pain began several hours after lunch and did not intensify with eating dinner, although his appetite was poor. He is mildly nauseated and feels warm but denies emesis, diarrhea, dyspnea, chest pain, fever, chills, or previous episodes of a similar pain. The pain neither radiates nor changes with his position.

He has hypertension and gout and drinks 2 oz of whiskey per day. There is no history of gastrointestinal illness or abdominal surgery.

He appears moderately uncomfortable, but he is afebrile, with normal vital signs. The bowel sounds are decreased but present. Although he allows you to palpate his abdomen, you elicit tenderness in the epigastrium without rebound. The Murphy sign is negative. The liver and gallbladder are not palpable, a rectal examination causes no pain, and the stool is negative for occult blood.

On laboratory examination, the white blood cell count is 12700/ μ L (the automated differential shows increased neutrophil levels). He has normal electrolyte, transaminase, bilirubin, amylase, and lipase levels and normal renal function. The alkaline phosphatase level is slightly elevated, at 155 U/L. An electrocardiogram reveals no evidence of ischemia, and the troponin I level is normal.

The emergency physician regards acute cholecystitis as the leading diagnosis. Six months ago, the physician took a 2-day course in ultrasonography and has since performed 40 bedside ultrasonographic tests on patients with abdominal pain, the first 10 of which were proctored by a radiologist to assess competency. The emergency physician performs a focused right upper quadrant ultrasonography in the present patient. He finds no evidence of gallstones, and the point of maximal tenderness does not localize to the gallbladder (ie, there is no sonographic Murphy sign). He recommends that the patient be discharged home, with follow-up in your clinic. However, the patient is still uncomfortable, and you have not established a diagnosis. Have you effectively ruled out acute cholecystitis?

UPDATED SUMMARY ON ACUTE CHOLECYSTITIS

Original Review

Trowbridge RL, Rutkowski NK, Shojania KG. Does this patient have acute cholecystitis? *JAMA*. 2003;289(1):80-86.

UPDATED LITERATURE SEARCH

We repeated the original search strategy that targeted any study involving diagnosis, physical examination, sensitivity and specificity, reproducibility of results, decision support techniques, and other relevant methodologic terms, with any of the following text or keywords: "gallbladder," "gall stones," or "cholecystitis." The updated PubMed search included the years 1998 through September 2004, and we included a more robust search for systematic reviews according to a published strategy.¹ This search yielded 337 articles published since November 11, 2001. An independent search of the OVID database with slight differences in the methodologic terms identified an additional 34 English-language studies published from 2002 to September 2004.

NEW FINDINGS

- The clinician's gestalt is the most important piece of evidence from the clinical evaluation. The single findings with the highest diagnostic value remain Murphy sign (positive likelihood ratio, 2.8) and right upper quadrant tenderness (negative likelihood ratio, 0.4), although the confidence intervals (CIs) for both values cross 1, as documented in the original review.
- Bedside ultrasonography performed by physicians with brief formal training courses may be useful when the result is the combined absence of a sonographic Murphy sign and any evidence of gallstones. Additional studies of bedside ultrasonography by nonradiologists are required.

Details of the Update

The focus of this review remains acute calculous cholecystitis. Studies focusing predominantly on acalculous cholecystitis were excluded.

IMPROVEMENTS IN THE DATA PRESENTED IN THE ORIGINAL PUBLICATION

No new data were found that modify the original results, although we added data on bedside ultrasonography performed by nonradiologists.

CHANGES IN THE REFERENCE STANDARD

Surgical findings combined with pathology or clinical follow-up in patients who do not undergo surgery remain the reference standard for acute cholecystitis.

RESULTS OF LITERATURE REVIEW

Patients reproducibly report biliary symptoms when questioned again 2 weeks after an initial assessment, using an extensive questionnaire addressing the details of their symptoms across various domains—pain, association with eating, changes in bowel habits, and fever, among others.² Physicians concurred with patients' self-reported symptoms to a substantial extent (κ scores > 0.6 and much higher in several cases). Two exceptions were history of fever and radiation to the right shoulder. For these findings, physicians concurred with only moderate agreement ($\kappa = 0.52$ and $\kappa = 0.46$, respectively).

The bedside ultrasonography examination performed by a nonradiologist is an emerging approach to cholecystitis diagnosis.^{3,4} We had not previously included this test, so we conducted a supplemental search for additional articles addressing the utility of bedside ultrasonography. With no date restriction, we found 6 studies, although only 1 study³ was of sufficiently high quality to warrant abstraction and inclusion in the update (Table 12-3). All studies used nonconsecutive convenience samples, but the 5 additional studies excluded from the update also had bias because of nonindependence of reference standard (ie, the decision to undergo confirmatory testing explicitly depended on the results of bedside ultrasonography).⁴⁻⁸ In addition, these studies did not attempt to diagnose

Table 12-3 Bedside Ultrasonographic Findings for Acute Cholecystitis

Finding	LR+ (95% CI)	LR- (95% CI)
Bedside ultrasonography evidence of gallstones and a positive sonographic Murphy sign ^a	2.7 (1.7-4.1)	0.13 (0.04-0.39)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

^aRequires special training and validation of competence.

acute cholecystitis; they only evaluated agreement between bedside ultrasonography and formal ultrasonography with respect to specific radiologic findings.

The single included study showed that physicians with brief training and moderate experience in bedside ultrasonography could adequately visualize the gallbladder in most patients.³ Even among patients with definitive bedside ultrasonography results, defined as the presence of both cholelithiasis and a sonographic Murphy sign, the positive predictive value was only 70%. Thus, patients with positive findings on bedside ultrasonography require confirmatory radiologic investigations before proceeding to surgery. The negative predictive value is 90%. For 30 patients in this study, bedside ultrasonography detected no sonographic Murphy sign and no cholelithiasis. Had these patients not undergone formal ultrasonography, there would have been a 26% reduction in ultrasonography use by the emergency department, at a cost of missing 1 case of cholecystitis. It is tempting to regard this miss rate of 97% as clearly adequate to rule out cholecystitis, but the 95% CI for 1 of 30 extends from 0.6% to 17%. Among patients for whom the pretest suspicion of cholecystitis is low, a definitely negative bedside ultrasonography result probably would be adequate to decide against formal ultrasonography, especially if adequate clinical follow-up is in place.

Evidence From Guidelines

There are no governmental agency guidelines that address the diagnosis of acute cholecystitis.

CLINICAL SCENARIO—RESOLUTION

You decide to observe the patient in the hospital for at least 24 hours. The patient's pain improves with intravenous morphine, and he is admitted to the medical service. Overnight, he requires increasing doses of morphine for pain control, but the electrocardiogram output and cardiac enzyme levels remain normal. The following morning, the abdominal tenderness has worsened, and the white blood cell count has increased to $19\,200/\mu\text{L}$, but the serum amylase level remains normal. Given the persistent concern for acute cholecystitis, he undergoes formal abdominal ultrasonography in the radiology department, which reveals substantial pericholecystic fluid and a positive sonographic Murphy sign. A laparoscopic cholecystectomy reveals an acutely inflamed gallbladder with a small stone impacted in the cystic duct. Pathology confirms the diagnosis of acute cholecystitis.

ACUTE CHOLECYSTITIS—MAKE THE DIAGNOSIS

No single clinical finding, or known combination of clinical history and physical examination findings, efficiently establishes a diagnosis of acute cholecystitis. Thus, clinicians must rely on their clinical gestalt. Bedside ultrasonography requires additional study, and clinicians must receive proper training, followed by demonstration of their proficiency.

PRIOR PROBABILITY

Approximately 5% of emergency department patients with abdominal pain have cholecystitis. Women and Native Americans have a higher risk of cholecystitis. Patients with increased risk of cholecystitis include those with chronic hemolytic disease (eg, sickle cell disease) or recent rapid weight loss.

POPULATION FOR WHOM ACUTE CHOLECYSTITIS SHOULD BE CONSIDERED

Patients with abdominal pain.

DETECTING THE LIKELIHOOD OF ACUTE CHOLECYSTITIS

See Table 12-4.

Table 12-4 Likelihood Ratios for Acute Cholecystitis

Finding (No. of Studies)	LR+ (95% CI)	LR- (95% CI)
Clinical gestalt ^a	≈25-30	
Murphy's sign (n = 3)	2.8 (0.8 to 8.6)	0.5 (0.2 to 1.0)
Right upper quadrant tenderness (n = 4)	1.6 (1.0 to 2.5)	0.4 (0.2 to 1.1)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

^aThe LR is imputed from the baseline pretest probability (5%), the sensitivity and specificity of ultrasonography (0.88 and 0.80, respectively), and the false-positive rate of diagnosis.

REFERENCE STANDARD TESTS

Surgical findings combined with pathology or clinical follow-up in patients who do not undergo surgery remain the reference standard for acute cholecystitis.

REFERENCES FOR THE UPDATE

- Shojania KG, Bero LA. Taking advantage of the explosion of systematic reviews: an efficient MEDLINE search strategy. *Eff Clin Pract.* 2001;4(4):157-162.
- Romero Y, Thistle JL, Longstreth GF, et al. A questionnaire for the assessment of biliary symptoms. *Am J Gastroenterol.* 2003;98(5):1042-1051.^a
- Rosen CL, Brown DF, Chang Y, et al. Ultrasonography by emergency physicians in patients with suspected cholecystitis. *Am J Emerg Med.* 2001;19(1):32-36.^a
- Kendall JL, Shimp RJ. Performance and interpretation of focused right upper quadrant ultrasound by emergency physicians. *J Emerg Med.* 2001;21(1):7-13.
- Lanoix R, Leak LV, Gaeta T, Gernsheimer JR. A preliminary evaluation of emergency ultrasound in the setting of an emergency medicine training program. *Am J Emerg Med.* 2000;18(1):41-45.
- Durston W, Carl ML, Guerra W. Patient satisfaction and diagnostic accuracy with ultrasound by emergency physicians. *Am J Emerg Med.* 1999;17(7):642-646.
- Schlager D, Lazzareschi G, Whitten D, Sanders AB. A prospective study of ultrasonography in the ED by emergency physicians. *Am J Emerg Med.* 1994;12(2):185-189.
- Jehle D, Davis E, Evans T, et al. Emergency department sonography by emergency physicians. *Am J Emerg Med.* 1989;7(6):605-611.

^aFor the Evidence to Support the Update for this topic, see <http://www.JAMAevidence.com>.

This page intentionally left blank

EVIDENCE TO SUPPORT THE UPDATE:

12

Cholecystitis

TITLE A Questionnaire for the Assessment of Biliary Symptoms.

AUTHORS Romero Y, Thistle JL, Longstreth GF, et al.

CITATION *Am J Gastroenterol.* 2003;98(5):1042-1051.

QUESTION What are the reproducibility, concurrent validity, and discriminative ability of a questionnaire designed to elicit patients' self-reported biliary symptoms?

DESIGN Prospective, independent, consecutive sample of blinded patients and investigators.

SETTING Referral gastroenterology practice at a major teaching institution, Rochester, Minnesota.

PATIENTS Two hundred forty-five adults (aged ≥ 18 years) referred to an outpatient clinic.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

A Biliary Symptoms Questionnaire (BSQ) was developed according to a review of the literature and the experience of the investigators using previously developed questionnaires for irritable bowel syndrome (IBS) and gastroesophageal reflux disease (GERD) as templates. The 114-question instrument was administered to subjects on initial presentation and then again after a 2-week interval. After the initial survey, subjects also underwent a structured interview conducted by investigators, who then completed their own BSQ according to the interview findings. Finally, investigators reviewed 10 BSQs of patients with known diagnoses (as determined by clinical follow-up and gastroenterologist opinion) and decided whether IBS, GERD, or biliary disease was the most likely diagnosis. A shortened BSQ was tested for reproducibility.

MAIN OUTCOME MEASURES

Agreement was expressed as simple agreement (%) and agreement beyond chance (κ). The domains assessed were as follows:

1. Agreement between the serial surveys administered to the patient (reproducibility)
2. Agreement between patient-reported symptoms and physician-reported symptoms (concurrent validity)
3. Agreement between investigator diagnosis according to the BSQ and gastroenterologist clinical diagnosis (discriminative validity)

MAIN RESULTS

Patients exhibited reasonable consistency throughout the 2-week test-retest period (see Table 12-5). In addition, physicians concurred with patients' self-reported symptoms with moderate or better agreement. Patient reproducibility and physician concurrence were almost perfect for complaints of upper abdominal pain ($\kappa = 0.94$ for both) and for jaundice ($\kappa = 0.94$ and $\kappa = 0.84$ for reproducibility and concurrence, respectively). Moderate agreement was observed for radiation of the pain ($\kappa = 0.47$ and 0.46 for reproducibility and concurrence, respectively). For fever, patients reported the symptom with substantial reproducibility ($\kappa = 0.79$), but physicians concurred with only moderate agreement ($\kappa = 0.52$). Although the questionnaire performed reasonably well in terms of discriminative ability ($\kappa = 0.58$), the limited sam-

Table 12-5 Questionnaire Results for Reproducibility and Concurrent Validity

	Reproducibility, κ (95% CI)	Concurrent Validity, κ (95% CI)
Emesis	0.95 (0.85 to 1)	0.73 (0.60 to 0.87)
Jaundice	0.94 (0.83 to 1)	0.84 (0.71 to 0.97)
Pain in upper abdomen	0.94 (0.83 to 1)	0.94 (0.86 to 1)
Nausea	0.81 (0.65 to 0.98)	0.75 (0.61 to 0.88)
Fever	0.79 (62 to 0.97)	0.52 (0.36 to 0.68)
Biliary symptoms ^a	0.72 (-0.03 to 0.95)	0.64 (0.15 to 0.95)
Radiation to right shoulder	0.47 (-0.15 to 1)	0.46 (0.21 to 0.72)

Abbreviation: CI, confidence interval.

^aThe results for biliary symptoms reflect the median agreement across all 18 questions identified as biliary (as opposed to gastroesophageal reflux disease or irritable bowel syndrome), including stabbing upper abdominal pain, cramping upper abdominal pain, radiation to the back, radiation to the shoulder blade, periodicity of pain episodes, daytime or nocturnal occurrence, and pain improved with movement, among others.

ple size (only 10 patients) and presentation of only 3 diagnostic choices (biliary colic, GERD, and IBS) severely limit this finding.

CONCLUSIONS

LEVEL OF EVIDENCE Level 1 for reproducibility and concurrent validity. Level 4 for discriminative validity (nonindependent sample with small numbers).

STRENGTHS The reproducibility and concurrent validity sections were well designed.

LIMITATIONS The study was designed primarily to assess the utility of a questionnaire as a research tool rather than to assess the variability in patient and physician reporting of abdominal symptoms. In testing the discriminative validity of the questionnaire, a small sample (10) of patients was used. In addition, investigators were given only 3 possible diagnoses to choose from—biliary pain, GERD, and IBS—which likely resulted in a significant overestimation of the discriminative ability of the questionnaire. The shortened BSQ was tested only for reproducibility, not concurrent validity or discriminative ability.

This study evaluated the reproducibility and concurrent validity of a questionnaire aimed at evaluating those with possible biliary colic. Although a few conclusions may be inferred regarding the variability in reporting of abdominal symptoms, the main focus of the study was to validate the instrument as a research tool. Patients appeared to be reasonably consistent in reporting most abdominal symptoms over time, and physicians generally concurred in their assessments of patients' symptoms.

Reviewed by Robert L. Trowbridge, MD

TITLE Ultrasonography by Emergency Physicians in Patients With Suspected Cholecystitis.

AUTHORS Rosen CL, Brown DF, Chang Y, et al.

CITATION *Am J Emerg Med.* 2001;19(1):32-36.

QUESTION How well do the assessments of emergency physicians using bedside ultrasonography (BUS) agree with the results of formal ultrasonography and clinical follow-up in the evaluation of suspected cholecystitis?

DESIGN Prospective, independent, convenience sample.

SETTING Emergency department at a major teaching hospital, Boston, Massachusetts.

PATIENTS One hundred sixteen adults (aged ≥ 18 years) who presented with abdominal pain and were suspected of having cholecystitis.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

Fifteen full-time emergency physicians underwent a 5-hour course, including didactic learning and hands-on training, on the use of an ultrasonographic machine to identify the gallbladder, detect gallstones, and elicit a sonographic Murphy sign.

The bedside ultrasonographic findings were compared not only with formal ultrasonography by radiologists but also clinical follow-up, including the results of other non-invasive tests for cholecystitis, operative reports, pathology, and use of telephone follow-up 1 month after emergency department visit to ascertain subsequent episodes of abdominal pain requiring medical attention emergency visits.

MAIN OUTCOME MEASURES

Agreement between BUS and formal ultrasonography in the detection of gallstones or presence of sonographic Murphy sign (ie, sensitivity and specificity of bedside ultrasonography, using formal ultrasonography as reference standard).

MAIN RESULTS

Among 116 patients, the physician performing BUSs could not visualize the gallbladder adequately in 6 (5.2%) cases. Four of these 6 cases were diagnosed as cholecystitis on formal ultrasonography. The authors explicitly state their interest in focusing on cases in which bedside ultrasonography appears to provide a definitive answer. Definitive BUS results were defined as both findings present or both absent (ie, both gallstones and sonographic Murphy sign present or both absent). Of the 116 patients, 70 (60%) had definitive findings (see Table 12-6). Although we do not show it here, the

Table 12-6 Likelihood Ratio of Bedside Ultrasonography

Test	Sensitivity	Specificity	LR+ (95% CI)	LR- (95% CI)
Definitive bedside ultrasonography compared with clinical follow-up for detection of cholecystitis	91%	66%	2.7 (1.7-4.1)	0.13 (0.04-0.39)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

authors presented data on the sensitivity and specificity of formal ultrasonography among patients with definitive results on BUSs. The negative likelihood ratio was similar to that above, but the positive likelihood ratio was much higher, at 14.

CONCLUSIONS

LEVEL OF EVIDENCE Level 3, since not all patients referred for formal ultrasonography were selected to undergo BUS.

STRENGTHS Radiologists performed formal ultrasonography without knowing the results of bedside ultrasonography. Distinct comparisons with formal ultrasonography and clinical follow-up provide useful information because cases not detected by formal ultrasonography would not be expected to be detected by BUS. Appropriately designed analysis, including adjusting for clustering effects.

LIMITATIONS Convenience sample. It was not clear how clinicians decided to choose which patients they referred for right upper quadrant ultrasonography. Unconsciously or not, physicians may have selected cases in which bedside ultrasonography was likely to perform well. The 3 physicians with the most training and previous experience were investigators in the study, and they contributed almost half

of the patients. The remaining physicians each contributed 10 or fewer patients; 2 physicians contributed only 1 patient each.

This study evaluated the potential effect of performing BUS on requests for formal ultrasonography to evaluate suspected acute cholecystitis. The limitations of the study (above) are important, but other well-designed aspects of the design and presentation of the results allow us to draw some reasonable conclusions.

Physicians with brief training and moderate experience in bedside ultrasonography can adequately visualize the gallbladder in the majority of patients (95% in this study). Approximately 60% of patients had definitive BUS results, defined as the presence of both cholelithiasis and a sonographic Murphy sign. Among such patients, the positive predictive value of only 70% means positive results require confirmation with formal ultrasonography. The negative predictive value is 90%. The authors point out that, for 30 patients, bedside ultrasonography detected no sonographic Murphy sign and no cholelithiasis. Had these patients not been sent for formal ultrasonography, there would have been a 26% reduction in ultrasonographic use by the emergency department, at a cost of missing 1 case of cholecystitis.

Reviewed by Kaveh G. Shojania, MD

This page intentionally left blank

CHAPTER 13

Does the Clinical Examination Predict Airflow Limitation?

Donald R. Holleman Jr, MD

David L. Simel, MD, MHS

CLINICAL SCENARIOS—DO THESE PATIENTS HAVE AIRFLOW LIMITATION?

In each of the following cases, the clinician needs to decide whether the patient has airflow limitation. In case 1, a 63-year-old man who has smoked 2 packs of cigarettes per day for the past 47 years presents with decreased exercise tolerance caused by shortness of breath. In case 2, a 35-year-old woman complains of coughing, wheezing, and shortness of breath every autumn. In case 3, an 18-year-old man is brought to an emergency department, with extreme difficulty breathing that began earlier that evening.

WHY IS IT IMPORTANT TO DETECT AIRFLOW LIMITATION BY CLINICAL EXAMINATION?

Airflow limitation is a disorder known by many names, including airway obstruction and obstructive airways disease. Recognizing airflow limitation can lead to appropriate treatment and can yield important prognostic information. Patients with symptomatic airflow limitation may benefit by treatment with oral or inhaled bronchodilators, oral or inhaled glucocorticoids, or antibiotics. Recognition of this disorder also triggers environmental controls and preventive services, such as vaccination against pneumococcus and influenza.

Screening is advocated for target disorders in which early intervention favorably affects patient outcomes. Physicians do not screen for airflow limitation because early intervention has not been shown to alter the disease course. Therefore, clinicians are likely to want to confirm or rule out disease in patients presenting with pulmonary symptoms, such as cough or dyspnea, rather than screen for unrecognized disease in asymptomatic individuals.

The 3 clinical scenarios illustrate cases in which recognizing airflow limitation by the clinical examination is important. In the first case, recognizing airflow limitation might lead to the diagnosis of pulmonary emphysema, more intensive counseling on smoking cessation, vaccination against influenza and pneumococcal infection, and bronchodilator therapy to improve exercise tolerance. In the second case, recognizing airflow limitation might lead to the identification of environmental irritants or allergens responsible for symptoms. In the third case, recognizing airflow limitation would lead to the diagnosis of asthma and to acute, potentially lifesaving therapy with bronchodilators and systemic glucocorticoids. Recognizing airflow limitation clinically may have time, cost, and convenience advantages compared to routine pulmonary function testing.

Spirometry is the test of choice for confirming a diagnosis of airflow limitation. Both the forced expiratory volume in 1 second (FEV_1) and forced vital capacity (FVC) values are reduced in patients with airflow limitation; because the FEV_1 is affected more than the FVC, the ratio of FEV_1 to FVC (FEV_1/FVC) also

decreases. The reduced FEV₁/FVC is the hallmark of airflow limitation. Although emphysema and chronic bronchitis represent permanent reductions in airflow, asthma is a disorder characterized by increased responsiveness of the bronchial tree to a variety of stimuli, leading to intermittent airflow limitation.¹ In patients with asthma, provocative testing, such as methacholine challenge, may be necessary to bring about airflow limitation between symptomatic episodes.

The reference standard for airflow limitation is the measurement of the FEV₁ and the FVC by spirometry. An FEV₁/FVC lower than the fifth percentile for age, height, and sex is considered abnormal.² However, a normal FEV₁/FVC during an asymptomatic period does not rule out intermittent airflow limitation. For most patients, the fifth percentile of FEV₁/FVC is approximately 70%, but using this single value to diagnose airflow limitation is discouraged.²

We performed an English-language MEDLINE search, using the following Medical Subject Headings: (EXP Medical History Taking OR EXP Physical Examination) AND (EXP Lung Diseases, Obstructive). The titles and abstracts of the 1022 articles retrieved from the above MEDLINE search were reviewed independently by the 2 authors. If either reviewer chose an article as possibly useful, the article was reviewed for content. The authors had excellent agreement ($\kappa = 0.85$) on the 158 articles chosen for full review. If the article contained results of the clinical examination predicting airflow limitation, the article was reviewed for quality. References from appropriate articles were reviewed for additional references. Nineteen articles evaluating the clinical examination for airflow limitation³⁻²¹ used the accepted definition or a similar spirometric definition of disease. Others used a variety of definitions, including FEV₁ only²²⁻²⁷ or other, less-accepted or unclear definitions.²⁸⁻³⁷ We chose to include articles using reference standards that are not currently accepted because they were otherwise methodologically sound or they provided the only data available for some of the clinical

examination findings. The reference standards used in studies evaluating operating characteristics for individual clinical examination items are listed in Table 13-1. Because all studies used reference standards of current airflow limitation, the results in this review can be used only to predict airflow limitation at the evaluation. Patients with asthma may be overlooked if examined between attacks.

PATHOPHYSIOLOGIC CHARACTERISTICS OF AIRFLOW LIMITATION

Understanding the physiologic characteristics of pulmonary airflow helps to explain the clinical examination findings in airflow limitation. The airways are a branching system of tubes that link the outside atmosphere with the lung parenchyma. During inspiration, the thoracic cavity actively expands. As the chest volume increases, the intrathoracic pressure decreases. Because the airways are open to the atmosphere, air flows into the airways to equalize the intrathoracic pressure with the atmospheric pressure. Therefore, during inspiration, the pressure inside the airways is greater than the pressure in the surrounding lung. This pressure exerts a force on the inner wall of the airway, increasing the airway diameter during inspiration.

At end inspiration, the chest no longer expands, and the intrathoracic-to-atmospheric pressure difference disappears. During expiration, the thoracic cavity passively contracts. As the chest volume decreases, the intrathoracic pressure increases and exceeds the atmospheric pressure. Because the airways communicate with the atmosphere, the pressure inside the airways is lower than the pressure in the surrounding lung. This pressure difference exerts a force on the outer wall of the airway, decreasing the airway diameter during expiration. The resistance to airflow is inversely and exponentially related to the diameter of the airway, so small decreases in airway diameter lead to large increases in resistance.

During inspiration and expiration, the diameter of the airway varies around its static, resting diameter. In airflow limitation, the resting airway diameter is abnormally small. In emphysema, the lung parenchyma is destroyed. This leads to a decrease in the tethering forces that maintain airway diameter, resulting in decreased resting airway diameter. In asthma, the smooth muscle that surrounds the airway is hyperreactive to various stimuli. When one of these stimuli is present, the smooth muscle contracts. This leads to decreased resting diameter of the airway. In chronic bronchitis, there is increased mucus production in the airways. There may also be decreased mucus clearance caused by ciliary dysfunction. The resulting increased intra-airway mucus coats the inner wall of the airway. This leads to decreased resting diameter of the airway. Thus, in airflow limitation syndromes, the resistance to airflow is increased throughout the respiratory phase. Because of the further physiologic decrease in airway diameter during expiration, it is significantly more difficult to empty the lungs than to fill them. This leads to air trapping and to lung hyperinflation that can be demonstrated by an abnormally large residual volume on pulmonary function testing.

The touted physical examination findings for airflow limitation arise either from the difficulty in emptying the lungs or from the resulting hyperinflation. The prolonged expiratory

Table 13-1 Reference Standards Used in Studies Yielding Operating Characteristics for Individual Clinical Examination Items

Reference Standard	References
FEV ₁ < fifth percentile and FEV ₁ /FVC < fifth percentile ^a	14
FEV ₁ /FVC < fifth percentile	11
FEV ₁ /FVC < 0.70	5-8, 16, 18, 22
FEV ₁ /FVC < 0.75 and FVC < 80% of predicted	9
FEV ₁ < 75% of predicted and FEV ₁ /FVC < 0.80	20
FEV ₁ < 70% of predicted	23, 24
FEV ₁ < 2 L	25, 26
FEV ₁ < fifth percentile	37
FEV ₁ < 60% of predicted or FEV ₁ /FVC < 0.60	17
Roentgenography, total lung capacity, and residual capacity	33
FEV ₁ /FVC < 0.6 or history	31
Diagnosis of asthma	32
Normal spirometry	30

Abbreviations: FEV₁, forced expiratory volume in 1 second; FVC, forced vital capacity.

^aThe definition recommended by the American Thoracic Society.¹

phase, wheezing, rhonchi, and match test are signs of abnormally high resistance to airflow during expiration. Decreased breath sounds, barrel chest, hyperresonance, decreased cardiac and hepatic dullness, absent or subxiphoid cardiac apical impulse, decreased chest expansion, and decreased diaphragmatic movement are signs of hyperinflation. Use of accessory muscles results from both the increased work of expiration and pulmonary hyperinflation.

HOW TO ELICIT SYMPTOMS AND SIGNS OF AIRFLOW LIMITATION

A concise evaluation for airflow limitation includes a focused medical history and physical examination.

History

The history should elicit background features and specific symptoms.

Background Information

The most important background features are exposure to cigarette smoke and to occupational or environmental pollutants. The duration of cigarette exposure can most easily be elicited by asking at what age the patient started smoking and in what year he or she quit. Although pack-years is the traditional measure of cigarette exposure, quantifying years of exposure works at least as well.¹³ The patient's personal and family history of atopic diseases is also associated with increased likelihood of asthma.

Symptoms

The most important symptoms to elicit from patients with suspected airflow limitation are wheezing, coughing, and sputum production. In fact, chronic bronchitis is defined by sputum production for at least 3 consecutive months in at least 2 consecutive years.¹

Physical Examination

The physical examination for airflow limitation should include inspection, measuring vital signs, palpation, percussion, auscultation, and expiratory airflow.

Inspection

While assessing the patient's overall appearance, the clinician should observe for the presence of a barrel chest. If the anteroposterior diameter appears greater than normal, the patient has a barrel chest deformity. This finding may be more an illusion than a true deformity because the antero-posterior dimensions have not been shown to be increased in patients with clinically defined barrel chests.³⁸

Vital Signs

While measuring blood pressure, the clinician can determine whether there is pulsus paradoxus. This maneuver may be most helpful in patients with suspected acute airflow limitation. During tidal breathing, the sphygmomanometer is inflated to above the systolic blood pressure. The cuff pressure is slowly released until the first Korotkoff sound is heard only during expiration;

this systolic blood pressure value is noted. The cuff pressure is further reduced until the first Korotkoff sound is heard throughout inspiration; the systolic blood pressure at this point is also noted. The systolic blood pressure is normally lower during inspiration than during expiration. The normal difference is accentuated when the patient has airflow limitation. If the difference between these 2 pressures is at least 15 mm Hg, the patient has pulsus paradoxus.

Palpation

Palpation should include locating the cardiac apical impulse. Chest palpation should be performed with the patient supine and disrobed from the waist up. A sheet or gown should be used to maintain patient comfort and privacy; however, palpation should be performed with the hand directly on the chest wall. When the chest volume is increased because of hyperinflation, the cardiac apex shifts to a more central location and either may not be palpable or may be palpable in the subxiphoid area.

Percussion

The chest should be percussed to determine the quality of the sound that resonates. Percussion of the chest wall should be performed by placing a digit (usually the second or third) of the nondominant hand firmly against the chest wall parallel to and between the ribs. The second and third digits of the dominant hand are flexed slightly at the metacarpophalangeal and proximal and distal interphalangeal joints to form a slight arch with the 2 fingertips even. The fingertips of the dominant hand tap the distal interphalangeal joint of the nondominant hand with a firm pecking motion. If the sound is more hollow than normal, the chest is hyperresonant.

Auscultation

Clinicians should auscultate the chest for wheezes, rhonchi, and breath sound intensity. Chest auscultation should be performed in a quiet room with the patient disrobed from the waist up. The warmed stethoscope diaphragm should be placed with moderate pressure on the patient's chest to ensure good sound transmission. The chest should be auscultated bilaterally over the lower, middle, and upper lung fields posteriorly, anteriorly, and along the midaxillary line. Patients should be breathing heavily, but not forcefully. Wheezing will be heard as high-pitched musical tones especially during expiration. Rhonchi are lower-pitched wheezes.³⁹ The intensity of breath sounds should be observed. Although elaborate scoring systems for breath sound intensity^{9,26} and for wheezing¹⁶ have been developed, they are not clearly better than the customary normal vs abnormal dichotomization.

Measures of Airflow

Measures of expiratory airflow include the forced expiratory time^{13,17} and the match test.^{16,24,25} To perform a forced expiratory time test, the patient must take a deep breath and forcefully exhale until no more air can be expelled. During this maneuver, the patient must keep mouth and glottis fully open as if the patient were yawning. While the patient is performing the forced expiration, the clinician listens over the larynx or lower trachea with a stethoscope and times the duration of audible airflow. To obtain the best results, the forced expiratory time should be measured with a stopwatch and recorded to the nearest 0.1 second. An alternative

maneuver is the match test. During this test, the patient performs a forced expiration exactly as in the forced expiratory time maneuver. However, the clinician holds a burning match 10 cm from the patient's widely open mouth. If the match is still burning after the forced expiration, the test result is positive. Others have used a candle for this test. However, one needs a match to light a candle, and we can find no benefit in carrying around both except for those who frequently practice in the dark. Also, to avoid malpractice claims and personal injury, we do not recommend this test in patients receiving supplemental oxygen!

PRECISION OF HISTORY AND SYMPTOMS FOR AIRFLOW LIMITATION

The observer agreement for smoking history, dyspnea, coughing, wheezing, chronic bronchitis, and orthopnea has been described with the κ statistic.¹³ Two physicians almost always agree on the smoking history ($\kappa = 0.95$). Physicians agree frequently on the presence or absence of wheezing ($\kappa = 0.61$), chronic bronchitis ($\kappa = 0.55$), dyspnea ($\kappa = 0.44-0.48$), and coughing ($\kappa = 0.46$).

ACCURACY OF MEDICAL HISTORY AND SYMPTOMS FOR AIRFLOW LIMITATION

Table 13-2 summarizes the operating characteristic estimates for airflow limitation, obtained for each historical item and symptom, after pooling data from referenced studies.

Background Information

The best background information for diagnosing airflow limitation is exposure to cigarette smoke. Although patients

who have smoked are only slightly more likely to have airflow limitation,^{5,6,13} never having smoked cigarettes is moderately well associated with decreased likelihood of disease.^{5,6,13} Perhaps more useful is the fact that the number of years the patient has smoked correlates well with the likelihood of disease (Figure 13-1).¹³ Patients with at least a 70-pack-year history of smoking are much more likely to have airflow limitation.¹⁶

Age is related to airflow limitation. Asthma is more common in the young, whereas chronic bronchitis and emphysema are more common in older patients. The prevalence of airflow limitation appears to be lowest between ages 10 and 30 years.⁴⁰ The higher prevalence at younger ages is due to asthma, which frequently remits after childhood. The higher prevalence in the older age group is probably due to 2 factors. First, age is a proxy for exposure to toxins, especially cigarette smoke. When smokers and nonsmokers are analyzed separately, the prevalence of airflow limitation does not appear to increase significantly with age in nonsmokers.⁴¹ Second, in adults, most airflow limitation is a chronic disease, so new incident cases are added faster than attrition from mortality, except in the very old. Therefore, advancing age is associated with increased likelihood of airflow limitation in adult smokers, but airflow limitation should not be considered a normal process of aging.

Symptoms

Symptoms of chronic bronchitis,^{13,19} sputum production of at least one-fourth of a cup when present,¹⁶ or wheezing^{13,36} are associated with a moderate increase in the likelihood of airflow limitation. However, symptoms of cough^{5,13} or exertional dyspnea^{13,36} are associated with only a slight increase in the likelihood of airflow limitation. Orthopnea is not useful in diagnos-

Table 13-2 Composite Operating Characteristics of History Items Predicting Airflow Limitation

Item	Grade of Recommendation ^a	References	Sensitivity, %	Specificity, %	LR+	LR-
<i>Smoking history</i>						
≥70 vs <70 pack-years	B	17	40	95	8.0	0.63
Ever vs never	A	6, 7, 14	92	49	1.8	0.16
Sputum production ≥ 1/4 cup	B	17	20	95	4	0.84
Symptoms of chronic bronchitis	A	14, 20	30	90	3.0	0.78
Wheezing	B	14	51	84	3.8	0.66
<i>Exertional dyspnea</i>						
Grade 4 vs 3 or less	A	20	03	99	3.0	0.98
Any vs none	A	20	27	88	2.2	0.83
Coughing	B	14	51	71	1.8	0.69
Any dyspnea	B	14	82	33	1.2	0.55

Abbreviations: LR+, positive likelihood ratio; LR-, negative likelihood ratio.

^aThe recommendation grading scheme was provided by David L. Sackett, MD, and Charles H. Goldsmith, PhD. Grade A: independent, blind comparison of sign or symptom with a gold standard of diagnosis among a large number of consecutive patients suspected of having the target condition. Grade B: independent, blind comparison of sign or symptom with a gold standard of diagnosis among a small number of consecutive patients suspected of having the target condition. Grade C: independent, blind comparison of sign or symptom, with a gold standard of diagnosis among nonconsecutive patients suspected of having the target condition; or nonindependent comparison of sign or symptom with a gold standard of diagnosis among samples of patients who obviously have the target condition plus, perhaps, individuals with normal results; or nonindependent comparison of sign or symptom with a standard of uncertain validity (see Table 1-7).

ing airflow limitation, because its positive likelihood ratio ($LR+$) and negative likelihood ratio ($LR-$) are not significantly different from 1.¹³ No single symptom effectively rules out airflow limitation. The absence of dyspnea^{5,13,36} or of exertional dyspnea^{13,36} is only moderately useful in ruling out disease.

PRECISION OF THE SIGNS OF AIRFLOW LIMITATION

κ Statistics or correlation coefficients have generally been used to describe the precision of physical examination items for airflow limitation.^{13,16,17,28,34,35,37}

Inspection

Precision has not been studied for most inspection items, and physicians agree only part of the time that a patient has a cough ($\kappa = 0.29$),¹³ which can probably be explained largely by patients having paroxysms of coughing. They may cough during one, but not the other, examination.

Vital Signs

The precision of pulsus paradoxus has not been well studied.

Palpation

Physicians agree only part of the time on the results of palpating for an absent apical impulse ($\kappa = 0.39$).³⁴ Physician agreement on whether a patient has a subxiphoid apical impulse may be no greater than chance ($\kappa = 0.0-0.3$).^{13,16} However, the low prevalence of this finding may lead to underestimating the chance-corrected agreement.

Percussion

Physicians appear to agree infrequently on the results of chest percussion. However, only hyperresonance ($\kappa = 0-0.42$)^{16,37} and diaphragmatic excursion ($\kappa = -0.04$; $r = 0.24$) have been studied.^{16,35}

Auscultation

Physicians agree frequently on the results of auscultation for wheezing ($\kappa = 0.43-0.93$),^{13,16,37} whereas they agree less frequently on breath sound intensity ($\kappa = 0.23-0.47$)^{13,16,28,37} and crackles ($\kappa = 0.30-0.63$).³⁷

Measures of Airflow

Physicians frequently obtain the same results when measuring forced expiratory time (intraclass correlation, 0.81; $\kappa = 0.7$)^{13,17} or interpreting the match test ($\kappa = 0.39$).¹⁶ Agreement on the forced expiratory time is better if a stopwatch is used instead of a second hand.

ACCURACY OF THE SIGNS OF AIRFLOW LIMITATION

Table 13-3 summarizes the operating characteristic estimates for airflow limitation, obtained for each sign, after pooling data from referenced studies.

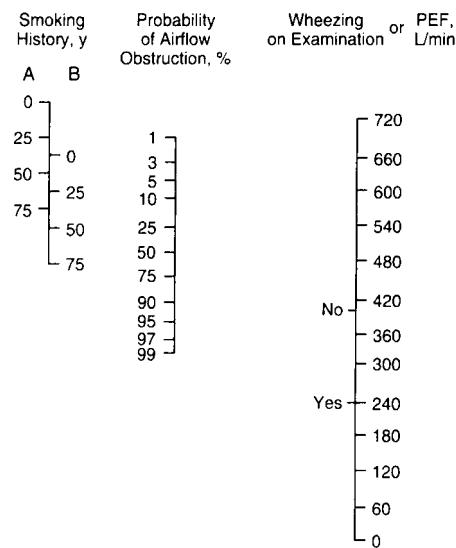


Figure 13-1 Predicting Probability of Airflow Obstruction at the Bedside

Choose the number of years the patient smoked cigarettes under the "Smoking History" heading; use scale A if the patient reports no symptoms of wheezing or scale B if the patient reports symptoms of wheezing. Under "Wheezing on Examination," select "No" if wheezing was absent or "Yes" if wheezing was present (alternatively, the best of 3 peak expiratory flow [PEF] rates could be chosen under the "PEF" heading). With a straightedge, connect the points chosen on the "Smoking History" and "Wheezing on Examination" lines. Read the probability of airflow limitation where the straightedge intersects the line under the "Probability of Airflow Obstruction" heading.

Reprinted from Hollerman et al,¹³ with the permission of the *Journal of General Internal Medicine*.

Inspection

A barrel chest^{31,32} predicts airflow limitation. However, the evidence for this association comes largely from one study in asthmatic children. Recent studies using currently accepted reference standards have failed to include this finding. Therefore, the value of the barrel chest sign in adults is not well supported. Other inspection items (accessory muscle use, excavated supraclavicular fossae, and coughing) have not been studied in a large enough sample of patients to determine the extent of their usefulness in diagnosing airflow limitation.^{13,32,36} In other words, their likelihood ratio confidence intervals are wide and include 1.⁴² Decreased chest expansion and kyphosis have been studied only in patients with known disease,³² so their usefulness has not yet been determined. Patients who do not use accessory muscles^{32,36} or who do not have excavated supraclavicular fossae are only slightly less likely to have airflow limitation.³⁶ Patients without a barrel chest^{31,32} or who do not cough¹³ are significantly less likely to have airflow limitation but the clinical importance of the absence of these findings is negligible. Therefore, the only inspection item we can recommend is looking for a barrel chest. The presence of this finding, especially in children, virtually rules in airflow limitation.

Table 13-3 Composite Operating Characteristics of Physical Examination Items Predicting Airflow Limitation

Item ^a	Grade of Recommendation ^b	References	Sensitivity, %	Specificity, %	LR+	LR-
Wheezing	A	14, 17, 34	15	99.6	36	0.85
Barrel chest	B ^c	32	10	99	10	0.90
Decreased cardiac dullness	B	17	13	99	10	0.88
Match test	B	17, 25, 26	61	91	7.1	0.43
Rhonchi	B	31, 32	8	99	5.9	0.95
Hyperresonance	B	17	32	94	4.8	0.73
Forced expiratory time, s ^d	A	14, 18				
>9					4.8	
6-9					2.7	
<6					0.45	
Subxiphoid cardiac apical impulse	B	14, 17	8	98	4.6	0.94
Pulsus paradoxus (>15 mm Hg)	C	8, 23, 24	45	88	3.7	0.62
Decreased breath sounds	B	14, 17	37	90	3.7	0.70
Accessory muscle use	C	33, 37	24	100	e	0.70
Excavated supraclavicular fossae	C	37	31	100	e	0.69

Abbreviations: LR+, positive likelihood ratio; LR-, negative likelihood ratio.

^aListed in order of decreasing LR+.

^bSee Table 1-7 for a summary of Evidence Grades and Levels.

^cThis recommendation includes only children. Sensitivity in adults was 4% (grade C).¹²

^dBecause the forced expiratory time test has 3 levels, LR-, sensitivity, and specificity cannot be calculated.

^eThis item was studied in too few subjects to yield meaningful results.

Vital Signs

The presence of pulsus paradoxus of at least 15 mm Hg is associated with only a moderate increase in the likelihood of airflow limitation, and the absence of this sign is associated with only a slight reduction in the likelihood of disease.^{7,22,23} Other vital signs have not been studied and cannot be recommended for use in determining the likelihood of airflow limitation.

Palpation

Palpating a subxiphoid cardiac apical impulse is associated with a moderate increase in the likelihood of airflow limitation. However, the absence of this finding is not useful.^{13,16} Absent apical impulse has been studied only in patients with known disease,³² so its usefulness has not yet been determined. Therefore, according to current evidence, we recommend palpating the subxiphoid region for the cardiac apical impulse. We recommend this despite the reportedly low observer agreement because the low prevalence of this finding may lead to underestimates of the chance-corrected agreement.

Percussion

Chest hyperresonance on percussion is associated with a moderate increase in the likelihood of disease.^{16,32} Neither decreased cardiac dullness nor decreased diaphragmatic movement has been studied in enough patients to determine

definitively the extent of usefulness.¹⁶ However, patients with decreased cardiac dullness are more likely to have airflow limitation. Decreased liver dullness has been studied only in patients with known disease,³² so its usefulness has not yet been determined. Patients without chest hyperresonance are only slightly less likely to have airflow limitation.^{16,32} Normal cardiac dullness and normal diaphragmatic movement are likely not useful for decreasing the likelihood of airflow limitation.¹⁶ We recommend percussing the chest for the resonance sound. Hyperresonance over the precordium may be particularly useful for increasing the likelihood of airflow limitation.

Auscultation

Objective wheezing, or wheezing observed on physical examination, is the most potent predictor of airflow limitation. Patients with wheezing almost certainly have airflow limitation.^{13,15,16,37} However, this is true only of wheezing on unforced expiration. Forced expiration is associated with increased sensitivity of wheezing, and with decreased specificity. The current literature suggests that the presence or absence of wheezing on forced expiration is of no value in diagnosing or ruling out airflow limitation.^{15,20} Additionally, the sensitivity of wheezing increases with the severity of airflow limitation.¹³ Studies that recruited patients referred for spirometry^{15,36} yielded sensitivities greater than those found in unreferred populations.^{13,16} Although the sensitivity of wheezing varies greatly (10%-50%) by study population,

the LR+ and LR- change little. Rhonchi were associated with a moderate increase in the likelihood of airflow limitation in 2 studies^{30,31}; however, because neither study explicitly defined rhonchi and because there is significant variability in how physicians define rhonchi,⁴³ this result must be interpreted cautiously. Decreased breath sounds are associated with only a moderate increase in the likelihood of disease.^{13,16,32} Absent wheezing,^{13,15,16,36} normal breath sound intensity,^{13,16,32} or absent rhonchi^{30,31} are associated with only a moderate decrease in the likelihood of disease. We recommend auscultating the chest for wheezes and for breath sound intensity. Patients with wheezing should be considered to have airflow limitation, and patients with decreased breath sound intensity should be considered somewhat more likely to have airflow limitation. Patients without wheezing or with normal breath sound intensity should be considered somewhat less likely to have this disorder. Neither the presence nor absence of crackles (rales) helps with the diagnosis of airflow limitation.^{8,13,29}

Measures of Airflow

Patients who are unable to extinguish a lighted match held 10 cm from the open mouth are significantly more likely to have airflow limitation than patients who are able to extinguish a match. The ability to extinguish a match is associated with a moderate decrease in the likelihood of disease.^{16,24,25} The forced expiratory time^{4,5,10,11,13,16-18} is a continuous variable that can range from a few tenths of a second to more than 20 seconds. Unfortunately, each of the 4 best studies of forced expiratory time^{10,13,16,17} used different methods. Two studies^{10,16} used average expiratory time, which makes bedside use cumbersome. Of the other 2 studies, one used the shortest expiratory time of 3 trials;¹³ the other, the longest expiratory time of 2 trials.¹⁷ Because the ability to discriminate between patients with and without airflow limitation is the same regardless of whether the shortest or longest time is used,¹³ there is no clear advantage to one method over the other. To allow pooling of results, one of the studies¹³ was reanalyzed with the longest rather than the shortest time. When the longest expiratory time is chosen, a result less than 6 seconds was associated with a modest decrease in the likelihood of airflow limitation; a result between 6 and 9 seconds was associated with a modest increase in the likelihood of airflow limitation; and a result greater than 9 seconds was associated with a great increase in the likelihood of airflow limitation. A forced expiratory time of approximately 9 seconds predicts an FEV₁/FVC of 70%,⁸ a level suggesting the diagnosis of airflow limitation.

Peak expiratory flow rates predict airflow limitation (Figure 13-1).¹³ However, 2 studies have shown that peak expiratory flow adds little to the clinical examination for airflow limitation.^{13,16} In one study,¹⁶ peak expiratory results improved the accuracy of the clinical examination for only 1 of the 4 physicians studied. In the other study,¹³ peak expiratory flow was equivalent to auscultating for wheeze, but more difficult to assess. Therefore, we cannot recommend routine peak flow measurements in the diagnosis of airflow limitation. Peak flow

measurements may be useful in assessing benefit from therapy, especially for asthma.

CAN THE CLINICAL EXAMINATION PREDICT SEVERITY OF AIRFLOW LIMITATION?

Stubbing et al³ found that the number of positive findings (tracheal descent during inspiration, sternomastoid contraction, scalene contraction, supraclavicular fossae excavation, supraclavicular fossae recession, intercostal recession, or costal margin movement) predicted the severity of airflow limitation in patients with known disease. These findings tended to be present only if the FEV₁ was less than 50% of the predicted value. The American Thoracic Society^{1,2} found that the number of positive findings (barrel chest, low diaphragm, decreased diaphragmatic excursion, decreased breath sounds, prolonged expiratory phase, wheezing, noisy inspiration, or crackles) predicted the severity of airflow limitation ($r = 0.6$). The literature suggests that, as airflow becomes more limited, more physical examination findings become apparent.

ACCURACY OF THE OVERALL CLINICAL IMPRESSION FOR PREDICTING AIRFLOW LIMITATION

Three studies^{14,17,33} evaluated the accuracy of the overall clinical impression or a clinician's ability to integrate all aspects of the clinical examination in forming an impression about the likelihood of airflow limitation. Clinicians' overall impressions¹³ (graded as moderate to severe limitation [LR+ = 4.2], mild [LR+ = 0.82], or none [LR+ = 0.42]), predicted any airflow limitation only moderately well. However, Badgett et al¹⁶ found that clinicians' impressions (blinded to medical history but not physical examination) predicted moderate to severe airflow limitation somewhat better (LR+ = 7.3; LR- = 0.53) and about as well as some of the individual findings in Table 13-3. On the other hand, Fletcher³² evaluated the clinical impressions of 6 physicians and found sensitivities ranging from 15% to 95% for airflow limitation. Therefore, clinicians' ability to diagnose airflow limitation clinically is variable, but accuracy seems to improve as the severity of airflow limitation increases.

COMBINATIONS OF INDIVIDUAL FINDINGS

Six studies (Table 13-4) assessed the usefulness of combining clinical examination items to predict airflow limitation. Unfortunately, as with individual findings, combinations of findings do not effectively rule out airflow limitation. The best combination is never having smoked, no reported wheezing, and no wheezing on examination (Figure 13-1; LR-, 0.18).¹³ Other combinations have LR- values ranging from 0.33 to 0.77. Even the best combination is no better than smoking history alone (LR-, 0.16). Therefore, combinations of findings are more helpful for ruling in than for ruling out this disorder. In fact, a patient with any combination

Table 13-4 Combinations of Clinical Examination Items Predicting Airflow Limitation

Clinical Examination Item	Interpretation	Relation to Airflow Limitation	Reference Standard
Years of cigarette exposure, patient-reported wheezing, objective wheezing ¹³	See Figure 13-1	LR+ varies (see Figure 13-1) LR- = 0.18	FEV ₁ /FVC and FEV ₁ < fifth percentile
Patient-reported chronic obstructive pulmonary disease, ≥ 70 pack-years of cigarette smoking, decreased breath sounds ¹⁶	≥2 Findings present <2 Findings present	LR+ = 34 LR- = 0.34	FEV ₁ < 60% of predicted or FEV ₁ /FVC < 0.60
Dyspnea, subjective wheezing, objective wheezing, accessory muscle use, excavation of supraclavicular fossae, and distention of external jugular veins ²⁶	No. of findings present	$r = -0.64$	Ratio of FEV ₁ to predicted FEV ₁
Breath sound intensity, use of scalene muscle, objective wheezing, and rales during cough ²⁷	No. of findings present	Negatively correlated with FEV ₁	FEV ₁
Decreased breath sounds, objective wheezing, rales, and prolonged expiratory time ³³	All 4 findings present <4 Findings present	LR+ = 3.3 LR- = 0.44	Abnormal FVC, FEV ₁ , or maximal midexpiratory flow
History by questionnaire, standardized physical examination ²¹	Any abnormal finding No abnormal findings	LR+ = 1.4 LR- = 0.77	FEV ₁ /FVC < 0.70

Abbreviations: FEV₁, forced expiratory volume in 1 second; FVC, forced vital capacity; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

of 2 findings (≥ 70 -pack-year history of smoking, history of chronic obstructive pulmonary disorder, or decreased breath sounds) can be considered to have airflow limitation.¹⁶

CLINICAL SCENARIOS—RESOLUTIONS

CASE 1 In case 1, the patient reported a 47-year smoking history but no other environmental or occupational exposures. He complained of episodes of wheezing but had no wheezing on examination. According to Figure 13-1, he has a 65% chance of having airflow limitation.

CASE 2 In case 2, the patient was asymptomatic during the office visit. She had never smoked cigarettes and had no exposure to environmental or occupational pollutants. She did not have a barrel chest, her apical impulse was normally located, and her chest was not hyperresonant. Her breath sounds were normal in intensity, without wheezing or rhonchi. Her forced expiratory time was 2 seconds. Because of clinical examination findings, you conclude that she does not have airflow limitation at the office visit. However, because of her medical history, you suspect that she has intermittent airflow limitation secondary to environmental allergens.

CASE 3 In case 3, the patient had no smoking history or previous episodes of dyspnea. His chest was hyperresonant, and he had diffuse expiratory wheezes. His forced expiratory time was 12 seconds, and he had pulsus paradoxus of 32 mm Hg. You diagnose acute bronchospasm and begin appropriate bronchodilator and glucocorticoid therapy.

THE BOTTOM LINE

Guidelines for using the clinical examination to diagnose airflow limitation are as follows:

- No single item or combination of items from the clinical examination rules out airflow limitation. However, the

best finding associated with decreased likelihood of airflow limitation is a history of never having smoked cigarettes (especially in patients without a history of wheezing and without wheezing on examination).

- The best findings associated with increased likelihood of airflow limitation are objective wheezing, barrel chest, positive match test result, rhonchi, hyperresonance, forced expiratory time greater than 9 seconds, and subxiphoid apical impulse.
- A finding of a barrel chest (in children) or wheezing virtually rules in airflow limitation.
- Any 2 of the following virtually rule in airflow limitation: 70 pack-years or more of smoking, decreased breath sounds, or history of chronic obstructive pulmonary disorder.
- Three findings predict the likelihood of airflow limitation in men (Figure 13-1): years of cigarette smoking, subjective wheezing, and either objective wheezing or peak expiratory flow rate.

Author Affiliations at the Time of the Original Publication

Medical Service, Lexington Veterans Affairs Medical Center and Department of Medicine, University of Kentucky, Lexington (Dr Holleman); and the Center for Health Services Research in Primary Care, Durham Veterans Affairs Medical Center, and Department of Medicine and Center for Health Care Policy Research and Education, Duke University, Durham, North Carolina (Dr Simel).

Acknowledgments

This study was supported in part by the Andrew W. Mellon Foundation, New York, New York.

We thank Marilyn Schapira, MD, and Bob Badgett, MD, for providing raw data from their studies that allowed us to calculate composite operating characteristics. We also thank Bob Badgett, MD, and Joseph Govert, MD, for their careful reviews.

REFERENCES

1. American Thoracic Society. Standards for the diagnosis and care of patients with chronic obstructive pulmonary disease (COPD) and asthma. *Am Rev Respir Dis.* 1987;136(1):225-243.
2. American Thoracic Society. Lung function testing: selection of reference values and interpretative strategies. *Am Rev Respir Dis.* 1991;144(5):1202-1218.
3. Stubbing DG, Mathur PN, Roberts RS, Campbell EJM. Some physical signs in patients with chronic airflow obstruction. *Am Rev Respir Dis.* 1982;125(5):549-552.
4. Lal S, Ferguson AD, Campbell EJM. Forced expiratory time: a simple test for airways obstruction. *BMJ.* 1964;1(5386):814-817.
5. Scheinhorn DJ. Screening test for airflow limitation. *South Med J.* 1982;75(4):434-438.
6. Catlett GF, Kidera GJ. Detection of respiratory impairment in pilots. *Aerospace Med.* 1969;40(11):1252-1257.
7. Govindaraj M. Pulsus paradoxus in respiratory disease. *Indian J Chest Dis.* 1970;12(3):110-115.
8. Workum P, Holford SK, Delbono EA, Murphy RLH. The prevalence and character of crackles (rales) in young women without significant lung disease. *Am Rev Respir Dis.* 1982;126(5):921-923.
9. Bohadana AB, Peslin R, Uffholtz H. Breath sounds in the clinical assessment of airflow obstruction. *Thorax.* 1978;33(3):345-351.
10. Kern DG, Patel SR. Auscultated forced expiratory time as a clinical and epidemiologic test of airway obstruction. *Chest.* 1991;100(30):636-639.
11. Macdonald JB, Cole TJ, Seaton A. Forced expiratory time: its reliability as a lung function test. *Thorax.* 1975;30(5):554-559.
12. Van Schayck CP, van Weel C, Harbers HJM, van Herwaarden CLA. Do physical signs reflect the degree of airflow obstruction in patients with asthma or chronic obstructive pulmonary disease? *Scand J Prim Health Care.* 1991;9(4):232-238.
13. Holleman DR, Simel DL, Goldberg JS. Diagnosis of obstructive airways disease from the clinical examination. *J Gen Intern Med.* 1993;8(2):63-68.
14. Suprenant EL, Vance JW. Evaluation of methods for the early detection of chronic obstructive ventilatory diseases. *Dis Chest.* 1967;52(6):760-766.
15. Marini JJ, Pierson DJ, Hudson LD, Lakshminarayan S. The significance of wheezing in chronic airflow obstruction. *Am Rev Respir Dis.* 1979;120(5):1069-1072.
16. Badgett RG, Tanaka DJ, Hunt DK, et al. Can moderate chronic obstructive pulmonary disease be diagnosed by historical and physical findings alone? *Am J Med.* 1993;94(2):188-196.
17. Schapira RM, Schapira MM, Funahashi A, McAuliffe TL, Varkey B. The value of the forced expiratory time in the physical diagnosis of obstructive airways disease. *JAMA.* 1993;270(6):731-736.
18. Teklu B, Pierson DJ, Fair K, Schoene RB. The match test revisited: blowing out a candle as a screening test for airflow obstruction. *J Fam Pract.* 1990;31(5):557-562.
19. Foxman B, Lohr KN, Brook RH, Goldberg GA, Rosenthal M, Sloss E. *Conceptualization and Measurement of Physiologic Health for Adults: Chronic Obstructive Airway Disease.* Santa Monica, CA: RAND; 1982. Publication R-2262/8-1-HHS.
20. King DK, Thompson BT, Johnson DC. Wheezing on maximal forced exhalation in the diagnosis of atypical asthma: lack of sensitivity and specificity. *Ann Intern Med.* 1989;110(6):451-455.
21. Mannino DM, Etzel RA, Flanders WD. Do the medical history and physical examination predict low lung function? *Arch Intern Med.* 1993;153(16):1892-1897.
22. Rebuck AS, Tomarken JL. Pulsus paradoxus in asthmatic children. *CMAJ.* 1985;112(6):710-711.
23. Bilgi C, Jones RL, Sproule BJ. Relation between pulsus paradoxus and pulmonary function in patients with chronic airways obstruction. *CMAJ.* 1977;117(12):1389-1392.
24. Marks A, Bocles J. The match test and its significance. *South Med J.* 1960;53:1211-1216.
25. Snider TH, Stevens JP, Wilner FM, Lewis BM. Simple bedside test of respiratory function. *JAMA.* 1959;170:1631-1632.
26. Pardee NE, Martin CJ, Morgan EH. A test of the practical value of estimating breath sound intensity. *Chest.* 1976;70(3):341-344.
27. Pardee NE, Winterbauer RH, Morgan EH, Allen JD, Olson DE. Combinations of four physical signs as indicators of ventilatory abnormality in obstructive pulmonary syndromes. *Chest.* 1980;77(3):354-358.
28. Nairn JR, Turner-Warwick M. Breath sounds in emphysema. *Br J Dis Chest.* 1969;63(1):29-37.
29. Thacker RE, Kramer SS. The prevalence of auscultatory crackles in subjects without lung disease. *Chest.* 1982;81(6):672-674.
30. Lefcove NM, Wonnacott TH. The prevalence of chronic respiratory disease in the male physicians of London, Ontario. *CMAJ.* 1970;102(4):381-385.
31. McNicol KN, Williams HE, Gillam GL. Chest deformity, residual airways obstruction and hyperinflation, and growth in children with asthma. *Arch Dis Child.* 1970;45(244):783-798.
32. Fletcher CM. The clinical diagnosis of pulmonary emphysema: an experimental study. *Proc R Soc Med.* 1952;45(9):577-584.
33. Hepper NG, Hyatt RE, Fowler WS. Detection of chronic obstructive lung disease. *Arch Environ Health.* 1969;19(6):806-813.
34. Schilling RSF, Hughes JPW, Dingwall-Fordyce I. Disagreement between observers in an epidemiological study of respiratory disease. *BMJ.* 1955;1(4905):65-68.
35. Williams TJ, Ahmad D, Morgan WKC. A clinical and roentgenographic correlation of diaphragmatic movement. *Arch Intern Med.* 1981;141(7):878-880.
36. Bohadana AB, Mohankumar T. Symptoms and signs in the assessment of chronic airflow obstruction. *Indian J Chest Dis Allied Sci.* 1982;24(2-3):133-142.
37. Mulrow CD, Dolmatch BL, Delong ER, et al. Observer variability in the pulmonary examination. *J Gen Intern Med.* 1986;1(6):364-367.
38. Kilburn KH, Asmundsson T. Anteroposterior chest diameter in emphysema: from maxim to measurement. *Arch Intern Med.* 1969;123(4):379-382.
39. Forgacs P. The functional basis of pulmonary sounds. *Chest.* 1978;73(3):399-405.
40. Dodge R, Cline MG, Burrows B. Comparisons of asthma, emphysema and chronic bronchitis diagnoses in a general population sample. *Am Rev Respir Dis.* 1986;133(6):981-986.
41. Sherrill DL, Lebowitz MD, Knudson RJ, Burrows B. Smoking and symptom effects on the curves of lung function growth and decline. *Am Rev Respir Dis.* 1991;144(1):17-22.
42. Simel DL, Samsa GP, Matchar DB. Likelihood ratios with confidence: sample size estimation for diagnostic test studies. *J Clin Epidemiol.* 1991;44(8):763-770.
43. Wilkins RL, Dexter JR, Murphy RL, DelBono EA. Lung sound nomenclature survey. *Chest.* 1990;98(4):886-889.

This page intentionally left blank

Prepared by David L. Simel, MD, MHS

Reviewed by Sheri Keitz, MD

CLINICAL SCENARIO

A 51-year-old business executive comes to the clinic for a health checkup. He has no specific complaints, other than those he attributes to the vagaries of reaching middle age. You know the patient well and thus are aware that he smokes cigarettes. He has been smoking one-half to 1 pack a day since college. His neck and chest configuration are normal. There is no dyspnea, and he has never complained of either shortness of breath or cough other than during flulike illnesses. As part of his examination, you listen to his chest and hear no wheezes. You do not think he has airflow limitation, but is it time for spirometry testing, given his smoking history?

UPDATED SUMMARY ON OBSTRUCTIVE AIRWAYS DISEASE

Original Review

Holleman DR Jr, Simel DL. Does the clinical examination predict airflow limitation? *JAMA*. 1995;273(4):313-319.

UPDATED LITERATURE SEARCH

Our literature search used the parent search strategy for the Rational Clinical Examination series, combined with the subject headings “lung diseases,” “obstructive/di,” “pulmonary disease,” “chronic obstructive/di,” or “airway obstruction/di” published in English from 1994 to August, 2004. The results yielded 131 titles for which we reviewed the abstracts. As in the original Rational Clinical Examination article, we focused on studies that assessed clinical findings in a population of nonemergency primary care patients with irreversible airflow limitation, rather than the acutely dyspneic patient. The abstracts were reviewed to identify studies that might allow us to assess the sensitivity and specificity of patient symptoms or signs. We found 18 original articles for further review. We retained articles ($n = 5$) that included a population of patients without a previous diagnosis of obstructive airways disease who had their disease status verified by spirometry after a clinical evaluation.

We also crossed the clinical subject headings with “meta-analysis,” “ROC curve,” and the text word “systematic review” in both MEDLINE and the Cochrane databases. We retrieved articles referenced in Table 13-2 of The Rational Clinical Examination article for assessment of quality, along with examining files that we retained from the original Rational Clinical Examination article. These additional searches led us to 4 articles, 3 of which gave us insight into estimates for the prior probability of disease.

NEW FINDINGS

- The single best finding for identifying adults with obstructive airways disease is a history of > 40 pack-years of smoking
- Findings for obstructive airways disease in combination are much better than individual symptoms or signs

Details of the Update

On reviewing the initial Rational Clinical Examination article, we realized that we did not consider the potential effect of studies that included patients with a known diagnosis of obstructive airways disease. When obstructive airways disease is known and treated, some signs might improve (eg, wheezing, bedside measures of airflow such as forced expiratory time or peak flow), whereas other findings might not be affected by treatment (eg, maximum laryngeal height). It is also possible that including patients with a known diagnosis of obstructive airways disease minimizes the independent importance of the risk factors that led to the diagnosis (eg, smoking). For generalizability, the most promising studies should either analyze patients separately for those with a known diagnosis of obstructive airways disease or enroll a population that is independent of whether a previous diagnosis was made so that the examining clinicians would have no information about previous diagnosis and would be examining a variety of patients with and without obstructive airways disease. Therefore, we reassessed the studies that reported on combinations of findings from the original Rational Clinical Examination article.

One promising study did an excellent job of assessing interobserver variability.¹ The study reported good precision for the presence of wheezing ($\kappa = 0.69$) and for reduced breath sounds ($\kappa = 0.47$). However, we had included the univariate and multi-

Table 13-5 Univariate Findings for Obstructive Airways Disease in Patients With No Prior Obstructive Airways Disease Diagnosis

Finding (No. of Combined Studies)	LR+ (95% CI)	LR- (95% CI)
Wheezing (n = 5 ^{a,b}) ²⁻⁶	4.4 (1.6-12)	0.88 (0.84-0.92)
Maximum laryngeal height ≤ 4 cm ⁴	4.2 (2.3-7.9)	0.7 (0.5-0.9)
Decreased breath sounds (n = 2) ^{3,6}	2.6 (1.9-3.6)	0.66 (0.49-0.69)
Forced Expiratory Time, s ⁵		
≥9	6.7 (2.1-21)	
6-9	1.8 (0.77-4.0)	
<6	0.6 (0.5-0.8)	
Forced Expiratory Time Adjusted for Age, s ⁹		
≥6 And patient ≥60 y	3.4 (2.2-5.2)	
≥6 And patient <60 y	2.1 (1.3-3.5)	
<6 And patient ≥60 y	0.33 (0.23-0.47)	
<6 And patient <60 y	0.57 (0.34-0.95)	
Smoking Status, Pack-Years ^{a,4}		
>40	12 (2.7-50)	
20-40	0.8 (0.4-1.6)	
<20	0.5 (0.3-0.9)	
Overall Clinical Prediction of Disease (n = 2) ^{3,10c}		
Moderate-severe disease	5.6 (3.1-10)	
Mild disease	2.3 (0.55-9.7)	
No disease	0.59 (0.51-0.68)	

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

^aFor data from Straus et al,⁴ we used only the data for patients without obstructive airways disease.

^bAuscultated.

^cOverall clinical prediction of disease (or gestalt) is listed with univariate measures here because it is a single assessment by the clinician, without an explicit list of criteria, as opposed to the multivariate methods in Table 13-6 that use the specified variables.

variate data for our accuracy estimates in the original Rational Clinical Examination article, even though the study had only 15 patients with obstructive airways disease (of a total of 92 patients). The data were pooled from each physician, resulting in a reported sample size of approximately 340 observations. With only 15 affected patients, our confidence in the sensitivity from this study should have been much less, and the multivariate model may be overfit to the small number of affected patients.

IMPROVEMENTS IN THE DATA PRESENTED IN THE ORIGINAL PUBLICATION

Because of our uncertainty in combining the results from a study with only 15 affected patients,¹ we updated Table 13-3 from the original article. We added new information, updating the meta-analysis for auscultated wheezing²⁻⁶ and decreased breath sounds.^{2,4} In addition, we updated the results for forced expiratory time because our initial report combined data from a study with a univariate likelihood ratio (LR) for forced expiratory time with the results of forced expiratory time adjusted for age.

CHANGES IN THE REFERENCE STANDARD

The diagnosis of obstructive airways disease depends on a spirometric reference standard. Spirometry has 2 functions: it confirms airflow limitation and it confirms the lack of reversibility with bronchodilators. Patients with reversible airflow limitation have an asthmatic component to their airways disease. It is important that the spirometry be performed as soon as possible after the physical examination.

The spirometric criteria for obstructive airways require a forced expiratory volume in 1 second (FEV₁)/forced vital capacity (FVC) ratio of 0.70 or less, combined with a post-bronchodilator FEV₁ less than 80% of the patient's predicted value.^{7,8} Using different criteria results in a different prevalence of disease.⁷ Although specialty groups are coming to consensus and understanding of how these spirometric definitions differ, some authors of original articles have evaluated the effect of differing reference standards on the findings from the medical history and physical examination. Fortunately, there is no clinically meaningful difference on the LRs when different spirometric reference standards are used.⁴ As long as your patients are similar to those in the studies reviewed and your pulmonary laboratory uses one of the above definitions, the results of this literature review for the clinical examination would apply to your patients.

RESULTS OF LITERATURE REVIEW

The most diagnostically useful single finding was not an item from the physical examination but from the patient medical history—a finding that the patient smoked more than 40 pack-years (see Table 13-5). In the absence of appropriate analyses to assess the independence of findings, the best diagnostic strategy for diagnosing chronic airways obstruction is using the single most diagnostically useful positive LR or 1/negative LR.¹¹

Two studies allow us to assess the overall clinical impression according to whether the clinicians thought that the patient had moderate to severe, mild, or no disease (see Table 13-5). These results are the clinician's gestalt, assessing how he or she integrates all the available information. Unfortunately, clinicians do not accurately identify patients with mild disease (LR, 2.3; 95% confidence interval [CI], 0.55-9.7). The results for the clinical gestalt are not much better than would be obtained from using only the smoking history or using the information from any single physical examination finding. Given the diagnostic difficulty, it is appropriate to determine whether a more formal weighting of the data in a statistical (rather than intuitive) model can improve performance.

These data show the important influence of analyzing combinations of findings (see Table 13-6). Although the univariate data are mostly unimpressive, combinations of just a few findings greatly improve the diagnostic efficiency. Most clinicians will want the most parsimonious model. By parsimonious, we mean the model that has the smallest number of variables while yielding the best accuracy. The first and fourth models in Table 13-6 have the highest diagnostic odds ratios. What becomes readily apparent is that whereas the univariate

models have almost no ability to decrease the odds of obstructive airways disease from baseline, these 2 models appear quite efficient and may be able to rule out the disease.

EVIDENCE FROM GUIDELINES

Whereas all guidelines advocate for counseling patients to stop smoking, neither the US Preventive Health Services Task Force nor Canadian Task Force for Preventive Health Care evaluated the evidence for screening strategies for obstructive airways disease. The Global Initiative for Chronic Obstructive Lung Disease, sponsored by the National Heart, Lung, and Blood Institute, together with the World Health Organization, concluded that the benefits were unknown for a strategy of screening either the general population or the smaller population of smokers.¹²

CLINICAL SCENARIO—RESOLUTION

The clinical findings do not help this patient. Smoking as a single risk factor is not particularly helpful, although the quantity smoked in either pack-years or years of tobacco use is valuable information. This patient does not yet exceed the threshold that produces the highest likelihood ratios (>40 pack-years of smoking or having smoked for more than 55 years). The absence of wheezing and the finding of a normal neck do not appreciably lower the odds of obstructive airways disease. None of these findings moves us much from the baseline risk of 10% for an adult man.

What about combinations of findings? The results of the multivariate models give us the posterior odds of disease after applying the adjusted likelihood ratios. You can use the model with the variables self-reported obstructive airways disease, smoked more than 40 pack-years, aged 45 years or older, and maximum laryngeal height of 4 cm or less. His data yield posterior odds of $0.5 \times 0.8 \times 1.3 \times 0.16$, or 0.08. You have decreased the probability from the 10% baseline risk to about 7.7%. Because your clinical intuition is that the probability is higher than his baseline risk, you decide to check a different clinical model.

The other recommended model uses smoking status in terms of the number of years of use, symptoms of wheezing, and auscultated wheezing. His data yield posterior odds for the combined variables of $3.5 \times 0.26 \times 0.25$, or 0.23, which increases the probability from 10% to approximately 18%. With continued smoking, his likelihood of obstructive airways disease will increase more, whether or not he develops symptoms or signs.

The prediction models suggest that his risk of obstructive airways disease is about at the baseline risk to a little higher. On the other hand, you have a sense that he is going to develop the disease, and with a few more pack-years of smoking, the results of the 2 models increase precipitously and converge at approximately 40% to 45%, even in the absence of signs or symptoms. You might choose to get spirometry to prognosticate or to use the results as a motivational strategy to get him to stop smoking.

Table 13-6 Multivariate Findings for Obstructive Airways Disease

Model	Odds Used in Combination, Derived From a Model (Factor Present)	Odds Used in Combination, Derived From a Model (Factor Absent)
Combination of Patients With Known and Unknown OAD ⁴		
Smoked > 40 pack-years	8.3	0.8
Self reported history of OAD	7.3	0.5
Maximum laryngeal height ≤ 4 cm	2.8	0.16
Age ≥ 45 y	1.3	0.4
Posterior odds, all 4 findings present	220	
Posterior odds, all 4 findings absent		0.8
Combination of Patients With Known and Unknown OAD ⁵		
Forced expiratory time ≥ 9 s	4.6	0.8
Self-reported history of OAD	4.4	0.5
Wheezing	2.9	0.8
Posterior odds, all 3 findings present	59	
Posterior odds, all 3 findings absent		0.32
Patients Without Known OAD ⁴		
Smoked > 40 pack-years	12	0.9
Maximum laryngeal height ≤ 4 cm	3.6	0.7
Age ≥ 45 y	1.4	0.5
Posterior odds, all 3 findings present	58	
Posterior odds, all 3 findings absent		0.32
Patients Selected Without Consideration of OAD ¹⁰		
Smoked > 55 y	10	
Smoked 30–55 y	3.5	
Smoked < 30 y	0.23	
Auscultated wheezing	4.1	0.25
Self-reported wheezing	3.8	0.26
Posterior odds, all 3 findings present (smoked > 55 y)	156	
Posterior odds, all 3 findings absent (smoked < 30 y)		0.02

Abbreviation: OAD, obstructive airways disease.

OBSTRUCTIVE AIRWAYS DISEASE—MAKE THE DIAGNOSIS

PRIOR PROBABILITY

A systematic review identified 32 sources of information from studies done worldwide on the prevalence of obstructive airways disease.¹³ Nine of the 32 studies used a spirometric reference standard, similar to what is advocated for clinical practice; 8 of these had data that allowed us to compare the overall prevalence and sex-specific prevalence. The summary overall prevalence was 7.1% (95% CI, 5.2%-9.3%). Men (11%; 95% CI, 8.5%-14%) had about twice the rate as women (6%; 95% CI, 3%-10%) (see Table 13-7).

Table 13-7 Prior Probability of Obstructive Airways Disease Differs by Sex

	Prior Probability, %
Men	11
Women	6

POPULATION FOR WHOM OBSTRUCTIVE AIRWAYS DISEASE SHOULD BE CONSIDERED

All adults, especially those who smoke and are aged 45 years or older.

DETECTING THE LIKELIHOOD OF OBSTRUCTIVE AIRWAYS DISEASE

See Table 13-8.

Table 13-8 Likelihood Ratios for Best Single Findings and for Multivariate Models

Single Best Findings That Are the Easiest to Measure	Likelihood Ratio
Smoking status, > 40 pack-years	12
Auscultated wheezing or laryngeal height ≤ 4 cm	≈4
To “Rule In” Obstructive Disease, Must Use a Multivariate Model ^a	Posterior Odds of Disease, Probability (%)
Smoking > 55 y and wheezing symptoms and auscultated wheezing	156 (99)
History of OAD and smoking > 40 pack-years and age ≥ 45 y and laryngeal height ≤ 4 cm	220 (99)
To “Rule Out” Obstructive Disease, Must Use a Multivariate Model ^a	Posterior Odds of Disease, Probability (%)
Smoking < 30 y and no wheezing symptoms and no auscultated wheezing	0.02 (1.5)
No history of OAD and smoking < 40 pack-years and age < 45 y and laryngeal height > 4 cm	0.03 (3)

Abbreviation: OAD, obstructive airways disease.

^aSee multivariate table (Table 13-6) for other combinations of findings.

REFERENCE STANDARD TESTS

Spirometry using the pulmonary laboratories definition for the presence of obstructive airways disease.

REFERENCES FOR THE UPDATE

- Badgett RG, Tanaka DJ, Hunt DK, et al. Can moderate chronic obstructive pulmonary disease be diagnosed by historical and physical findings alone? *Am J Med*. 1993;94(2):188-196.
- Hepper NG, Hyatt RE, Fowler WS. Detection of chronic obstructive lung disease. *Arch Environ Health*. 1969;19(6):806-813.
- Holleman DR, Simel DL, Goldberg JS. Diagnosis of obstructive airways disease from the clinical examination. *J Gen Intern Med*. 1993;8(2):63-68.
- Straus SE, McAlister FA, Sackett DL, Deeks JJ, for the CARE-COAD1 Group. The accuracy of patient history, wheezing, and laryngeal measurements in diagnosing obstructive airways disease. *JAMA*. 2000;283(14):1853-1857.^a
- Straus SE, McAlister FA, Sackett DL, Deeks JJ, for the CARE-COAD2 Group. Accuracy of history, wheezing, and forced expiratory time in the diagnosis of chronic obstructive pulmonary disease. *J Gen Intern Med*. 2002;17(9):684-688.^a
- Garcia-Pachon E. Paradoxical movement of the lateral rib margin (Hoover sign) for detecting obstructive airway disease. *Chest*. 2002;122(2):651-655.^a
- Viegi G, Pedreschi M, Pistelli F, et al. Prevalence of airways obstruction in a general population. *Chest*. 2000;117(5 suppl 2):S339-S345.
- Iqbal A, Schloss S, George D, Isonaka S. Worldwide guidelines for chronic obstructive pulmonary disease: a comparison of diagnosis and treatment recommendations. *Respirology*. 2002;7(3):233-239.
- Schapira RM, Schapira MM, Funahashi A, McAuliffe TL, Varkey B. The value of the forced expiratory time in the physical diagnosis of obstructive airways disease. *JAMA*. 1993;270(6):731-736.
- Melby H, Araaas I, Hana J, Hensrud A. Improving pulmonary auscultation as a tool in the diagnosis of bronchial obstruction results of an educational intervention. *Scan J Prim Health Care*. 1998;16(3):160-164^a
- Holleman DR, Simel DL. Quantitative assessments from the clinical examination: how should clinicians integrate the numerous results? *J Gen Intern Med*. 1997;12(3):165-171.^a
- Pauwels RA, Buist AS, Calverley PM, Jenkins CR, Hurd SS; GOLD Scientific Committee. Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease: NHLBI/WHO Global Initiative for Chronic Obstructive Lung Disease (GOLD) Workshop summary. *Am J Respir Crit Care Med*. 2001;163(5):1256-1276.
- Halbert RJ, Isonaka S, George D, Iqbal A. Interpreting COPD prevalence estimates: what is the true burden of disease? *Chest*. 2003;123(5):1684-1692.

^aFor the Evidence to Support the Update for this topic, see <http://www.JAMAEvidence.com>.

EVIDENCE TO SUPPORT THE UPDATE:

13

Chronic Obstructive Airways Disease

TITLE Office Spirometry Significantly Improves Early Detection of Chronic Obstructive Pulmonary Disease in General Practice: The DIDASCO Study.

AUTHORS Buffels J, Degryse J, Hayrman J, Decramer M.

CITATION *Chest*. 2004;125(4):1394-1399.

QUESTION Does a simple patient questionnaire identify patients with obstructive airways disease?

DESIGN Consecutive patients, prospective.

SETTING Twenty general practitioners in Belgium.

PATIENTS The patients were aged 35 to 70 years and visiting their general practitioner during a 12-week period. Subjects using bronchodilators or inhaled steroids were excluded.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

See [Table 13-9](#). All patients with at least 1 positive answer on the questionnaire were considered to have a positive result and underwent spirometry within 1 week. A random sample of individuals with no complaints (10%) underwent spirometry. The data were appropriately adjusted for this planned verification bias. It is not clear whether the spirometry and its interpretation was blinded to the questionnaire results.

Table 13-9 Simple Questions for Obstructive Airways Disease

Do you have any of the following complaints?

1. Cough lasting for at least 2 weeks
2. Breathing difficulties during mild exercise or at night
3. Wheezing
4. Any kind of nasal allergy or hay fever
5. Have you had 1 or more of these complaints during the past year
6. Have you ever had to visit your physician for a wheezing or long-lasting cough

MAIN OUTCOME MEASURES

Obstructive airways disease was identified by spirometry. Patients with newly found airways obstruction then underwent assessment for reversibility with a bronchodilator. Those with reversible findings were considered to have asthma rather than obstructive airways disease. All general practitioners used the same model of spirometer, underwent training for its use, and had their precision assessed. The difference between the pulmonary function laboratory- and office-based tests was only 2.2%.

MAIN RESULTS

See [Table 13-10](#). One hundred thirty-five patients had newly diagnosed obstructive airways disease. After adjusting for verification bias, the number of new diagnoses by spirometry was extrapolated to 216 of 2923. Of the patients with disease, less than 10% had moderate to severe or worse obstructive airways disease.

CONCLUSIONS

LEVEL OF EVIDENCE Level 3.

STRENGTHS This is one of the few studies for any clinical examination finding that recognized verification bias, planned for it appropriately, and adjusted for it appropriately in the analysis.

LIMITATIONS The general practitioners knew these patients and may likely have known their airways disease and smoking status before conducting the spirometry tests. Nonetheless, the results of the office-based data showed excellent precision

Table 13-10 Likelihood Ratios for at Least 1 Positive Answer on the Questionnaire

Test	Sensitivity	Specificity	LR+ (95% CI)	LR- (95% CI)
Positive questionnaire (at least 1 positive answer)	0.58	0.79	2.7 (2.4-3.1)	0.53 (0.45-0.61)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

compared with those of a pulmonary laboratory. Furthermore, the practitioners had to show competence in use of the instrument.

The investigators intentionally did not include a question about pack-years of smoking, because they believed that the clinical evidence indicated that all smokers older than 45 years should have spirometry testing. In the discussion, the authors observed that the diagnoses were all new, even though these were patients followed in their practice.

A simple symptom-based clinical model, without eliciting patient risk factors, was not particularly useful for patients with predominantly unrecognized, mild obstructive airways disease. That finding is not surprising but is useful. The accuracy of the multivariate model that used the answers to all the questions is not provided.

Reviewed by David L. Simel, MD, MHS

Table 13-11 Likelihood Ratios That Are Independently Useful for Diagnosis of Obstructive Airways Disease

Model Variable	LR When Finding Is Present	LR When Finding Is Absent
Smoke > 55 y	10	
Smoke 30-55 y	3.5	
Smoke < 30 pack-years	0.23	
Auscultated wheezing	4.1	0.25
Self-reported wheezing	3.8	0.26

Abbreviation: LR, likelihood ratio.

approaches were statistically similar. The third and fourth approaches that use a multivariate model to reduce the number of variables were the optimal approaches because they reduced the amount of data that the physician had to collect while still having a higher accuracy from the receiver operating characteristic curve.

CONCLUSIONS

In this reanalysis of the data originally included in The Rational Clinical Examination article, we get a better understanding of the concept of statistical independence as it applies to likelihood ratios. The patient report of previous wheezing was as useful as, and independent of, the information obtained by auscultated wheezing.

Reviewed by David L. Simel, MD, MHS

REFERENCE FOR THE EVIDENCE

- Holleman DR, Simel DL, Goldberg JS. Diagnosis of obstructive airways disease from the clinical examination. *J Gen Intern Med*. 1993;8(2):63-68.

MAIN OUTCOME MEASURES

The main outcome measure was the area under the curve (a measure of accuracy) for 4 approaches to using a list of symptoms and signs for obstructive airways disease. The approaches were (a) use all the likelihood ratios for the individual findings and multiply them serially without any regard to independence; (b) pick the single best likelihood ratio applicable to the patient; (c) identify the most important variables from a logistic model and then use the unadjusted, raw likelihood ratios from method (a); or (d) use the adjusted likelihood ratios for the variables found significant in a multivariate model.¹

MAIN RESULTS

See Table 13-11. Using all the information but ignoring independence was the worst approach and required the most amount of clinician time to collect the data. The other

TITLE Improving Pulmonary Auscultation as a Tool in the Diagnosis of Bronchial Obstruction—Results of an Educational Intervention.

AUTHORS Melbye H, Aaraas I, Hana J, Hensrud A.

CITATION *Scand J Prim Health Care*. 1998;16(3):160-164.

QUESTION Does an educational intervention consisting of audiovisual review of lung sounds and a didactic review emphasizing diminished breath sounds, crackles, and wheezes over the usefulness of rhonchi improve the prediction of the forced expiratory volume in 1 second (FEV₁) percentage?

DESIGN Before-after study of an educational intervention.

SETTING Five primary care practices in Norway.

PATIENTS Convenience sample of general practice patients with a 1:3 ratio of patients with vs without pulmonary symptoms. There were 354 patients enrolled in the phase before the intervention and 343 patients in the second phase after the intervention.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

The physicians recorded physical findings on a prespecified list and then estimated the FEV₁ (<60%, 60%-79%, or ≥80% predicted). The diagnostic standard spirometry test was performed after the clinical examination, blinded to the clinical results.

MAIN OUTCOME MEASURE

Accuracy of FEV₁ percentage prediction.

MAIN RESULTS

See Table 13-12. Before the educational intervention, the accuracy of predicting the correct FEV₁ percentage range was 0.68 (SE, 0.04). After the educational intervention, the accuracy was 0.71 (SE, 0.04). These accuracy outcomes were derived from the area under the receiver operating characteristic curve, using data in the article. There was no statistical difference in the accuracy before vs after the intervention ($P = .21$). Therefore, we combined the data before the educational intervention with the data after the intervention, displaying the ability to predict the presence of disease using a cut point of FEV₁ less than 80% predicted.

CONCLUSIONS

LEVEL OF EVIDENCE Level 1 randomized trial of an educational intervention.

STRENGTHS Independent assessment of the outcome.

Table 13-12 Serial Likelihood Ratios for the FEV₁ to the Diagnosis of Obstructive Airways Disease

Clinical Prediction of FEV ₁ Percentage, %	LR for OAD
<60 predicted	11 (3.9-32)
60-79 predicted	6.5 (4.2-9.9)
≥80	0.61 (0.53-0.7)

Abbreviations: FEV₁, forced expiratory volume at 1 second; LR, likelihood ratio; OAD, obstructive airways disease.

LIMITATIONS Practitioners likely knew their affected patients. The cut point for the FEV₁ in this study was that of the British Thoracic Society, which is slightly higher than other recommendations.

These practitioners were good at identifying the patients with more significant disease. However, they were not as good at ruling out disease. The educational intervention that emphasized the physical findings noted above had little effect on these general practitioners. The conclusions are that the clinical findings were not useful, the clinical findings are useful but the educational intervention was not effective, or these providers already knew the patients who had obstructive airways disease and those who did not. Overall, these clinicians were already good diagnosticians in being able to identify those with disease (see likelihood ratio [LR] for those they predicted would have obstructive airways disease). However, they were not as good at identifying those patients with normal results because the LR when they predicted normality was only 0.61. These data are consistent with the data for the physical examination components that show that individual findings do not rule out obstructive airways disease. They are also consistent with the overall assessment reported in another study for moderate to severe disease (LR, 4.2), mild disease (LR, 0.82), and no disease (LR, 0.42), in which the clinicians did not know the patients.¹

Reviewed by David L. Simel, MD, MHS

REFERENCE FOR THE EVIDENCE

- Holleman DR, Simel DL, Goldberg JS. Diagnosis of obstructive airways disease from the clinical examination. *J Gen Intern Med*. 1993;8(2):63-68.

TITLE Paradoxical Movement of the Lateral Rib Margin (Hoover Sign) for Detecting Obstructive Airway Disease.

AUTHOR Garcia-Pachon E.

CITATION *Chest*. 2002;122(2):651-655.

QUESTION Do the Hoover sign and other clinical measures of obstructive airways disease have good reliability and accuracy?

DESIGN Prospective, independent comparison with spirometry.

SETTING Pulmonary consultation clinic.

PATIENTS Consecutive patients ($n = 172$) referred to a pulmonary clinic who were older than 40 years and smokers more than 20 pack-years, had a previous diagnosis or self-report of obstructive pulmonary disease, were receiving inhalants for more than 6 months, or had dyspnea. Thus, all patients were those in whom a diagnosis of obstructive airways disease could be considered.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

The examiners were a pulmonologist and a first-year resident. The examinations were done independently. Each resident was specifically trained for 1 week before the patients were enrolled. A blinded technician performed spirometry tests independently. Standard definitions of obstructive airways disease were used.

In patients without disease, the lower thoracic rib cage moves upward and outward with inspiration. Hoover sign refers to a paradoxical indrawing of the lateral ribs with inspiration, attributed to a fixed and flattened diaphragm.

MAIN OUTCOME MEASURES

Sensitivity, specificity, likelihood ratio (LR), and κ statistics.

MAIN RESULTS

See Table 13-13. Of the 172 patients, 64 (37%) met spirometric criteria for obstructive airways disease.

The results for the positive likelihood ratio (LR+) and negative LR were statistically similar for all findings. We report these as meta-analytically combined results, according to the results given.

CONCLUSIONS

LEVEL OF EVIDENCE Level 1.

STRENGTHS Comparison between a trainee and staff pulmonologist.

Table 13-13 Agreement (κ) and Likelihood Ratios for Findings of Obstructive Airways Disease

Test	κ (95% CI)	LR+ (95% CI)	LR- (95% CI)
Hoover sign	0.74 (0.63-0.86)	4.6 (3.1-6.9)	0.50 (0.4-0.61)
Rhonchi	0.38 (0.13-0.64)	3.0 (1.6-5.7)	0.87 (0.78-0.98)
Reduced breath sounds	0.51 (0.37-0.65)	2.7 (2.0-3.7)	0.54 (0.43-0.68)
Wheeze	0.67 (0.51-0.84)	1.3 (0.78-2.3)	0.95 (0.83-1.1)
Overall clinical impression	0.61 (0.49-0.73)	3.6 (2.7-4.7)	0.25 (0.17-0.37)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

LIMITATIONS The study was done in a pulmonary clinic from new consultations, although the frequency of obstructive airways disease is comparable to that of previous studies. The specific training (especially for the Hoover sign) of the resident likely improved the reliability, which strengthens the study even though the results may not generalize to individuals without similar training.

The results for the reliability of wheezing and reduced breath sounds are almost identical to that found in a larger study of reliability.¹ It is reassuring that a first-year resident with only 1 week of specific training on the pulmonary examination can develop an overall clinical impression that agrees with that of their pulmonology instructor.

The setting for this study (a pulmonology clinic) was different from that of other studies. However, the patients were all referral patients and unknown to the examiners before the study. The prevalence of disease in this population was similar to that of other studies of the physical examination. The prevalence of moderate to severe disease (22% of all patients and 59% of those with obstructive airways disease) was similar to that of a previous study done in a pulmonary laboratory² but much lower than that of a study done that recruited patients with similar qualifying characteristics independently of a referral.¹

At least among a group of patients referred to a pulmonary clinic, the overall clinical impression was more efficient (highest diagnostic odds ratio at 14) than any individual finding. Holleman et al³ found a similar LR+ for detecting moderate to severe disease (LR, 4.2; 95% confidence interval [CI], 2.2-8), but their ability to rule out disease in a less severely affected population not referred to a pulmonary clinic was not as efficient (LR for the clinical impression of normality was 0.42; 95% CI, 0.25-0.70). The ability of the clinicians in the multinational study to come up with a useful overall clinical impression was poor.

Hoover sign needs to be confirmed with different examiners and in different populations to make sure that it is reproducible and does not vary with disease prevalence. The results reported here for wheezing can be combined with other studies, although this study reports statistically worse LRs.

Reviewed by David L. Simel, MD, MHS

REFERENCES FOR THE EVIDENCE

- Badgett RG, Tanaka DJ, Hunt DK, et al. Can moderate chronic obstructive pulmonary disease be diagnosed by historical and physical findings alone? *Am J Med.* 1993;94(2):188-196.
- Schapira RM, Schapira MM, Funahashi A, McAuliffe TL, Varkey B. The value of the forced expiratory time in the physical diagnosis of obstructive airways disease. *JAMA.* 1993;270(6):731-736.
- Holleman DR, Simel DL, Goldberg JS. Diagnosis of obstructive airways disease from the clinical examination. *J Gen Intern Med.* 1993;8(2):63-68.

TITLE The Accuracy of Patient History, Wheezing, and Laryngeal Measurements in Diagnosing Obstructive Airway Disease.

AUTHORS Straus SE, McAlister FA, Sackett DL, Deeks JJ, for the CARE-COAD1 Group.

CITATION *JAMA.* 2000;283(14):1853-1857.

QUESTION How do we determine the accuracy of a variety of signs for obstructive airways disease?

DESIGN Prospective.

SETTING Multinational, primary care.

PATIENTS Consecutive patients from each site, sorted into one of 3 groups: known chronic obstructive airways disease ($n = 76$ [25%]), suspected chronic airways disease ($n = 114$ [37%]), neither known nor suspected chronic airways disease ($n = 119$ [39%]).

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

The patients were assessed clinically with the physical examination findings of laryngeal height, laryngeal descent, and wheezing. Videotaped instructions for assessing laryngeal height and descent were shared with each investigator (<http://www.carestudy.com/CareStudy/COAD3/Intro.asp>, accessed April 20, 2008). The clinical examination findings were recorded independently and blinded to the spirometry results. Likewise, the spirometrists were blinded to the clinical assessment.

MAIN OUTCOME MEASURES

Univariate and multivariate likelihood ratios. Obstructive airways disease with analyses analyzed to determine the effect of different case definitions of obstructive airways disease. A multivariate analysis assessed combinations of variables.

MAIN RESULTS

See **Tables 13-14** and **13-15**.

Table 13-14 Likelihood Ratios of Univariate Findings for Patients Without Known Obstructive Airways Disease

	LR+ (95% CI)	LR- (95% CI)
Smoking Status, Pack-Years		
>40	12 (2.7-50)	
20-40	0.8 (0.4-1.6)	
<20	0.5 (0.3-0.9)	
Age, y		
≥65	1.9 (1.3-2.8)	
45-64	1.5 (1.1-2.2)	
<45	0.3 (0.2-0.5)	
Finding		
Wheezing	2.1 (1.2-3.5)	0.9 (0.7-1.0)
Maximum laryngeal height ≤ 4 cm	4.2 (2.3-7.9)	0.7 (0.5-0.9)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

Table 13-15 Likelihood Ratios for Multivariate Findings for All Patients vs Those Without Known Obstructive Airways Disease^a

	LR+	LR-
All Patients		
Self-reported OAD	7.3	0.5
Smoked > 40 pack-years	8.3	0.8
Age ≥ 45 y		
Maximum laryngeal height ≤ 4 cm	2.8	0.8
Patients Without Known OAD		
Smoked > 40 pack-years	12	0.9
Age ≥ 45 y	1.4	0.5
Maximum laryngeal height ≤ 4 cm	3.6	0.7
All 3 factors present vs none present	58.5	0.32

Abbreviations: LR+, positive likelihood ratio; LR-, negative likelihood ratio; OAD, obstructive pulmonary disease.

^aFor the multivariate models, the LRs appropriate to an individual patient's results can be multiplied to determine the LR specific to that patient.

CONCLUSIONS

LEVEL OF EVIDENCE Level 1.

STRENGTHS Assessment of different case definitions for obstructive airways disease. The results for patients without a prior OAD diagnosis can be compared to the entire population.

LIMITATIONS None.

As in other studies, this high-quality study showed that smoking status dominates the clinical symptoms and signs.

According to a receiver operating characteristic analysis of all patients, they chose a cut point of greater than 40 pack-years. The results also show that no single finding can be used to rule out obstructive airways. Because of the lack of ability of single findings to prove normality, the investigators appropriately examined combinations of findings. Before assessing that, they determined that using a variety of accepted spirometric definitions for obstructive airways disease did not alter the univariate likelihood ratios in a clinically meaningful way.

The assessment of the threshold value for smoking status was based on all patients. This brings up an important point not just about smoking status but also about the physical findings. In practice, including patients with known obstructive airways disease in a study of the clinical examination may not give the results that clinicians need; once you know the patient has obstructive airways disease, the physical findings no longer matter for diagnosis. Fortunately, the investigators include a separate analysis for patients without known obstructive airways disease.

Including patients with known obstructive airways disease affects the results for sensitivity in various ways. In general, including more severely ill patients (or those with disease that is more obvious) would be expected to inflate the sensitivity and make the negative likelihood ratio appear optimistically low. However, this may not always be the case. For example, wheezing might be one finding that physicians “treat” when they know their patients have obstructive airways disease. Thus, patients with known obstructive airways disease who are under treatment might proportionately wheeze less than untreated, affected patients. The effect on sensitivity from including vs excluding such patients will create variability in outcomes as a function of the relative proportion of such patients and the pattern of their disease severity. Similarly, a finding such as abnormal laryngeal height might be fixed and not appear until more severe disease is present and not change with treatment. In this study, the point estimate for maximum laryngeal height (≤ 4 vs > 4 cm) appeared better than wheezing, although the confidence intervals overlapped. It is likely that future studies will show that either wheezing or laryngeal height is useful, but not both. Because of the overlap in their LRs, some multivariate models could have wheezing, whereas others might have laryngeal height, depending on the spectrum of disease. The multivariate models comparing a population of patients in which 25% have known obstructive airways disease, vs those in whom the status is unknown, show that the ability to rule in or rule out disease is not as efficient because information is lost. The model for patients with disease status unknown seems to be the most relevant for clinicians who are trying to establish a diagnosis.

Reviewed by David L. Simel, MD, MHS

TITLE Accuracy of History, Wheezing, and Forced Expiratory Time in the Diagnosis of Chronic Obstructive Pulmonary Disease.

AUTHORS Straus SE, McAlister FA, Sackett DL, Deeks JJ, for the CARE-COAD2 Group.

CITATION *J Gen Intern Med.* 2002;17(9):684.

QUESTION What is the accuracy of a variety of signs for obstructive airways disease?

DESIGN Prospective.

SETTING Multinational, primary care.

PATIENTS Consecutive patients from each site, sorted into one of 3 groups: known chronic obstructive airways disease ($n = 66$ [41%]), suspected chronic airways disease ($n = 43$ [27%]), and neither known nor suspected chronic airways disease ($n = 52$ [32%]).

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

The patients were assessed clinically with the physical examination findings of wheezing and forced expiratory time. The clinical examination findings were recorded independently and blinded to the spirometry results. Likewise, the spirometrists were blinded to the clinical assessment.

MAIN OUTCOME MEASURES

Univariate and multivariate likelihood ratios (LRs). Obstructive airways disease with analyses analyzed to determine the influence of different case definitions of obstructive airways disease. A multivariate analysis assessed combinations of variables.

MAIN RESULTS

See [Tables 13-16](#) and [13-17](#).

Table 13-16 Likelihood Ratios for Univariate Findings for All Patients

Test	LR+ (95% CI)	LR- (95% CI)
Wheezing	4.0 (1.6-9.9)	0.8 (0.7-0.9)
Smoking Status, Pack-Years		
>40	3.3 (1.5-7.1)	
20-40	2.0 (1.0-4.0)	
<20	0.7 (0.4-1.2)	
Age, y		
>65	1.6 (1.1-2.3)	
≤65	0.7 (0.5-0.9)	
Forced Expiratory Time, s		
>9	6.7 (2.1-21)	
6-9	1.8 (0.77-4.0)	
<6	0.6 (0.5-0.8)	

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

Table 13-17 Multivariate Findings for All Patients^a

Test	LR+	LR-
Forced expiratory time ≥ 9 s	4.6	0.8
Self-reported OAD	4.4	0.5
Wheezing	2.9	0.8

Abbreviations: LR+, positive likelihood ratio; LR-, negative likelihood ratio; OAD, obstructive airways disease.

^aFor the multivariate model, the LRs appropriate to a patient's results can be multiplied to determine the LR specific to that patient. Smoking status was not independently significant once these 3 variables were considered.

CONCLUSIONS

LEVEL OF EVIDENCE

Level 1.

STRENGTHS Assessment of different case definitions for obstructive airways disease.

LIMITATIONS Forty-one percent of the patients had known obstructive airways disease.

The results in this study were similar to those in the first study reported by the same group of authors. The authors argue that it makes sense to include the patient's self-report of a previous diagnosis in a logistic model. It does make sense because the patient's report may or may not be cor-

rect. However, the results may not generalize as well to patients who are unaware of their status simply because effective treatment may affect the physical examination findings (eg, wheezing).

The adjusted LR for a previous diagnosis of obstructive airways disease was 4.4, with a negative likelihood ratio (LR-) of 0.5. Although Holleman et al¹ did not assess patients for a previous diagnosis of obstructive airways disease, they did collect symptoms of chronic bronchitis. The results are consistent in that the independent LR for chronic bronchitic symptoms was 3.8, with an LR- of 0.66 in the Holleman et al study.¹

In this study, there was a high prevalence of patients with known obstructive airways disease. The higher prevalence of disease appears to have rendered the additional information about smoking status useless when evaluated in combination with other findings. Nonetheless, the results here support those reported by Holleman et al¹ that forced expiratory time adds information to wheezing status.

Reviewed by David L. Simel, MD, MHS

REFERENCE FOR THE EVIDENCE

- Holleman DR, Simel DL, Goldberg JS. Diagnosis of obstructive airways disease from the clinical examination. *J Gen Intern Med*. 1993;8(2):63-68.

This page intentionally left blank

CHAPTER 14

Does This Patient Have Clubbing?

Kathryn A. Myers, MD, EdM, FRCPC

Donald R. E. Farquhar, MD, SM, FRCPC

CLINICAL SCENARIOS

CASE 1 A respiratory therapist asks you to see her asymptomatic 76-year-old mother in consultation because she is concerned that her mother has clubbing. The patient has increased curvature of the nails, and you wonder whether other physical examination techniques can help you decide whether clubbing is present.

CASE 2 While performing a routine physical examination on a 65-year-old female smoker with chronic obstructive pulmonary disease (COPD), you detect changes in the fingers suggestive of clubbing. You recall an association between clubbing and certain types of pulmonary disease, and you wonder whether any further diagnostic evaluation of this patient is warranted.

WHY IS THE CLINICAL EXAMINATION IMPORTANT?

Clubbing is one of those phenomena with which we are all so familiar that we appear to know more about it than we really do.¹

—Samuel West, 1897

The association of clubbing with a host of infectious, neoplastic, inflammatory, and vascular diseases has captured the imagination of clinicians since Hippocrates first described clubbing in a patient with empyema in the fifth century BC.² Although clubbing can be a benign hereditary condition, the diagnostic implications in an adult are such that its detection should prompt consideration of the underlying etiology (Table 14-1).^{3,4} In the pediatric population, clubbing usually represents the progression of established diseases, such as cystic fibrosis or uncorrected cyanotic congenital heart disease.

Digital clubbing is characterized by the enlargement of the terminal segments of the fingers or toes that results from the proliferation of the connective tissue between the nail matrix and the distal phalanx. Although most often symmetrical, clubbing can be unilateral or even unidigital.^{5,6} Clubbing can occur in isolation or in association with hypertrophic osteoarthropathy.^{7,8} Hypertrophic osteoarthropathy, a systemic disorder affecting bone and joints, is most commonly associated with bronchogenic carcinoma, but it can occur in association with extrapulmonary malignancies, as well as nonmalignant pulmonary diseases.⁹ Pachydermoperiostosis is a rare, congenital form of hypertrophic osteoarthropathy. Congenital clubbing, which usually has its onset in childhood, may represent a limited form of pachydermoperiostosis.⁵

Unlike such physical findings as ascites and splenomegaly, the clinical impression of clubbing cannot be verified by simple imaging tests. Throughout the past century, many investigators have described possible reference standards for

Table 14-1 Conditions Associated With Acquired Clubbing

Neoplastic intrathoracic disease
Bronchogenic carcinoma
Malignant mesothelioma
Pleural fibroma
Metastatic osteogenic sarcoma
Suppurative intrathoracic disease
Lung abscess
Bronchiectasis
Cystic fibrosis
Empyema
Chronic cavitary mycobacterial or fungal infection
Diffuse pulmonary disease
Idiopathic pulmonary fibrosis
Asbestosis
Pulmonary arteriovenous malformations
Cardiovascular disease
Cyanotic congenital heart disease
Infective endocarditis
Arterial graft sepsis ^a
Brachial arteriovenous fistula ^b
Hemiplegic stroke ^b
Gastrointestinal disease
Inflammatory bowel disease
Celiac disease
Hepatobiliary disease
Cirrhosis (particularly biliary and juvenile)
Metabolic disease
Thyroid acropachy

^aAssociated with clubbing distal to graft sepsis.

^bAssociated with unilateral clubbing.

diagnosis of clubbing, including water displacement of the terminal phalanges, measurement of nail curvature using a device called an unguisometer, and measuring nail angles and ratios with plaster casts or shadow projections of fingers.¹⁰⁻¹⁵ None has been accepted as a criterion standard of diagnosis, and all are cumbersome and impractical as a method of verifying the clinical impression of clubbing. Therefore, physicians must rely solely on their skills in clinical examination to detect clubbing.

Pathophysiology

Normally, the nailbed thickness is less than 2.0 mm. Clubbed fingers studied at autopsy show not only a thickness greater than 2.0 mm but also a lower density of nailbed connective tissue.¹⁶ Morphologic findings include the presence of primitive fibroblasts, increased numbers of eosinophils and lymphocytes, and increased caliber and number of blood vessels. Genetic predisposition, vagally mediated neural mechanisms, and the direct effect of tissue hypoxia or of circulating vasodilators that elude metabolism in the lung through right-to-left shunting have all been proposed to explain the

morphology. Although there is experimental and clinical evidence to support each of these hypotheses, it has not been possible to formulate a comprehensive theory of pathogenesis applicable to all clinical circumstances.^{5,17-19}

Symptoms

Clubbing is almost always painless, unless it is associated with hypertrophic osteoarthropathy. Symptoms of hypertrophic osteoarthropathy include periarticular pain and swelling, most often in the wrists, ankles, knees, and elbows. Accordingly, the presentation of hypertrophic osteoarthropathy can be confused with such primary rheumatologic disorders as rheumatoid arthritis.⁵ Many patients with clubbing express unawareness of any abnormality in their fingers. In one series of patients with clubbing, only 32 of 116 patients were aware of the onset of the changes in their nails, and only 2 reported painful fingers or joints.²⁰

Signs

Identification of advanced clubbing, which is characterized by so-called drumstick fingers, poses little difficulty for clinicians. By contrast, the subtleties of the earlier stages of clubbing may lead to animated bedside debate among medical students, residents, and experienced physicians. The 2 approaches for identifying clubbing on physical examination are visual inspection and palpation of the cuticle for increased sponginess.^{16,21,22}

Inspection

General Appearance

Inspection of the fingers for clubbing can reveal abnormalities in the nailfold angles and in the shape, depth, and width of the terminal phalanges. In addition to the obvious changes in the shape of the terminal phalanges in established clubbing (Figure 14-1A), close inspection of the cuticle may reveal a shiny and smooth appearance. Lovibond²³ described a lilac hue of the nail fold in clubbing, caused by increased vascularity in the connective tissue. Although the increased nail curvature seen in clubbed fingers has been studied extensively using chord-arc measurements and unguisometers, it is not easily measured at the bedside. Moreover, nail curvature tends to become more pronounced with age and can occur in the absence of other signs of clubbing.^{5,24}

Nailfold Angles

Inspection of clubbed fingers reveals a number of abnormalities in the angles made by the nail as it exits from the terminal phalanx. Lovibond²³ popularized this as the profile sign in his 1938 report on the diagnosis of clubbed fingers. He observed that in normal fingers, the nail projects from the nail bed at an angle of about 160 degrees, but this angle approached 180 degrees in clubbed fingers (Figure 14-1B). Later, the hyponychial angle was proposed as a more reliable sign than the profile angle in the assessment of clubbing (Figure 14-1B).¹¹

Phalangeal Depth Ratio

Estimation of the phalangeal depth ratio (PDR) can be used to identify clubbing (Figure 14-1C).¹⁴ In the normal finger,

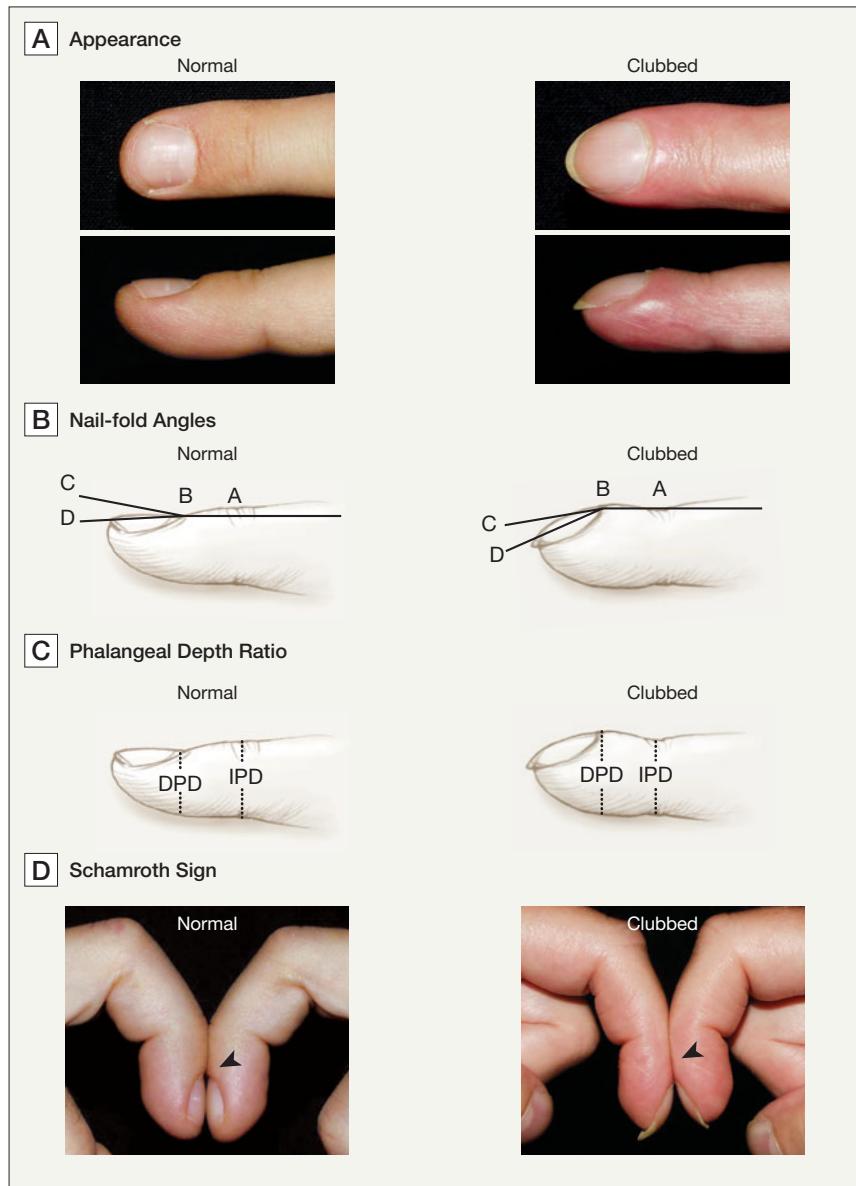


Figure 14-1 Appearance on Inspection for Clubbing

A, Normal finger viewed from above and in profile, and the changes occurring in established clubbing viewed from above and in profile. B, The finger on the left demonstrates normal profile (ABC) and normal hyponychial (ABD) nailfold angles of 169 degrees and 183 degrees, respectively. The clubbed finger on the right shows increased profile and hyponychial nailfold angles of 191 degrees and 203 degrees, respectively. C, Distal phalangeal finger depth (DPD)/interphalangeal finger depth (IPD) represents the phalangeal depth ratio. In normal fingers, the IPD is greater than the DPD. In clubbing, this relationship is reversed. D, Schamroth sign. In the absence of clubbing, opposition of the index fingers nail to nail creates a diamond-shaped window (arrowhead). In clubbed fingers, the loss of the profile angle because of the increase in tissue at the nail bed causes obliteration of this space (arrowhead).

the distal phalangeal depth is smaller than the interphalangeal depth. As connective tissue deposition expands the pulp in the terminal phalanx, this ratio becomes reversed. The PDR appears to be independent of age, sex, and ethnicity in randomly selected populations.^{14,25} A similar ratio using distal and interphalangeal width can be determined, but it has not been studied as extensively as the PDR.

Although the PDR was originally described using plaster casts and shadowgrams, subsequent studies have reported the use of calipers on live fingers. To perform this measurement, the calipers should touch but not compress the tissue at the distal phalanx and the interphalangeal joint of the index finger during measurement. Baughman et al²⁶ estimated that this technique takes no longer than 1 minute to perform. Visual estimation for the reversal of the PDR has been suggested as a simple bedside technique for clubbing, but the precision of this method has not been tested.

Schamroth Sign

In 1976, Schamroth²⁷ reported a new clinical sign that incorporated 2 of the clinical features of clubbing (Figure 14-1D). Normal fingers create a diamond-shaped window when the dorsal surfaces of terminal phalanges of similar fingers are opposed. In the clubbed finger, the diamond becomes obliterated because of the loss of the profile angle and the increase in the soft tissue at the cuticle. Since its original description, this technique has become popular with physicians as a quick test to establish the presence of clubbing. The precision and accuracy of this sign, however, have not been formally tested.²⁸

Palpation

On palpation of the base of the nail bed, the examiner perceives that the nail is floating within the soft tissue and, in advanced cases, may even be able to feel the proximal edge of the nail. This sign is best elicited by gently rocking the nail.

The examiner grips the sides of the subject's finger between the thumb and middle finger of each hand. Exerting downward pressure with his or her own index fingers, the examiner then rocks the distal and proximal ends of the subject's nail, using the nail bed as a fulcrum.

METHODS

We used the MEDLINE database to search for English-language articles related to the clinical evaluation of clubbing that were published between January 1966 and April 1999. The Medical Subject Heading (MeSH) "hypertrophic osteoarthropathy," followed by the text word "clubbing," was used in the following search strategy: "physical examination or physical exams," "medical history taking," "professional competence," "sensitivity" and "specificity" or "sensitivity and specificity," "reproducibility of result," "observer variation," "diagnostic tests," "routine," "decision support techniques," and "Bayes theorem." This strategy resulted in a limited number of articles.

To expand the search, the titles and abstracts of all articles retrieved using the MeSH heading "hypertrophic osteoarthropathy" or the text words "clubbing" and "Hippocratic fingers" were evaluated by each author independently. According to this review, relevant publications were retrieved and their bibliographies were evaluated for additional material. We also examined standard textbooks of physical diagnosis for information on the physical examination for clubbing. We attempted to contact the authors of articles in which more than 1 observer made a determination of clubbing to obtain additional data about precision of the examination for clubbing. Studies selected for data extraction were those in which quantitative or qualitative assessment for clubbing was described in a series of patients. Although our expanded electronic search identified 567 articles related to clubbing, only 16 studies met the criteria for inclusion in our analysis.

Study Characteristics

Clubbing differs from other physical signs evaluated in The Rational Clinical Examination series in that the lack of an accepted objective diagnostic criterion standard precludes meaningful assessment of the accuracy of clinical examination. However, our review of the literature on clubbing permitted us to evaluate quantitative indices used to distinguish clubbed from normal fingers, precision of physicians' bedside clinical examination for clubbing, and accuracy of clubbing as a marker of selected diseases. We chose to limit our review of the quantitative indices of clubbing to studies of nail-fold angles and the PDR because of their potential applicability at the bedside.

Data Analysis

Pooled weighted averages were calculated for quantitative measurements of nailfold angles and PDRs from data in studies of normal and diseased populations. Using data available in 2 articles on the precision of clubbing, we calculated κ

statistics using the Stata Statistical Package (version 3.0; Computing Resource Center, Santa Monica, California). Sensitivities, specificities, and likelihood ratios (LRs) of clubbing as a marker of specific underlying disease were calculated from original data when possible.

RESULTS

Quality of the Evidence

By consensus and using criteria previously developed for this series, we appraised the quality of the evidence contained in the articles that we retrieved.²⁹ For reasons of selection bias, small sample size, and lack of an independent, blind comparison of the physical sign with a criterion standard, we classified all of the included studies as level 4, leading to grade C recommendations.²⁹

Quantitative Indices of Clubbing in Normal and Disease States

Using plaster casts, shadowgraphs, and calipers, nailfold angles and the PDR have been measured in normal populations and in subjects with diseases associated with clubbing. The precision of these quantitative techniques is high. Using the shadowgraph method, Kitis et al³⁰ examined the precision of measuring nailfold angles. Duplicate measurements of 51 subjects showed a difference of 0.2 degrees in the mean of both the hyponychial and profile angles, with SDs of 4.6 degrees and 4.3 degrees, respectively. Although Waring et al¹⁵ found that the measurement of the PDR with calipers on live fingers rather than plaster casts resulted in a loss of precision, Baughman et al²⁶ investigated intrarater reliability and found an SD of only 0.0008. In the same study, 2 observers independently measured the ratio in 20 subjects, and the maximal difference in PDR was 0.03.

Published data pertaining to the measurement of nailfold angles and the PDR in disease-free individuals are summarized in Table 14-2. The pooled weighted mean values for the profile and hyponychial angle are 167 degrees and 179 degrees, respectively. The pooled weighted mean PDR is 0.90. Do these measurements help distinguish those with from those without clubbing? The range was available for only 45 of the 161 disease-free subjects in whom the profile angle was measured, and none exceeded 176 degrees. In studies of hyponychial angles, none of the 171 disease-free subjects had angles greater than 192 degrees. The PDR has been reported in 359 disease-free subjects, and in only 1 did it exceed unity.

Table 14-3 shows the nailfold angles and PDRs in patients with diseases associated with clubbing. In such chronic diseases as cystic fibrosis and cyanotic congenital heart disease, the nailbed angles and the PDRs are significantly higher than those found in disease-free populations. In case series of asthma and COPD, PDRs are slightly higher than normal values. However, it is impossible to exclude the possibilities that these series may have included patients with other pulmonary disorders associated with clubbing or that some patients were selected because they had clubbing.

Table 14-2 Reported Values for Profile Angle, Hyponychial Angle, and Phalangeal Depth Ratio in Disease-Free Subjects

Source, y	Technique	Population	No. of Subjects	Mean Profile Angle, Degrees (SD)
Bentley et al, ¹³ 1976	Shadowgraph	Healthy subjects from a surgical clinic (age not specified)	25	168 (3.7)
Kitis et al, ³⁰ 1979	Shadowgraph	Healthy hospital employees	116	166 (4.3)
Sinniah and Omar, ³¹ 1979	Shadowgraph	Healthy children (source population not specified)	20	171 (5.5)
Pooled weighted mean			161	167 (4.4)
Hyponychial Angle, Degrees (SD)				
Regan et al, ¹² 1967	Plaster casts, planimeter	Healthy manual workers	10	186 (2.0)
Bentley et al, ¹³ 1976	Shadowgraph	Healthy manual workers	25	180 (4.2)
Kitis et al, ³⁰ 1979	Shadowgraph	Healthy manual workers	116	178 (4.6)
Sinniah and Omar, ³¹ 1979	Shadowgraph	Healthy manual workers	20	181 (5.2)
Pooled weighted mean			171	179 (4.5)
Phalangeal Depth Ratio, μm (SD)				
Waring et al, ¹⁵ 1971	Plaster casts, micrometer	Children and adults (source population not specified)	160	0.90 (0.04)
Sly et al, ²⁵ 1973	Plaster casts, micrometer	Adults (medical center personnel and relatives of patients attending pediatric allergy clinic)	60	0.90 (0.04)
Paton et al, ³² 1991	Plaster casts, micrometer	Children and adults (random sample from people playing in nearby park)	85	0.89 (0.04)
Baughman et al, ²⁶ 1998	Live fingers, calipers	Adults (medical center personnel)	54	0.92 (0.05)
Pooled weighted mean			359	0.90 (0.04)

Abbreviation: SD, standard deviation.

Table 14-3 Reported Values for Quantitative Measures of Clubbing in Disease States

Source, y	No. of Subjects	Technique	Quantitative Measure	Mean (SD)
Asthma ^a				
Waring et al, ¹⁵ 1971	45	Plaster casts	DPD/IPD ratio	39/45 <1.0 ^a
Sly et al, ²⁵ 1973	119	Plaster casts	DPD/IPD ratio	0.91 (0.05)
Bentley et al, ¹³ 1976	25	Shadowgraph	Profile angle; hyponychial angle	171 degrees (4.1 degrees); 185 degrees (6.4 degrees)
Paton et al, ³² 1991	20	Plaster casts	DPD/IPD ratio	0.91 (0.05)
Chronic Obstructive Pulmonary Disease ^b				
Baughman et al, ²⁶ 1998	54	Live fingers, calipers	DPD/IPD ratio	0.94 (0.06)
Bronchogenic Carcinoma ^b				
Baughman et al, ²⁶ 1998	109	Live fingers, calipers	DPD/IPD ratio	0.98 (0.1)
Cystic Fibrosis				
Waring et al, ¹⁵ 1971	45	Plaster casts	DPD/IPD ratio	38/45 > 1.0 ^{a,b}
Bentley et al, ¹³ 1976	50	Shadowgraph	Profile angle; hyponychial angle	179 degrees (6.2 degrees); 195 (8.3 degrees)
Lemen et al, ³³ 1978	18	Plaster casts	DPD/IPD ratio ^c	1.010 (0.016)
Pitts-Tucker et al, ³⁴ 1986	73	Shadowgraph	Hyponychial angle	192 degrees
Paton et al, ³² 1991	44	Plaster casts	DPD/IPD ratio	1.0 (0.08)
Cyanotic Congenital Heart Disease				
Waring et al, ¹⁵ 1971	27	Plaster casts	DPD/IPD ratio	18/27 > 1.0 ^{a,b}
Bentley et al, ¹³ 1976	25	Shadowgraph	Profile angle; hyponychial angle	180 degrees (4.8 degrees); 196 degrees (2.5 degrees)
Asbestos Exposure				
Regan et al, ¹² 1967	50	Plaster casts, planimeter	Hyponychial angle	195 degrees (9.6 degrees)
Crohn Disease				
Kitis et al, ³⁰ 1979	200	Shadowgraph	Hyponychial angle	184 degrees (7.8 degrees)

Abbreviations: DPD, distal phalangeal depth; IPD, interphalangeal depth.

^aIndividual values not reported; proportion of patients with DPD/IPD ratio of greater than 1.0 reported.

^bValue reported in the table is for right index finger only.

^cPooled weighted average for right index finger only.

In summary, in disease-free subjects, a PDR of more than 1 is rare, the profile angle does not exceed 176 degrees, and the hyponychial angle does not exceed 192 degrees. To facilitate clinical use, we suggest accepting values of less than 180 degrees for the profile angle (a straight line) and less than 190 degrees for the hyponychial angle as describing normality.

PRECISION AND ACCURACY

Precision of the Clinical Examination for Clubbing

Four studies³⁵⁻³⁸ have reported the precision of physicians' bedside examination for clubbing (Table 14-4). Although several of the case series describing the prevalence of clubbing in various disease states used multiple examiners, none reported interrater reliability. We have excluded from this section reports of precision that used only casts or shadow-graphs for determination of precision because potentially important clinical information from inspection or palpation of the live finger was not available to the examiners.

In an attempt to challenge the prevailing wisdom that clubbing was easily recognized, Pyke³⁵ studied the precision of physicians' global assessment for the sign. He enlisted 12 physicians and 4 medical students to examine 12 patients for the presence of clubbing. He purposefully chose patients who exhibited the full range of findings from normal to advanced clubbing. Overall agreement was fair ($\kappa = 0.39$). From the reported data, it was impossible to determine the effect of training on the examiners' precision, but it was clear that the examiners used different criteria to identify clubbing. After completion of their assessments, Pyke³⁵ asked the examiners to define clubbing, and he received a wide variety of answers.

Rice and Rowlands³⁶ used several quantitative indices, including PDRs, to assemble 11 patients who exhibited a range of findings from normal to advanced clubbing. Nineteen clinicians, all internal medicine staff or resident physicians, examined the patients for clubbing. Clubbing was judged to be

present in 103 of the 209 subject examinations. As with Pyke's³⁵ findings, observer agreement was only fair ($\kappa = 0.36$).

Precision of physical examination for a variety of signs of pulmonary disease, including clubbing, was evaluated in a study in which 24 experienced physicians examined 4 patients each.³⁷ The precision of the examination for clubbing was moderate ($\kappa = 0.45$). Although several signs showed marginally greater precision (eg, wheezes, $\kappa = 0.51$), most signs had significantly lower precision (eg, displaced trachea, $\kappa = 0.01$; whispering pectoriloquy, $\kappa = 0.11$).

A 1965 study³⁸ contrasted other reports of the precision of the physical examination for clubbing. Of 21 pulmonary signs, clubbing exhibited the highest rate of interobserver agreement among 9 experienced physicians examining 20 patients ($\kappa = 0.90$).³⁹ This high level of precision may reflect either the experience of the examiners or a selection bias because the degree of clubbing in affected patients was not described. The use of cases of more advanced clubbing may have led to an overestimation of precision.

Accuracy of Clubbing as a Marker of Disease States

Determination of the accuracy of clinical examination techniques to detect clubbing has been confounded by incorporation bias that results when the clinical examination itself forms part or all of the diagnostic criterion standard. One example of such confounding is illustrated by the digital index of Vasquez et al.⁴⁰ This index, the sum of the ratios of the distal phalangeal finger depth and interphalangeal depth circumferences in all 10 fingers, has been reported to have a high sensitivity and specificity for clubbing. However, the index was evaluated in patients with cyanotic congenital heart disease, whose clubbing was so marked that it was "obvious by simple inspection."⁴⁰ Only 1 study³⁶ measured the accuracy of clinicians' bedside examination for clubbing against a priori diagnostic criteria derived from quantitative indices in disease-free populations and those with disease. Unfortunately, data were not given in sufficient detail to allow calculation of the sensitivity and specificity of the clinical examination. Hence, data on the accuracy of clinical examination compared with the quantitative indices to detect clubbing are limited.

An alternative approach is to consider the accuracy of the presence of clubbing as a marker of underlying disease. Because many patients with clubbing have pulmonary disease, a relevant clinical question is whether clubbing separates those with COPD from those who have clubbing associated with pulmonary malignancy. In this way, 1 study²⁶ assessed the usefulness of the PDR in distinguishing patients with documented lung cancer from control subjects and those with COPD. Using calipers, Baughman et al²⁶ measured the PDR in both right and left index fingers of 109 patients with known lung cancer, 55 patients with COPD, and 54 control subjects. Of the 54 control subjects, none had a PDR in excess of 1. In those patients who had a PDR greater than 1, 40 had lung cancer and 5 had COPD alone (LR, 3.9; 95% confidence interval [CI], 1.6-9.4). Seventy patients who had a PDR of 1 or less had lung cancer, and 49 with the same depth ratio had COPD

Table 14-4 Interobserver Agreement of Clinical Examination for Clubbing

Source, y	No. of Observers	Observers' Level of Experience	κ
Pyke, ³⁵ 1954	16	4 Medical students	0.39
		4 Medical registrars	
		4 Surgical registrars	
		4 Senior physicians	
Rice and Rowlands, ³⁶ 1961	19	Residents	0.36
		Fellows	
		Staff physicians	
Smyllie et al, ³⁸ 1965	9	5 Medical registrars	0.90
		4 Consultant physicians	
Spiteri et al, ³⁷ 1988	24	2 Senior house officers	0.45
		14 Medical registrars	
		8 Consultant physicians	

alone (LR, 0.7; 95% CI, 0.6-0.8). We reclassified 1 subject in the COPD group who had a pulmonary nodule detected on chest radiography at study entry, which was subsequently diagnosed as adenocarcinoma of the lung.

These data confirm, as expected, that although a normal PDR does not rule out lung cancer, an abnormal ratio implies an increased probability (LR, 3.9; 95% CI, 1.6-9.4) of underlying lung cancer. Only 3 of the patients with COPD had a PDR greater than 1.05, and none had a ratio greater than 1.1. Among individuals with lung cancer, there was no significant difference in the prevalence of clubbing (as defined by distal phalangeal finger depth/interphalangeal finger depth ratio > 1) among the different histologic subtypes of lung cancer.

Kitis et al³⁰ investigated the association of clubbing with the activity of inflammatory bowel disease in 327 patients. Clubbing was defined as a shadowgraph-measured hyponychial angle greater than 186 degrees, which corresponded to 1.65 SDs above the mean value found in a group of 116 healthy controls. Disease activity was determined using an index incorporating the results of various laboratory investigations. The LRs for clubbing as a marker of active Crohn disease were 2.8 (95% CI, 1.8-4.1) and 3.7 (95% CI, 1.4-9.4) for ulcerative colitis. The sensitivity and specificity values were 0.58 and 0.79 for Crohn disease vs 0.30 and 0.92 for ulcerative colitis, respectively.

CLINICAL SCENARIOS—RESOLUTION

In the first case, you find that the patient appears to have increased nail curvature. You use calipers to estimate a PDR of 0.90, and on inspection, you estimate a profile angle of about 160 degrees. According to your knowledge of these values in disease-free subjects, you inform the respiratory therapist that her mother does not have clubbing. On the other hand, you find that the second patient has a PDR of 1.1 and a profile angle of 180 degrees, findings that are unusual for disease-free subjects or patients with COPD alone. You conclude that a search for bronchogenic carcinoma (or other causes of clubbing) should be undertaken.

THE BOTTOM LINE

For generations, medical students and residents have been quizzed at the bedside about the diagnostic features of clubbing. Confident though their inquisitors may be in their own ability to detect clubbing, the literature shows that interobserver agreement is only fair to moderate and that the accuracy of techniques to detect clubbing has not been well established. Nevertheless, because nonhereditary clubbing is almost always a portent of serious disease, clinicians need to be as certain as possible about its presence.

Recognizing the limitations of the studies we have appraised, we recommend the following:

In cases of diagnostic uncertainty, the PDR may be helpful. This ratio can be measured with calipers at the bedside and in disease-free populations rarely exceeds 1.0. An increased

ratio should prompt a search for underlying disease. Although patients with COPD have slightly higher ratios than do disease-free subjects, it is unusual for the ratio to exceed 1.05. A value in excess of this in a patient with COPD should prompt a search for bronchogenic carcinoma. Because most clinicians do not have calipers, visual estimation of reversal of the PDR should be assessed.

Although the accuracy of clinicians' bedside estimation of nailfold angles has not been studied, the normal values for these angles have been established. A profile angle that approaches a straight line (180 degrees) is rare in disease-free subjects and, in our opinion, is easily identifiable at the bedside. Although the normal range of the hyponychial angle has also been defined, this angle is more difficult to estimate at the bedside.

No published evidence exists as to the diagnostic yield or the optimal strategy for investigating a patient with clubbing. Therefore, after completion of a thorough medical history and physical examination, clinical judgment must guide the choice of investigations.

Author Affiliations at the Time of the Original Publication

Department of Medicine, Division of General Internal Medicine, Queen's University, Kingston, Ontario, Canada.

Acknowledgments

The authors appreciate the expert advice offered by Joseph Govert, MD, and John Whited, MD, both of Duke University, during the preparation of this article.

REFERENCES

- West S. Two cases of clubbing of the fingers developing within a fortnight and four weeks respectively. *Trans Clin Soc London*. 1897;30:60.
- Adams F. *The Genuine Works of Hippocrates*. Vol 1. New York, NY: Wm Wood & Co; 1891:206.
- Horsfall FL. Congenital familial clubbing of the fingers and toes. *CMAJ*. 1936;34:145-149.
- Buchman D, Hrowat EA. Idiopathic clubbing and hypertrophic osteoarthropathy. *Arch Intern Med*. 1955;97(3):355-358.
- Mendlowitz M. Clubbing and hypertrophic osteoarthropathy. *Medicine (Baltimore)*. 1942;21:269-306.
- Alvarez AS, McNair D, Wildman J, Hewson JW. Unilateral clubbing of the fingernails in patients with hemiplegia. *Gerontol Clin*. 1975;17(1):1-6.
- Marie P. De l'ostéo-arthropathie hypertrophante pneumique. *Rev Med*. 1890;10:1.
- Bamberger E. Über knochenvänderungen bei chronischen lungen und herzkrankheiten. *Arch Klin Med*. 1891;18:193.
- Coury C. Hippocratic fingers and hypertrophic osteoarthropathy: a study of 350 cases. *Br J Chest*. 1960;54:202-209.
- Cudowitz L, Wraith DG. An evaluation of the clinical significance of clubbing in common lung disorders. *Br J Tuberc Dis Chest*. 1957;51(1):14-21.
- Stavem P. Instrument for estimation of clubbing. *Lancet*. 1959;2(7088):7-8.
- Regan GM, Tagg B, Thompson ML. Subjective assessment and objective measurement of finger clubbing. *Lancet*. 1967;1(7489):530-532.
- Bentley D, Moore A, Swachman H. Finger clubbing: a quantitative survey by analysis of the shadowgraph. *Lancet*. 1976;2(7978):164-167.
- Mellins RB, Fishman AP. Digital casts for the study of clubbing of the fingers. *Circulation*. 1966;33(1):143-145.
- Waring WW, Wilkinson RW, Wiebs RA, Faol BC, Hilman BC. Quantitation of digital clubbing in children. *Am Rev Respir Dis*. 1971;104(2):166-174.
- Bigler FC. The morphology of clubbing. *Am J Pathol*. 1958;34(2):237-261.
- Dickinson CJ. The aetiology of clubbing and hypertrophic osteoarthropathy. *Eur J Clin Invest*. 1993;23(6):330-338.

18. Racoceanu SN, Mendlowitz M, Suck AF. Digital capillary blood flow in clubbing. *Ann Intern Med.* 1971;75(6):933-935.
19. Martinez-Lavin M. Digital clubbing and hypertrophic osteoarthropathy. *J Rheumatol.* 1987;14(1):6-8.
20. Onadeko BO, Kolawole TM. The clinical and aetiological pattern of finger clubbing and hypertrophic osteoarthropathy in Nigerians. *Trop Geogr Med.* 1979;31(2):191-199.
21. Hansen-Flaschen J, Nordberg J. Clubbing and hypertrophic osteoarthropathy. *Clin Chest Med.* 1987;8(2):287-298.
22. Rabin CB. New or neglected physical signs in diagnosis of chest diseases. *JAMA.* 1965;194(5):546-550.
23. Lovibond JL. Diagnosis of clubbed fingers. *Lancet.* 1938;1:363-364.
24. Carroll DG. Curvature of the nails, clubbing of the fingers and hypertrophic osteoarthropathy. *Trans Am Clin Climatol Assoc.* 1972;83:198-208.
25. Sly RM, Ghazanshani S, Buranakul B, et al. Objective assessment for digital clubbing in Caucasian, negro and Oriental subjects. *Chest.* 1973;64 (6):687-689.
26. Baughman RP, Gunther KL, Buchsbaum JA, Lower EE. Prevalence of digital clubbing in bronchogenic carcinoma by a new digital index. *Clin Exp Rheumatol.* 1998;16(1):21-26.
27. Schamroth L. Personal experience. *S Afr Med J.* 1976;50(9):297-300.
28. Lampe RM, Kagan A. Detection of clubbing—Schamroth's sign. *Clin Pediatr (Phila).* 1983;22(2):125.
29. Holleman DR, Simel DL. Does the clinical examination predict airflow limitation? *JAMA.* 1995;273(4):313-319.
30. Kitis G, Thompson H, Allan RN. Finger clubbing in inflammatory bowel disease. *BMJ.* 1979;2(6194):825-828.
31. Sinniah D, Omar A. Quantitation of digital clubbing by shadowgram technique. *Arch Dis Child.* 1979;54(2):145-146.
32. Paton JY, Bautista DB, Stabile MW, et al. Digital clubbing and pulmonary function abnormalities in children with lung disease. *Pediatr Pulmonol.* 1991;10(1):25-29.
33. Lemen RJ, Gates AJ, Mathe AA, et al. Relationship among digital clubbing, disease severity and serum prostaglandin F2 and E concentrations in cystic fibrosis patients. *Am Rev Respir Dis.* 1978;117(4):639-646.
34. Pitts-Tucker TJ, Miller MG, Littlewood JM. Finger clubbing in cystic fibrosis. *Arch Dis Child.* 1986;61(6):576-579.
35. Pyke DA. Finger clubbing. *Lancet.* 1954;2(6834):352-354.
36. Rice RE, Rowlands PW. *A Quantitative Method for the Estimation of Clubbing* [thesis]. New Orleans, LA: Tulane University Medical School; 1961.
37. Spiteri MA, Cook DG, Clarke SW. Reliability of eliciting physical signs in examination of the chest. *Lancet.* 1988;1(8590):873-875.
38. Smyllie HC, Blendis LM, Armitage P. Observer disagreement in physical signs of the respiratory system. *Lancet.* 1965;2(7409):412-413.
39. Armitage P, Blendis LM, Smyllie HC. The measurement of observer disagreement in the recording of signs. *J R Stat Soc A.* 1966;129:98-109.
40. Vazquez-Abad D, Pineda C, Martinez-Lavin M. Digital clubbing. *J Rheumatol.* 1989;16(4):518-520.

Prepared by Kathryn A. Myers, MD, EdM

Reviewed by Marisa D'Silva, MD

CLINICAL SCENARIO

A 24-year-old male intravenous drug user presents to the emergency department with fatigue and weight loss. Examination shows cervical lymphadenopathy, a palpable liver, and signs of recent intravenous drug use. Should the clubbing raise concerns for other diagnoses that would explain his presentation?

UPDATED SUMMARY ON CLUBBING

Original Review

Myers KA, Farquhar DRE. Does this patient have clubbing? *JAMA*. 2001;286(3):341-347.

UPDATED LITERATURE SEARCH

We searched the MEDLINE database from May 1999 to July 2004, using the same search strategy used for the original review. This resulted in a limited number of articles, so all abstracts from a MEDLINE database search using the title word "clubbing" were reviewed. We also reviewed the Citation Index (ISI Web of Knowledge and Science Citation Index Expanded) and PubMed databases for relevant articles. This strategy resulted in 2 new articles related to the diagnosis of clubbing.

NEW FINDINGS

- Digital photography is an accurate, inexpensive, and easy method for calculation of the hyponychial angle.
- The upper limit of the hyponychial angle for healthy individuals is confirmed as approximately 192 degrees; the phalangeal depth ratio (PDR) is confirmed as less than 1 in healthy individuals.
- The PDR correlates with hypoxemia and airways obstruction in cystic fibrosis.

Details of the Update

Since the original review, 2 studies have been published that used quantitative methods to assess clubbing. Husarik et al,¹ using a software angle measurement application, measured the hyponychial angle of the right index finger on digital photo-

graphs. They determined that the hyponychial angle of healthy individuals does not exceed 192 degrees, confirming the results of the original review. Data are also reported for bronchogenic carcinoma ($n = 17$), human immunodeficiency virus (HIV) disease ($n = 19$), chronic hepatitis ($n = 21$), cirrhosis ($n = 19$), pneumonia ($n = 47$), heart failure ($n = 95$), ischemic heart disease ($n = 170$), and other disorders. For a variety of illnesses, the quantitatively measured hyponychial angle is significantly greater than seen in patients without an abnormal hyponychial angle. However, fewer than 25% of patients with these illnesses have hyponychial angles that exceed the normal upper range of 192 degrees, making the presence of clubbing an insensitive diagnostic marker. Patients with emphysema ($n = 9$) and acquired valvular heart disease ($n = 81$) were not different from patients without the disease ($P < .13$).

In a second study, Nakamura et al² examined the PDR of 100 healthy subjects and 100 patients with cystic fibrosis using plaster finger casts. The mean PDR of healthy controls was 0.90 (SD, 0.037), and no values exceeded unity, confirming the results of the original review. Among patients with cystic fibrosis, the presence of clubbing predicts those who will be hypoxic. Similarly, the absence of clubbing made hypoxemia much less likely.

No additional studies have compared quantitative measures of clubbing to physicians' bedside assessments, and no studies have evaluated the diagnostic yield or the optimal strategy for investigating a patient with clubbing.

IMPROVEMENTS IN THE DATA PRESENTED IN THE ORIGINAL PUBLICATION

Additional studies allow us to reestimate the normal PDR and the hyponychial angle. The normal PDR is 0.90 (95% confidence interval [CI], 0.89-0.90); the normal hyponychial angle is 181 degrees (95% CI, 178-183 degrees). Patients with a hyponychial angle greater than 192 degrees would be considered to have an abnormal hyponychial angle.

CHANGES IN THE REFERENCE STANDARD

During the past century, nail fold angle measurements using unguisometers, shadowgrams, or plaster casts of fingers have been proposed as the reference standard for clubbing. Because these methods are cumbersome and time consuming, the clin-

cal examination by experienced physicians has been accepted as establishing a diagnosis of clubbing.

RESULTS OF LITERATURE REVIEW

See Table 14-5.

Table 14-5 Univariate Findings Concerning Clubbing

Finding (n = 1)	Disorder	LR+ (95% CI)	LR- (95% CI)
Clubbing in cystic fibrosis	Hypoxemia	3.2 (1.9-6.4)	0.13 (0.06-0.27)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio, LR-, negative likelihood ratio.

EVIDENCE FROM GUIDELINES

No guidelines advocate the routine assessment of clubbing.

CLINICAL SCENARIO—RESOLUTION

Test results for HIV and hepatitis C virus were positive. Chest and abdominal imaging results were unremarkable. A thyroid-stimulating hormone level was normal. Although clubbing has been associated with both HIV and hepatitis C viral infections, it has also been associated with endocarditis. You reexamine the patient and find no heart murmur, fever, or stigmata of endocarditis, but you obtain blood cultures and echocardiography to rule out endocarditis.

CLUBBING—MAKE THE DIAGNOSIS

PREVALENCE OF CLUBBING

The probability of clubbing depends on the underlying illness. The frequency of a quantitatively measured hyponychial angle greater than 192 degrees has been reported in the illnesses listed in Table 14-6.¹

POPULATION FOR WHOM CLUBBING SHOULD BE CONSIDERED

Clubbing can occur in a variety of illnesses. It should be considered among patients with cystic fibrosis or bronchiectasis as a marker for chronic hypoxemia. In patients with clubbing that is not congenital, it would be reasonable to obtain a chest radiograph to look for pulmonary conditions associated with clubbing.

REFERENCE STANDARD TESTS

The pragmatic standard is examination by an experienced clinician, although laborious quantitative measures can be done as part of a research study.

Table 14-6 Prevalence of Hyponychial Angle Exceeding Upper Range of Normal Among Various Conditions

Condition	Prevalence, % (95% CI)
Pneumonia	23 (11-36)
HIV disease	16 (0-32)
Cirrhosis	16 (0-32)
Chronic obstructive pulmonary disease	15 (6-23)
Chronic hepatitis	14 (0-29)
Bronchogenic carcinoma	12 (0-27)
Pulmonary hypertension	10 (0-20)
Acquired valvular heart disease	9 (3-15)
Congestive heart failure	9 (4-15)
Ischemic heart disease	9 (5-14)
Solid tumor malignancy	6 (1-11)

Abbreviations: CI, confidence interval; HIV, human immunodeficiency virus.

REFERENCES FOR THE UPDATE

1. Husarik D, Vavricka SR, Mark M, Schaffner A, Walter RB. Assessment of digital clubbing in medical inpatients by digital photography and computerised analysis. *Swiss Med Wkly*. 2002;132(11-12):132-138.^a
2. Nakamura CT, Ng GY, Paton JY, et al. Correlation between digital clubbing and pulmonary function in cystic fibrosis. *Pediatr Pulmonol*. 2002;33(5):332-338.^a

^aFor the Evidence to Support the Update for this topic, see <http://www.JAMAevidence.com>.

EVIDENCE TO SUPPORT THE UPDATE:

14

Clubbing

TITLE Assessment of Digital Clubbing in Medical Inpatients by Digital Photography.

AUTHORS Husarik D, Vavricka SR, Mark M, Schaffner A, Walter RB.

CITATION *Swiss Med Wkly*. 2002;132(11-12):132-138.

QUESTION Can digital photography reliably assess the hyponychial angle of healthy controls and medical inpatients, and what is the range of angles associated with various medical diseases?

DESIGN The right index finger was digitally photographed, and a software angle measurement application was used to calculate the hyponychial angle. Three investigators performed the measurements on each finger. The patients' underlying medical diagnoses were obtained through chart review.

SETTING Medical inpatient ward in Switzerland; healthy controls (population not specified).

PATIENTS Five hundred fifteen patients admitted as general medical inpatients and 123 healthy controls.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

Digital photography and a software angle measurement application were used to assess the hyponychial angle. Interrater and intrarater reliability was calculated.

MAIN OUTCOME MEASURES

Hyponychial angles of patients and controls; angles by disease category as ascertained through chart review.

MAIN RESULTS

Measurement of the hyponychial angle with this technique demonstrated high intrarater and interrater reliability. Pro-

file angles and phalangeal depth ratios were not measured. The investigators did not compare their proposed angle for diagnosis of clubbing (192 degrees) with physician assessment of clubbing. Each patient's medical record was reviewed for diagnoses associated with clubbing. The investigators found that mean hyponychial angles were significantly increased compared with that of healthy controls for most diagnoses, even those not traditionally associated with clubbing, such as ischemic heart disease. When compared with other patients in their study, only those with chronic obstructive pulmonary disease, cystic fibrosis, cirrhosis, chronic hepatitis, and human immunodeficiency virus infection had significantly increased hyponychial angles. Approximately 15% of the patients with chronic obstructive pulmonary disease, cirrhosis, chronic hepatitis, and human immunodeficiency virus infection had hyponychial angles greater than 192 degrees.

CONCLUSIONS

LEVEL OF EVIDENCE Level 3.

STRENGTH Objective, reliable method for determine hyponychial angle.

LIMITATIONS Although the angles may have been determined in a blinded fashion from underlying disease status, it is not explicitly stated that the chart review was blinded to the presence/absence of clubbing.

Previous studies of clubbing have used cumbersome and impractical techniques, such as plaster finger casts and shadowgrams, to measure nailfold angles. This study reports a new, reliable technique for the quantitative assessment of clubbing. The data in healthy controls confirm that the hyponychial angle does not exceed 192 degrees. Although the mean hyponychial angle of patients with a wide variety of medical diagnoses exceeded that of controls, the prior probability of clubbing as defined by an angle greater than 192 degrees was less than 20% for most conditions. Physicians' bedside assessment of clubbing was not evaluated.

Reviewed by Kathryn A. Myers, MD

TITLE Correlation Between Digital Clubbing and Pulmonary Function in Cystic Fibrosis.

AUTHORS Nakamura CT, Ng GY, Paton JY, Keens TG, Witmer JC, Bautista-Bolduc D, Woo MS.

CITATION *Pediatr Pulmonol*. 2002;33(5):332-338.

QUESTION Does digital clubbing in patients with cystic fibrosis predict hypoxemia and airflow limitation?

DESIGN Plaster casts of 100 patients and 100 healthy controls were created to allow measurement of the phalangeal depth ratio (PDR) of the right index finger. The PDR was compared to oxygen levels and pulmonary function tests.

SETTING Los Angeles Childrens' Hospital.

PATIENTS Patients with cystic fibrosis and without rheumatologic or cyanotic congenital heart disease were included. Controls were recruited from unrelated visitors of hospital patients and employees and their families.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

The PDR of each subject was measured independently by 2 investigators using a micrometer on the right index finger cast. An average of the 2 measurements constituted the clubbing index. A clubbing index greater than 1.00 was defined as clubbing.

MAIN OUTCOME MEASURE

Correlation of clubbing with hypoxemia and pulmonary function (Table 14-7).

Table 14-7 Clubbing in Cystic Fibrosis Predicts Hypoxemia

Finding	LR+ (95% CI)	LR- (95% CI)
Clubbing in cystic fibrosis	3.2 (1.9-6.4)	0.13 (0.06-0.27)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

MAIN RESULTS

The PDR of healthy controls was 0.90 (SD, 0.04; range, 0.81-0.97), and none exceeded 1.0. Seventy-five of the patients with cystic fibrosis were defined as having clubbing (PDR > 1.0). Of the 25 patients without clubbing, the forced expiratory volume in 1 second (FEV₁) was 69% predicted, whereas those with clubbing had an FEV₁ of 45% predicted. The PDR was inversely correlated with hypoxemia ($r = -0.56$; $P < .001$) in patients with cystic fibrosis.

CONCLUSIONS

LEVEL OF EVIDENCE Level 3.

STRENGTHS Reliable, quantitative measure to assess clubbing.

WEAKNESS It is not explicitly stated that the clinicians were blinded to the hypoxemia status.

The presence of clubbing in patients with cystic fibrosis is associated with hypoxemia, and its absence made hypoxemia much less likely. The ability of physicians to detect clubbing at the bedside using the diagnostic standard of PDR greater than 1 was not evaluated by the investigators.

Reviewed by Kathryn A. Myers, MD

CHAPTER 15

Is This Patient Taking the Treatment as Prescribed?

Barbara J. Stephenson, RN

Brian H. Rowe, MD, MSc

R. Brian Haynes, MD, PhD

William M. Macharia, MD, MSc

Gladys Leon, MD, MSc

CLINICAL SCENARIOS

CASE 1 A 28-year-old woman presents to the emergency department in acute distress, with a 3-day history of worsening asthma. Her prescribed medications include an inhaled β_2 agonist and an inhaled steroid. When questioned, she breathlessly admits to “occasionally” missing her medications “maybe only once or twice.”

CASE 2 A 55-year-old man with posttraumatic seizure disorder has been taking phenytoin since his injury. His seizures were initially adequately controlled but he recently has been having weekly seizures. In an office visit he resentfully denies missing any of his medication.

THE IMPORTANCE OF CLINICAL EXAMINATION

Physicians should measure compliance for patients prescribed a self-administered treatment because noncompliance is common and physicians can help patients improve their compliance^{1,2} and increase the benefit they derive from therapy. Compliance with long-term self-administered medication therapy is approximately 50% for those who remain in care.³ There is a wide range of compliance among patients, from 0% to 100%. This average compliance rate of 50% provides only the most limited picture of compliance; in general, there is substantial undercompliance. Furthermore, compliance by individuals can vary considerably over time. Compliance rates for short-term self-administered therapies average about 75% initially but decrease to less than 25% for the completion of antibiotic therapy for acute infections. Aside from its potential for undermining the effectiveness of any treatment, noncompliance is associated with poorer prognosis.⁴

Table 15-1 depicts combinations of treatment outcome and compliance that present in clinical practice and need to be distinguished from one another to initiate the appropriate intervention. The bottom right cell, D, represents the most desirable state: high compliance with achievement of the treatment goal. Patients who fall into cell A (low compliance and suboptimal achievement of the treatment goal) are in need of efforts to promote compliance. Patients in cell B (high compliance without achieving the treatment goal) require more or better treatment, whereas those in cell C (achievement of the treatment goal despite low compliance) need less treatment prescribed or may actually have been misdiagnosed or mistreated and do not merit intervention to increase compliance, at least until the need for treatment is reassessed. The aim of compliance assessment, along with other diagnostic tests, is to categorize patients into the appropriate cells. When it has been determined that patients occupy cell A, B, or C, phy-

sicians may then alter treatment to attempt to move patients into cell D.

The 2 cases illustrate the importance of distinguishing between noncompliance and lack of therapeutic efficacy. On closer questioning, the first patient revealed that she had been using her inhaled steroid sporadically and gradually lost control of her asthma without change in extrinsic allergic stimuli. She initially provided 2 useful clinical clues to important deviation from her prescribed regimen: worsening symptoms while prescribed usually adequate therapy and admission of “occasional” noncompliance. Reinstituting her usual regimen and reinforcing the need for compliance, particularly with the inhaled steroid, improved her treatment results.

Case 2 required a different solution: the phenytoin levels eventually proved to be in the therapeutic range, confirming the patient’s compliance, and a second medication was required to improve long-term control. Accurate assessment of compliance by questioning could prevent overdosing the patient with additional therapy on the assumption of noncompliance and permit timely addition of the second therapy.

Although maintaining an adequate level of compliance is central to deriving benefit from any efficacious therapy, the degree of compliance necessary to achieve a measurable benefit from specific medications is variable. Haynes et al⁵ found that 80% compliance was necessary to achieve a reduction in blood pressure from antihypertensive therapy with the types and doses of medication that were prescribed by primary care physicians, whereas Markowitz⁶

reported that children receiving only a third of their prescribed penicillin had substantial protection from recurrences of rheumatic fever. The thresholds of compliance for acceptable therapeutic effects are not known for most regimens.

THE NATURE OF NONCOMPLIANCE

On a practical level, it is not difficult to imagine why noncompliant behavior occurs. Patients often find medical regimens complicated, inconvenient, embarrassing, or expensive. Particularly for chronic disorders, the short-term disadvantages frequently outweigh the long-term advantages.

At a theoretic level, the nature and determinants of noncompliant behavior are more complex and not well understood, although there are interesting models.⁷ Numerous studies of the “determinants” of compliance have led to the following generalizations.⁸ Sociodemographic factors such as age, sex, race, intelligence, and education have little to do with compliance. Low compliance is a problem with self-administered treatments for all disorders, but patients with psychiatric problems are less likely to comply, and those with (other) disabilities caused by disease are more likely to comply. Long waiting times at clinics and long gaps between appointments lead to patients’ missing appointments and dropping out of care. The more complex or costly the regimen and the longer its duration, the less the compliance.

MEASURING NONCOMPLIANCE

Most studies determining the limitations and strengths of clinical information about compliance include pill counts and measurement of serum levels of drugs or tracers. Special medication monitors can also reveal patterns of medication consumption that cannot be obtained by other means. None of these more accurate methods is likely to be handy to practitioners for most therapeutic regimens.

No clinical measurement of compliance approaches perfection, but clinical information can be used to narrow down the situations in which compliance measurement is most likely to be important for the care of the patient. A 3-step sequence will identify most noncompliers.

1. Nonattendance at appointments. Dropout rates are high with many treatments, and nonattendance at a scheduled appointment is the first step astray.
2. Lack (or loss) of responsiveness to a usually (or previously) adequate dose of treatment. These patients are most in need of further assessment of their compliance to separate problems of therapy from those of compliance. Cases 1 and 2 would qualify for this route.
3. For patients whose compliance is in doubt, particularly those who come to attention through steps 1 and 2, use the most appropriate method(s) from Table 15-2. Direct measures of medication consumption are most accurate, but they are available for only a small number of medications, can indicate spuriously high compliance

Table 15-1 Complications and Treatment Response^a

Achievement of the Treatment Goal	Compliance Rate	
	Low	High
No	A: The target group	B: Inadequate therapy?
Yes	C: Unnecessary therapy?	D: Ideal

^aBased on Sackett et al.¹

Table 15-2 Methods to Measure Compliance

Direct
Drug or metabolite levels
Blood
Urine
Saliva
Tracer compounds
Indirect
Appointment keeping
Therapeutic response
Self-report
Pill count
Pharmacy records
Medication event monitors

if the patient takes the prescribed dose only during the time leading up to assessments of drug levels, and do not apply to most nonmedication regimens, such as weight-loss diets. Even when available, they take time and money to obtain and are unlikely to help in such situations as the acute care of a patient in the midst of a crisis.

Questioning of patients is the most widely applicable method of measuring compliance. Careful questioning will identify more than half of those who are noncompliant without falsely labeling many of the compliers. Patients should be asked to indicate, without prompting, exactly what medications they are taking and when they are taking them. This may reveal a different understanding of and adherence to the regimen than was prescribed. For patients who report a generally correct understanding of their prescription, the details of any noncompliance should be sought. The method of asking likely affects the accuracy of the response. Studies assessing the value of patient self-report have used a nonjudgmental, nonthreatening approach, prefacing the question with a remark such as the following⁹: “People often have difficulty taking their pills for one reason or another.” The question also must be asked in a particular way: “Have you ever missed any of your pills?” If the answer is affirmative, then ask the patient to estimate how many pills he or she has missed during the previous day and week. The interview can also provide insight into the possible reasons for noncompliance. This valuable clinical information can allow prompt reevaluation of the current regimen if the information is interpreted appropriately. It is essential to take into account that even if the patient admits to missing any medication during the previous day or week, he or she will still tend to overestimate the actual rate of compliance (by an average of 17% in 1 study⁹).

Clinical measures of compliance sometimes can be supplemented or replaced by other methods. Some regimens produce telltale adverse effects, the absence of which suggests low compliance; for example, increased urinary frequency with the initiation of diuretics, dry mouth with anticholinergics, slow heart rate with β -blockers, dark stool with oral iron, and suppression of thyrotropin (thyroid-stimulating hormone) with thyroid hormone replacement.¹⁰ Blood level measurements are routinely available for some medications, and these can be used for monitoring compliance, particularly when the serum half-life is relatively prolonged.¹¹ When patients receive all their medications through a single pharmacy, pharmacy records can provide an indirect measure of compliance.¹² Medication event monitors, although providing unique information about the pattern of medication taking, are expensive and remain a research tool.^{13,14} Tracers, either harmless substances such as riboflavin¹⁵ or minute amounts of medications such as phenobarbital that can be easily measured,¹⁶ are also research tools.

For pill counts, drug and tracer levels, and surreptitious pill monitors, there are ethical issues to be addressed. Because they invade the patient’s privacy and can be used to usurp autonomy, when possible you should inform the patient of their intended use and ask for consent before using

them.¹⁷ Patients usually agree to monitoring if it is explained that the purpose of the assessments is to help better understand how they are taking their medicine.

ACCURACY OF CLINICAL MEASURES OF COMPLIANCE

Clinical judgment of compliance has been found wanting in almost every study in which it has been tested. Clinicians who believe that they are exceptions to this finding because they know their patients well should take heed of a study by Gilbert et al.¹⁸ Primary care physicians were asked to give compliance estimates only for patients they thought they knew well. The sensitivity of clinical judgment for detecting noncompliance was an embarrassing 10%, and overall performance by clinicians was not better than if they had flipped coins instead of applying their “clinical judgment.” Physicians should not trust their unaided judgment regarding the compliance by individual patients.

Studies have shown only a low-order correlation between nonattendance and noncompliance with self-administered treatments, but this is at least partly an artifact of nonattenders’ being frequently unavailable for compliance studies. For example, in one study, patients keeping all appointments appeared to be less compliant with antacid and anticholinergic medications for peptic ulcer therapy than patients who missed some appointments, but only 96 (60%) of the 160 patients had complete follow-up assessments.¹⁹ Richardson et al²⁰ confirmed both that attendance does not ensure compliance with medications and that compliance is even worse among nonattenders: of patients keeping more than 60% of their scheduled clinic appointments, 40% were found to be noncompliant with medication by urine metabolite measurement, whereas 95% of patients with lower appointment compliance demonstrated low compliance.

The patient’s response to therapy is also only weakly related to compliance for many treatments,⁹ but it can be useful when combined with other methods. For example, when Inui et al²¹ treated patients as noncompliant if they either admitted noncompliance or had uncontrolled pressures, this combined compliance test had a sensitivity of 83% and a specificity of 66%.

Questioning patients about their compliance is the most readily available, valid method of measuring compliance in clinical practice. To review the literature on self-report, we used previously published guidelines for collecting studies and preparing meta-analyses.²² We identified studies comparing self-report with other measures of compliance and uncovered many studies comparing self-report with pill counts. The 4 studies with the strongest research methods^{9,13,18,23} are summarized in Table 15-3. The results of compliance tests were considered positive if they uncovered noncompliance and negative if they verified compliance. In these studies, self-report yielded a sensitivity of 55%, a specificity averaging 87%, and a likelihood ratio for a positive test result of 4.4 on average. Patients’ reports of compliance with medication were less useful because the patients still may have been noncompliant (likelihood ratio for a negative test result, 0.5). In one study,⁹

Table 15-3 Pooled Data From Methodologically Strong Studies Comparing Pill Count With Self-report^{9,13,18,23}

Self-report	Pill Count	
	Noncompliant ^a	Compliant
Missed ≥ 1	152	34
None missed	122	232

^aNoncompliant was defined as taking less than 80% of pills,^{9,18} less than 100% of pills,¹³ and less than 75% of pills.²³

self-report outperformed several direct and indirect measures of compliance. Similarly, when Fletcher et al²⁴ compared the usefulness of interview, pill count, and measurement of serum drug levels of digoxin, they found that interviewing was the most useful method. Unfortunately, to our knowledge there are no studies to date assessing the agreement among clinicians on eliciting compliance information from patients in usual settings or of the effect on self-reports of repeatedly questioning patients about their compliance.

Counting the patient's pills is valid for single assessments at the patient's home if the purpose of the visit is not revealed in advance and if care is taken to determine the amount of medication that has been dispensed, the date the most recent prescription refill was begun, how much was left over from the previous prescription when the current prescription was begun, whether there has been any change in the prescription not noted on the pill container, and whether the patient has caches of pills in other locations or has shared them with relatives or friends.⁹ When all factors are taken into account, the pill count compares favorably with serum drug levels.¹⁸ However, pill counts of this rigor are impractical in most clinical settings, and pill counts performed on medications patients bring with them to clinic visits overrepresent compliance when compared with more tamper-proof methods, such as special pill containers that electronically monitor each dose as it is removed.^{13,14} The latter devices also show patterns of compliance that cannot be detected by simple pills counts, including increased compliance just before and after appointments and decreasing compliance between appointments.

Although the absence of common adverse effects may be an indication of noncompliance, the link between compliance and adverse effects is either unknown or relatively tenuous. For example, for patients prescribed diuretics, the sensitivity for noncompliance of reductions in serum potassium level was 82% but the specificity was only 48%.⁹

THE BOTTOM LINE

You can detect most noncompliant patients by watching for nonattenders, watching for nonresponders, and asking nonresponders about their compliance. In addition to clarifying problems of undertreatment and overtreatment, information about patient compliance permits the efficient application of effective methods of increasing compliance.^{1,2}

Author Affiliations at the Time of the Original Publication

From the Design, Measurement, and Evaluation Program (Drs Rowe, Haynes, Macharia, and Leon and Ms Stephenson) and the Health Information Research Unit (Dr Haynes), McMaster University, Hamilton, Ontario, Canada. Ms Stephenson is now with the Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario, Canada; Dr Rowe is now with the University of Ottawa, Northeastern Family Medicine Program, Sudbury, Ontario, Canada; Dr Macharia is now with the Department of Pediatrics, Faculty of Medicine, University of Nairobi, Kenya; Dr Leon is now with the Department of Dermatology, Faculty of Medicine, National Autonomous University of Mexico, Mexico City.

Acknowledgments

This study was supported by the Brian C. Decker Health Informatics Research Fund, Hamilton, Ontario, Canada; a Health Scientist Award from the National Health Research and Development Program, Canada (Dr Haynes); and clinical epidemiology fellowships from the International Clinical Epidemiology Network Program, Philadelphia, Pennsylvania (Drs Macharia and Leon).

REFERENCES

- Sackett DL, Haynes RB, Guyatt GH, Tugwell P. Helping patients follow the treatments you prescribe. In: Sackett DL, Haynes RB, Guyatt GH, Tugwell P, eds. *Clinical Epidemiology: A Basic Science for Clinical Medicine*. 2nd ed. Boston, MA: Little Brown & Co; 1991:249-281.
- Macharia WM, Leon G, Rowe BH, Stephenson BJ, Haynes RB. An overview of interventions to improve compliance with appointment keeping for medical services. *JAMA*. 1992;267(13):1813-1817.
- Sackett DL, Snow JC. The magnitude of compliance and noncompliance. In: Haynes RB, Taylor DW, Sackett DL, eds. *Compliance in Health Care*. Baltimore, MD: Johns Hopkins University Press; 1979:11-22.
- Haynes RB, Dantes R. Patient compliance and the design and interpretation of clinical trials. *Control Clin Trials*. 1987;8(1):12-19.
- Haynes RB, Gibson ES, Taylor DW, Bernholz C, Sackett D. Process versus outcome in hypertension: a positive result. *Circulation*. 1982;65(1):28-33.
- Markowitz M. Eradication of rheumatic fever: an unfulfilled hope. *Circulation*. 1970;41(6):1077-1084.
- Leventhal H, Cameron L. Behavioral theories and the problem of compliance. *Patient Educ Couns*. 1987;10:117-138.
- Haynes RB. Determinants of compliance: the disease and the mechanics of treatment. In: Haynes RB, Taylor DW, Sackett DL, eds. *Compliance in Health Care*. Baltimore, MD: Johns Hopkins University Press; 1979:49-62.
- Haynes RB, Taylor DW, Sackett DL, Gibson ES, Bernholz C, Mukherjee J. Can simple clinical measurements detect patient noncompliance? *Hypertension*. 1980;2(6):757-764.
- England ML, Hershman JM. Serum TSH concentration as an aid to monitoring compliance with thyroid hormone in hypothyroidism. *Am J Med*. 1986;292(5):264-266.
- Mollica JA. Monitoring compliance through analysis of drug and metabolite levels. *Control Clin Trials*. 1984;5(4 suppl):505-509.
- Sclar DA. Improving medication compliance: a review of selected issues. *Clin Ther*. 1991;13(4):436-440.
- Cramer JA, Mattson RH, Prevey ML, Scheyer RD, Ouellette VL. How often is medication taken as prescribed? *JAMA*. 1989;261(22):3273-3277.
- Rudd P, Ahmed S, Zachary V, Barton C, Bonduelle D. Improved compliance measures: applications in an ambulatory hypertensive drug trial. *Clin Pharmacol Ther*. 1990;48(6):676-685.
- Deuschle K, Jordahl C, Hobby G. Clinical usefulness of riboflavin-tagged isoniazid for self-medication in tuberculosis patients. *Am Rev Respir Dis*. 1960;82:1-10.

16. Wolff K, Hay A, Raistrick D, Calvert R, Feely M. Measuring compliance in methadone maintenance patients: use of a pharmacologic indicator to estimate methadone plasma levels. *Clin Pharmacol Ther.* 1991;50(2):199-207.
17. National Heart, Lung, and Blood Institute Working Group on Patient Compliance. Management of patient compliance in the treatment of hypertension. *Hypertension.* 1982;4(3):415-423.
18. Gilbert JR, Evans CE, Haynes RB, Tugwell P. Predicting compliance with a regimen of digoxin therapy in family practice. *CMAJ.* 1980;123(2):119-122.
19. Roth HP, Caron HS, Hsi BP. Measuring intake of a prescribed medication: a bottle count and a tracer technique compared. *Clin Pharmacol Ther.* 1970;7:49-54.
20. Richardson JL, Skilton DR, Krailo M, Levine AM. The effect of compliance with treatment on survival among patients with haematologic malignancies. *J Clin Oncol.* 1990;8(2):356-364.
21. Inui TS, Carter WB, Pecoraro RE. Screening for noncompliance among patients with hypertension: is self-report the best available measure? *Med Care.* 1981;19(10):1061-1064.
22. Oxman AD, Guyatt GH. Guidelines for reading literature reviews. *CMAJ.* 1988;138(8):697-703.
23. Stewart M. The validity of an interview to assess a patient's drug taking. *Am J Prev Med.* 1987;3(2):95-100.
24. Fletcher SW, Pappius EM, Harper SJ. Measurement of medication compliance in a clinical setting. *Arch Intern Med.* 1979;139(6):635-638.

This page intentionally left blank

Prepared by Hayden B. Bosworth, PhD

Reviewed by Amy Rosenthal, MD

CLINICAL SCENARIO

A 65-year-old woman prescribed a diuretic and an angiotensin-converting enzyme inhibitor continues to have inadequate blood pressure control. She volunteers that she takes her medicines exactly as instructed on her bottles but did not bring the bottles with her. You are considering adding another antihypertensive medication, but she is already taking 2 other medications for diabetes. Can you be confident that she is taking her medications as prescribed? Is there a way to get a better history of her medication adherence?

UPDATED SUMMARY ON COMPLIANCE AND MEDICATION ADHERENCE

Original Review

Stephenson BJ, Rowe BH, Haynes RB, Macharia WM, Leon G. The rational clinical examination: is this patient taking the treatment as prescribed? *JAMA*. 1993;269(21):2779-2781.

UPDATED LITERATURE SEARCH

We searched MEDLINE, Current Index to Nursing and Allied Health Literature, and Psychinfo using the following key terms: "compliance.mp." [mp = ti, ot, ab, rw, sh, it, tc, id], "adherence.mp." [mp = ti, ot, ab, rw, sh, it, tc, id], and (sensitivity and specificity).tw. We limited our search to English-language publications and obtained 30 publications in the period between 1993 and February, 2005. In addition, we reviewed the references of the selected publications and found 1 additional publication. We limited our review to only those that compared measures of medication adherence to Medication Event Monitoring System (MEMS) caps, a validated electronic monitor of pill adherence (see below), which resulted in 6 articles. However, 2 of these studies had fewer than 10 nonadherent patients and the quality levels for the studies were inadequate (quality score < 3) for inclusion.^{1,2}

NEW FINDINGS

- Complex questionnaires for assessing adherence may be no more efficient than the simple question, Have you missed any pills in the past week?

Details of the Update

Medications are the most common medical intervention. Adherence to medications is often suboptimal and nonadherence is associated with adverse health outcomes, as well as medical, social, and economic consequences.³⁻⁵ Nonadherence with therapeutic medication recommendations is prevalent. Across different definitions of nonadherence, approximately 50% of patients do not take their prescribed medications as recommended.⁶⁻⁹ The true rate of nonadherence may be higher because patients with a history of nonadherence are likely underrepresented in outcomes research.

Clinicians must frequently rely on their own judgment but unfortunately demonstrate no better than chance accuracy in predicting the medication adherence of their patients.¹⁰ Clinicians may conduct pill counts or review pharmacy records if available. The former method of assessing patient medication adherence is potentially problematic because, apart from being intrusive, it does not give any indication of when the medication was taken or whether it was thrown away and thus may result in overestimation of adherence. Pharmacy refill records provide a reliable and nonintrusive longitudinal measure of medication adherence when the patient receives all their medication from a centralized pharmacy such as that of the Department of Veterans Affairs or private sector health maintenance organizations. In addition, this method of assessing medication adherence requires extensive data tracking programs.

In general, patients tend to overestimate their medication adherence¹¹ and, unless a patient is not responding to therapy, it may be difficult to identify poor medication adherence. Asking patients about their medication use is often the most practical means of ascertainment, but it is prone to inaccuracy. A key validated question is, Have you missed any pills in the past week? and any indication of having missed 1 or more pills signals a problem with low adherence.¹² Compared to pill counts as the reference standard, asking nonresponders about their medication adherence by using this single question will detect 55% of those with less than complete adherence, with a specificity of 87% (positive likelihood ratio [LR], 4.3; 95% confidence interval [CI], 3.1-6.1; negative LR, 0.51; 95% CI, 0.44-0.58).¹⁰ Other practical measures to assess adherence include watching for those who do not respond to increments in treatment intensity and patients who fail to attend appointments. Additional practical meth-

ods include review of pill bottles and, when available, checking on fill dates and pill counts. Finally, simply asking the patients to describe their medication regimen such as when they take their medication and what it is for can often be informative.

IMPROVEMENTS IN THE DATA PRESENTED IN THE ORIGINAL PUBLICATION

The literature search was conducted without restriction to year, but we focused on studies that compared self-reported measures of adherence to electronic monitors of pill adherence. The update provides LRs for questionnaires designed to detect nonadherence to medication, using an alternative reference standard.

CHANGES IN THE REFERENCE STANDARD

The absence of a singular conceptual basis of medication adherence is problematic. Strategies to improve adherence can be evaluated only within the context of a given definition. Furthermore, comparative assessment of the adherence literature is difficult across studies using different definitions and methods of operationalizing adherence. A commonly used but arbitrary measure of optimal adherence has been identifying patients who take at least 80% of prescribed doses correctly.^{13,14} In other words, patients who take at least 80% of doses correctly are considered adherent. This level has not been validated in all circumstances and may vary, depending on several factors, including, for example, the half-life of the prescribed compound.¹⁵ Adherence to medication is not a dichotomy, and patients can demonstrate a wide variety of patterns of medication use.

The assessment of adherence is a complex task, and there is no gold standard, with the exception of actually observing an individual taking the prescribed medication. Researchers interested in measuring medication adherence often rely on one of 6 measures of adherence: pharmacy refills, pill counts, electronic measures (eg, MEMS caps), biologic indices, self-report, and physician judgments. Because of the disparate metrics used by investigators, comparison between methods (eg, self-report vs pharmacy records) or even across studies that use the same methods is difficult. Although there may not be a “best” measurement strategy to obtain an approximation of adherence behavior, strategies used must meet basic psychometric standards or acceptable reliability and validity properties.

Direct methods for assessing medication adherence include those that are more objective and require limited interpretation. Electronic measurement devices are considered the closest to a reference standard, and reviewed studies were limited to those that used electronic monitors. Electronic monitors, including the MEMS (AARDEX [APREX] Ltd, Union City, California) consist of a microprocessor placed in a medication container with a switch that is activated by the interruption of an electric current. When activated, the microprocessor records the date and time the bottle was opened. Several months of data can be stored on these units before they must be downloaded onto a computer. These

medication monitors can provide information on the pattern of drug intake, including the frequency and timing of medication dosing during a fairly extended period. Electronic monitors are not widely available and are expensive. They preclude the use of a pillbox to organize the medication being monitored by the electronic cap. In addition, some patients remove more than 1 dose per bottle opening to avoid carrying medication bottles when leaving home. These limitations may result in electronic monitoring underestimating a patient’s actual adherence. Electronic monitored adherence rates consistently range between 10% and 20% lower than rates assessed by other methods, including self-reports¹⁶ and pill counts.¹⁷

Indirect methods for measuring medication adherence involve interpretations that are more subjective and often based on an individual’s perception of adherence. Because indirect measurements of adherence, specifically self-report measures, continue to be the most commonly used measure because they are simple, inexpensive, and convenient to use,¹⁸ the current review will focus on the diagnostic properties of these measures.

There are 3 basic types of patient self-report: questionnaires, interviews (in person or by telephone), and self-monitoring logs (eg, diaries). Questionnaire-based measures include multi-item scales (summarized below), visual analog scales, or reports of missed doses. Maintaining confidentiality of the data and promoting a cooperative relationship between patients and the study team that collects the data can maximize the accuracy of patients’ self-reported adherence. These procedures make it less likely that patients will be defensive and deliberately distort their responses or that communication problems would otherwise render assessments inaccurate, as is particularly a concern when patient adherence reports are collected by health professionals themselves.¹⁹

RESULTS OF LITERATURE REVIEW

The results of the literature review are summarized in Table 15-4. The first study²⁰ examined a specific self-reported survey in comparison to MEMS caps among patients with human immunodeficiency virus infections. The self-report questionnaire was the Medication Adherence Self-Report Inventory, which consists of 12 items with 2 broad themes. The first section of this measure assessed the amount of medication actually taken, and the second part addressed the time of doses. The investigators selected the antiretroviral drug from the patient’s regimen that presented the greatest barrier to adherence (eg, higher pill burdens, dietary requirements, more frequent dose intervals). A second study²¹ examined the 19-item Compliance Questionnaire Rheumatology against MEMS caps among 81 patients with rheumatoid arthritis who were taking nonsteroidal anti-inflammatory drugs. The third study²² examined the relationship between the 6-item Simplified Medication Adherence Questionnaire and medication event monitoring among 40 patients using nelfinavir. The fourth study²³ examined the relationship between the 4-item Morisky measure and MEMS caps among 83 patients commencing tricyclic antidepressants.

Table 15-4 Likelihood Ratios of Self-reported Adherence Measures Compared to the Medication Event Monitoring System

Measure	Sensitivity, %	Specificity, %	LR+ (95% CI) ^a	LR- (95% CI) ^a
Medication Adherence Self-Report Inventory ²⁰	66	100	33 (4-317)	0.34 (0.23-0.47)
Compliance Questionnaire Rheumatology ²¹	62	95	17 (4.9-63)	0.39 (0.23-0.58)
Simplified Medication Adherence Questionnaire ²²	72	95	7.9 (2.4-29)	0.31 (0.14-0.58)
Morisky measure ²³	72	74	2.7 (1.6-4.4)	0.36 (0.18-0.64)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

^aThe LR+ is the likelihood ratio (LR) for medication nonadherence. For example, a patient with at least 1 positive answer on the Morisky measure would have a positive result, suggesting that the likelihood of an adherence problem increases by 2.7. LR- is the LR for medication adherence. For example, a patient with all negative answers would have an LR of 0.36, suggesting that the patient is less likely to have a problem with medication adherence.

There are inherent self-reported biases that are likely to exist,²⁴ such as halo effects (eg, overreporting adherence) or recall bias. Self-reported adherence represents “an upper limit” of the estimate of actual adherence because of social desirability. Despite the biases in using self-report measures of medication adherence, studies tend to show that patients are accurate when they say that they have not taken their medication.²⁵ Simply put, when patients state they are having problems taking their medication as prescribed, they are telling the truth. Patients’ claims of medication adherence tend to underestimate their true rate by approximately 20%.²⁴ Reasons for overreporting adherence may include the following: individuals might wish to give a socially desirable answer even though it deceives their physician, they might not understand their regimen and therefore not realize that they are not adhering, or they might forget instances of nonadherence.

Clinicians who rely on self-reports of adherence need to take steps to improve the accuracy of their assessment. Suggested steps include giving clear directions on how to take medications, providing education and encouragement regarding the need for both adherence and accurate reporting of adherence so that patients will not give socially desirable answers, asking nonjudgmental and nonthreatening questions about current medication use, and probing barriers to accurate reporting.²⁶ These steps should be taken routinely with all patients because there are few factors to help identify the patients at greatest risk of inaccurate reporting.²⁷

Because appointment nonadherence can be easily checked, it should serve as a warning to screen for medication nonadherence. It is useful to ask patients what they already know and believe about their medications, including how many pills they take, as well as the names and purpose of taking them. Inquiring about the most common adverse events, as well as when they are likely to occur, may prompt the patient to have a more open discussion about medication (and appointment) adherence. It is useful to ask patients what they already know and believe about the medications before and after explaining these points.

EVIDENCE FROM GUIDELINES

No guidelines give a standard approach to assessing or measuring adherence. Many guidelines for individual disorders address the need for assessing adherence.

CLINICAL SCENARIO—RESOLUTION

Despite her assertion that she is taking her medications just as instructed, you ask her whether she is having any problems taking her medications. You find that she is confused about when she should be taking her medications. After answering her question, you ask her to repeat the information. In addition, you ask whether you may explain the regimen to her husband. Finally, you provide a written reminder describing when each medication should be taken.

REFERENCES FOR THE UPDATE

1. Svarstad BL, Chewning BA, Sleath BL, Claesson C. The Brief Medication Questionnaire: a tool for screening patient adherence and barriers to adherence. *Patient Educ Couns*. 1999;37(2):113-124.
2. Butler JA, Peveler RC, Roderick P, Horne R, Mason JC. Measuring compliance with drug regimens after renal transplantation: comparison of self-report and clinician rating with electronic monitoring. *Transplantation*. 2004;77(5):786-789.
3. Benner JS, Glynn RJ, Mogun H, Neumann PJ, Weinstein MC, Avorn J. Long-term persistence in use of statin therapy in elderly patients. *JAMA*. 2002;288(4):455-461.
4. McDermott MM, Schmitt B, Wallner E. Impact of medication nonadherence on coronary heart disease outcomes: a critical review. *Arch Intern Med*. 1997;157(17):1921-1929.
5. DiMatteo MR. Variations in patients’ adherence to medical recommendations: a quantitative review of 50 years of research. *Med Care*. 2004;42(3):200-209.
6. Feldman R, Bacher M, Campbell N, Drover A, Chockalingam A. Adherence to pharmacologic management of hypertension. *Can J Public Health*. 1998;89(5):I16-I18.
7. Flack J, Novikov SV, Ferrario CM. Benefits of adherence to antihypertensive drug therapy. *Eur Soc Cardiol*. 1996;17(suppl A):16-20.
8. Mallion JM, Baguet JP, Siche JP, Tremel F, de Gaudemaris R. Compliance, electronic monitoring and antihypertensive drugs. *J Hypertens Suppl*. 1998;16(1):S75-S79.
9. Haynes RB, McKibbon KA, Kanani R. Systematic review of randomised trials of interventions to assist patients to follow prescriptions for medications. *Lancet*. 1996;348(9024):383-386.
10. Stephenson BJ, Rowe BH, Haynes RB, Macharia WM, Leon G. Is this patient taking the treatment as prescribed? *JAMA*. 1993;269(21):2779-2781.
11. Dunbar-Jacob J, Dwyer K, Dunning EJ. Compliance with antihypertensive regimen: a review of the research in the 1980s. *Ann Behav Med*. 1991;13(1):31-39.
12. Haynes RB, McDonald HP, Garg AX. Helping patients follow prescribed treatment: clinical applications. *JAMA*. 2002;288(22):2880-2883.

ASSESSING MEDICATION ADHERENCE—MAKE THE DIAGNOSIS

PRIOR PROBABILITY

Approximately 50% of patients do not take their medications as prescribed.

POPULATION FOR WHOM MEDICATION NONADHERENCE SHOULD BE CONSIDERED

- All patients should be assessed
- Patients not responding as expected to medication
- Patients receiving multiple or complicated regimens
- Patients who miss appointments
- Older patients
- Adolescents
- Patients with cognitive disorders
- Patients with psychiatric disorders
- Patients treated for asymptomatic diseases (eg, hypercholesterolemia, hypertension)

Given the high prevalence of medication nonadherence, ask all patients, “Have you missed any pills in the past week?” Any patient who answers yes should be considered nonadherent (**Table 15-5**). When patients answer no, a negative response to each of the Morisky questions makes it even more likely that the patient is adherent. Questionnaires about adherence may work better than clinical judgment.

Table 15-5 Detecting the Likelihood of Medication Nonadherence

	LR+ (95% CI) ^a	LR- (95% CI) ^b
Single question: Have you missed any pills in the past week?	4.3 (3.1-6.1)	0.51 (0.44-0.58)
Morisky questions (any one positive)	2.7 (1.6-4.4)	0.36 (0.18-0.64)
1. Do you ever forget to take your medication?		
2. Are you careless at times about taking your medicine?		
3. When you feel better, do you sometimes stop taking your medicine?		
4. Sometimes when you feel worse, do you stop taking your medicine?		

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

^aThe LR+ is the likelihood ratio for medication nonadherence.

^bThe LR- is the likelihood ratio for medication adherence.

REFERENCE STANDARD TESTS

There is no single best reference standard for measuring adherence for all medications, nor is there general agreement for the level of adherence that is considered optimal. Physicians must use their best judgment, tailored to their knowledge of each patient.

13. Haynes R, Sackett DL, Gibson ES, et al. Improvement of medication compliance in uncontrolled hypertension. *Lancet*. 1976;1(7972):1265-1268.
14. Psaty B, Koepsel TD, Wagner EH, LoGerfo JP, Inui TS. The relative risk of incident coronary heart disease associated with recently stopping the use of beta-blockers. *JAMA*. 1990;263(12):1653-1657.
15. Mallion JM, Dutrey-Dupagne C, Vaur L, et al. Benefits of electronic pill-boxes in evaluating treatment compliance of patients with mild to moderate hypertension. *J Hypertens*. 1996;14(1):137-144.
16. Kimmerling M, Wagner G, Ghosh-Dastidar B. Factors associated with accurate self-reported adherence to HIV antiretrovirals. *Int J STD AIDS*. 2003;14(4):281-284.
17. Waltherhouse DM, Calzone KA, Mele C, Brenner DE. Adherence to oral tamoxifen: a comparison of patient self report, pill counts and micro-electronic monitoring. *J Clin Oncol*. 1993;11(6):2547-2548.
18. Juncos L. Patient compliance and angiotensin converting enzyme inhibitors in hypertension. *J Cardiovasc Pharmacol*. 1990;15(suppl 3):S22-S25.
19. Hays R, DiMatteo MR. Key issues and suggestions for patient compliance assessment: source of information, focus of measures, and nature of response options. *J Compliance Health Care*. 1987;2:37-53.
20. Walsh JC, Mandalia S, Gazzard BG. Responses to a 1 month self-report on adherence to antiretroviral therapy are consistent with electronic data and virological treatment outcome. *AIDS*. 2002;16(2):269-277.^a
21. de Klerk E, van der Heijde D, Landewe R, van der Tempel H, van der Linden S. The compliance-questionnaire-rheumatology compared with electronic medication event monitoring: a validation study. *J Rheumatol*. 2003;30(11):2469-2475.^a
22. Knobel H, Alonso J, Casado JL, et al; for the GEEMA Study Group. validation of a simplified medication adherence questionnaire in a large cohort of HIV-infected patients: the GEEMA Study. *AIDS*. 2002;16(4):605-613.^a
23. George CF, Peveler RC, Heiliger S, Thompson C. Compliance with tricyclic antidepressants: the value of four different methods of assessment. *Br J Clin Pharmacol*. 2000;50(2):166-171.^a
24. Haynes R, Taylor DW, Sackett DL. Can simple clinical measurement detect patient noncompliance? *Hypertension*. 1980;2(6):759-782.
25. Ley P. *Communicating With Patients*. London, England: Croom Helm; 1988.
26. Roter DL, Hall JA. Physician's interviewing styles and medical information obtained from patients. *J Gen Intern Med*. 1987;2(5):325-329.
27. Bosworth HB. Medication treatment adherence. In: Bosworth HB, Oddone EZ, Weinberger M, eds. *Patient Treatment Adherence: Concepts, Interventions, and Measurement*. Mahwah, NJ: Lawrence Erlbaum Associates; 2005:147-194.

^aFor the Evidence to Support the Update for this topic, see <http://www.JAMAevidence.com>.

EVIDENCE TO SUPPORT THE UPDATE:

15

Compliance and Medication Adherence

TITLE The Compliance Questionnaire Rheumatology Compared With Electronic Medication Event Monitoring: A Validation Study.

AUTHORS de Klerk E, van der Heijde D, Landewe R, van der Tempel H, van der Linden S.

CITATION *J Rheumatol.* 2003;30(11):2469-2475.

QUESTION Is the Compliance Questionnaire Rheumatology (CQR) a valid measure of adherence compared with a gold standard electronic Medication Event Monitoring System (MEMS)?

DESIGN Prospective study comparing questionnaire responses with a MEMS.

SETTING Three outpatient referral centers for rheumatology.

PATIENTS Eighty-five patients who completed the questionnaire and electronic monitoring data were available.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

Prescriptions had to be first prescriptions (which is not the same as a prescription for a new diagnosis), and the instructions had to be “take as directed” (not “on demand”). In a case in which 2 drugs were started at the same time, the drug monitored was chosen to be the drug group in which the fewest patients were enrolled.

The development of the CQR has been described in detail.¹ CQR consists of 19 items, which were derived from a series of patient interviews and a focus group interview, and reflects statements that were made by patients regarding their drug-taking behavior. Patients were asked to indicate how much they agree with each statement on a 4-point Likert scale, with anchors “don’t agree at all” (scored 1), “don’t agree” (scored 2), “agree” (scored 3), and “agree very much” (scored 4). The CQR total score is calculated by summing the items, subtracting 19, and dividing by 0.57. This ensures that the CQR total score can vary from 0 (complete nonadherence) to 100 (perfect adherence).

The reference standard was MEMS.

MAIN OUTCOME MEASURES

Sensitivity and specificity.

MAIN RESULTS

Twenty-nine (34%) patients were not completely adherent (see Table 15-6).

CONCLUSIONS

LEVEL OF EVIDENCE Level 2.

STRENGTHS The CQR is a patient-oriented questionnaire that was designed to explore concepts related to patient adherence in antirheumatic drug regimens. The measure is easy to read and understand. Patients can complete the questionnaire in their own environment; an interviewer is not required. It has good psychometric properties.¹

LIMITATIONS The mean time to complete the questionnaire was 12 minutes. Approximately 20% of the sample had at least 1 missing value. Some of the questions are not applicable to all responders.

Table 15-6 Likelihood Ratios for the Compliance Questionnaire Rheumatology

Test	Sensitivity, %	Specificity, %	LR+ (95% CI) ^a	LR- (95% CI) ^b
Compliance Questionnaire Rheumatology (<80% on the scale)	62	96	17 (4.9-63)	0.39 (0.23-0.58)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

^aLR+ is the likelihood ratio for medication nonadherence.

^bLR- is the likelihood ratio for medication adherence.

REFERENCE FOR THE EVIDENCE

1. de Klerk E, van der Heijde D, van der Tempel H, van der Linden S. Development of a questionnaire to investigate patient compliance with anti-rheumatic drug therapy. *J Rheumatol.* 1999;26(12):2635-2641.

Reviewed by Hayden B. Bosworth, PhD

TITLE Compliance With Tricyclic Antidepressants: The Value of 4 Different Methods of Assessment.

AUTHORS George CF, Peveler RC, Heiliger S, Thompson C.

CITATION *Br J Clin Pharmacol.* 2000;50(2):166-171.

QUESTION What are the advantages and disadvantages of the 4 methods for studying adherence with antidepressants?

DESIGN As part of a larger randomized controlled trial, subjects were followed for up to 12 weeks after beginning to take antidepressants, but adherence was assessed at 6 weeks.

SETTING General practices.

PATIENTS Eighty-three patients aged 18 years or older who were beginning antidepressant treatment

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

At a 6-week interview, subjects were asked the 4 standard questions described by Morisky et al¹:

1. Do you ever forget to take your medication?
2. Are you careless at times about taking your medicine?
3. When you feel better, do you sometimes stop taking your medicine?
4. Sometimes when you feel worse, do you stop taking your medicine?

A yes answer was scored as 1, and the sum of yes answers constitutes the score. A score of 0 suggests no problems with medicine taking, whereas the maximum of 4 could indicate major difficulties and suggests poor adherence.

For each subject, antidepressant medication was dispensed in medication event monitoring system (MEMS) containers, sufficient for a period of 3 weeks. The MEMS cap contained a microprocessor that records the time the bottle is opened as a proxy for appropriate dosing. This was treated as the diagnostic standard, although the investigators also did pill counts.

MAIN OUTCOME MEASURES

Sensitivity and specificity compared with the MEMS measure at a threshold of 80%.

MAIN RESULTS

Among the subjects, 27 (32%) were nonadherent (see Table 15-7).

Pill counts indicated better adherence at the 80% pill count level (only 17 were nonadherent). However, the pill counts also found 29 additional patients who finished with more pills than were dispensed.

Table 15-7 Likelihood Ratios of the Morisky Scale for Medication Adherence

Test	Sensitivity, %	Specificity, %	LR+ (95% CI) ^a	LR- (95% CI) ^b
Morisky scale (nonadherence ≥ 1)	74	72	2.7 (1.6-4.4)	0.36 (0.18-0.64)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

^aLR+ is the likelihood ratio for medication nonadherence.

^bLR- is the likelihood ratio for medication adherence.

CONCLUSIONS

LEVEL OF EVIDENCE Level 2.

STRENGTHS The Morisky scale is 1 of the more common self-reported measures of medication adherence. It has been used for multiple diseases and is easy and quick to administer.

LIMITATIONS Depressed patients have a particular problem with adherence, especially as their symptoms improve.

REFERENCE FOR THE EVIDENCE

1. Morisky E, Green LW, Levine DM. Concurrent and predictive validity of a self-reported measure of medication adherence. *Med Care.* 1986;24(1): 67-74.

Reviewed by Hayden B. Bosworth, PhD

TITLE Validation of a Simplified Medication Adherence Questionnaire in a Large Cohort of HIV-Infected Patients: The GEEMA Study.

AUTHORS Knobel H, Alonso J, Casado JL, et al; for the GEEMA Study Group.

CITATION *AIDS.* 2002;16(4):605-613.

QUESTION How effective is the Simplified Medication Adherence Questionnaire (SMAQ) in identifying nonadherent patients?

DESIGN Prospective observational study of adherence.

SETTING A publicly funded specialist human immunodeficiency virus (HIV) clinic where all treatment was free.

PATIENTS A total of 40 HIV-seropositive adults who had commenced nelfinavir treatment in combination with other antiretroviral drugs. The study is a subset of a larger multicenter (69 hospitals) and nationwide (Spain) study.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

A group of physicians, nurses, pharmacists, psychologists, and patients, all with experience with antiretroviral treatment and adherence, developed the SMAQ. The questionnaire was based on the Morisky scale.¹ The research group then made the following change: item 3 “When you feel better, do you sometimes stop taking your medicine?”) was eliminated because many HIV-infected patients are asymptomatic. Three additional questions were incorporated, with the aim of obtaining more adherence-specific measurements. A modified version of a question used by Samet et al² to determine the number of missed doses during the previous 24 hours was used. The SMAQ result was considered “positive” when a positive response to any of the questions was provided.

The criterion validity assessment was carried out in a subset of 40 patients. The patients were provided with a MEMS cap bottle for each pack of nelfinavir prescribed.

MAIN OUTCOME MEASURES

Sensitivity and specificity of the SMAQ.

MAIN RESULTS

Among the patients, 18 (45%) were not adherent (see Table 15-8).

Table 15-8 Likelihood Ratios for the Simplified Medication Adherence Questionnaire

Test	Sensitivity, %	Specificity, %	LR+ (95% CI) ^a	LR- (95% CI) ^b
SMAQ	72	91	7.9 (2.4-29)	0.31 (0.14-0.58)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio; SMAQ, Simplified Medication Adherence Questionnaire.

^aLR+ is the likelihood ratio for medication nonadherence.

^bLR- is the likelihood ratio for medication adherence.

CONCLUSIONS

LEVEL OF EVIDENCE Level 2.

STRENGTHS The SMAQ showed a positive association to virologic outcome. The SMAQ's internal consistency and reproducibility were satisfactory and the measure is easy to implement.

LIMITATIONS Like all self-report measures, the questionnaire is limited by recall and social desirability bias.

REFERENCES FOR THE EVIDENCE

- Morisky E, Green LW, Levine DM. Concurrent and predictive validity of a self-reported measure of medication adherence. *Med Care*. 1986;24(1):67-74.
- Samet JH, Libman H, Steger KA, et al. Compliance with zidovudine therapy in patients infected with human immunodeficiency virus, type 1: a cross-sectional study in a municipal hospital clinic. *Am J Med*. 1992;92(5):495-502.

Reviewed by Hayden B. Bosworth, PhD

TITLE Responses to a 1-Month Self-report on Adherence to Antiretroviral Therapy Are Consistent With Electronic Data and Virologic Treatment Outcome.

AUTHORS Walsh JC, Mandalia S, Gazzard BG.

CITATION AIDS. 2002;16(2):269-277.

QUESTION Is the Medication Adherence Self-Report Inventory (MASRI) a valid measure of antiretroviral therapy compared with an objective measure of adherence?

DESIGN Prospective study comparing questionnaires responses with medication event monitoring system (MEMS) (MEMS TrackCap), pill count, and plasma human immunodeficiency virus (HIV) viremia.

SETTING A publicly funded specialist HIV clinic where all treatment was free.

PATIENTS Seventy-eight HIV-seropositive adults receiving stable combination antiretroviral therapy dispensed from the clinic's pharmacy.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

The MASRI consists of 12 items with 2 broad themes. The first section is related to the amount of medication actually taken. The second part of the MASRI addressed the timing of doses. Both 3-day and 2-week self-report assessments were used.

For each subject, the antiretroviral drug in the combination that presented the greatest barrier to adherence was selected (eg, higher pill burdens, dietary requirements, more frequent dose intervals). Subjects were given a bottle containing this drug, closed with a MEMS cap. These are pill bottle caps containing a microprocessor that records the time the bottle is opened as a presumptive dose. This was treated as the diagnostic standard.

MAIN OUTCOME MEASURES

Sensitivity and specificity.

MAIN RESULTS

See Table 15-9.

Table 15-9 Likelihood Ratios for the Medication Adherence Self-Report Inventory

Test	Sensitivity, ^a %	Specificity, ^a %	LR+ (95% CI) ^b	LR- (95%CI) ^c
MASRI (2 wk before ≤ 80% level of adherence)	66	100	33 (4-317)	0.34 (0.23-0.47)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio; MASRI, Medication Adherence Self-Report Inventory.

^aResults transformed from data in manuscript so that the LR+ is associated with an abnormal MASRI and increases the probability of nonadherence. The LR- would be a normal result on the MASRI and decreases the likelihood of nonadherence (ie, the patient is adherent).

^bLR+ is the likelihood ratio for medication nonadherence.

^cLR- is the likelihood ratio for medication adherence.

CONCLUSIONS

LEVEL OF EVIDENCE Level 2.

STRENGTHS The MASRI is one of the first adherence questionnaires for antiretroviral therapy to have been validated against an objective measure.

LIMITATIONS Study sample selected had higher adherence than that typically observed in the literature; subjects who admitted to deviating from instructions were excluded from analysis.

Reviewed by Hayden B. Bosworth, PhD

CHAPTER 16

Can the Clinical Examination Diagnose Left-Sided Heart Failure in Adults?

Robert G. Badgett, MD

Cynthia D. Mulrow, MD, MSc

Catherine R. Lucey, MD

CLINICAL SCENARIOS

CASE 1 Your first patient is a 65-year-old man with Canadian class II angina and hypertension. He takes daily aspirin, sublingual nitroglycerin, and a calcium-channel blocker. His examination findings are normal, but his electrocardiogram (ECG) shows inferolateral Q waves, and the chest radiograph shows cardiomegaly.

CASE 2 Your second case patient is an obese, 70-year-old woman who has had dyspnea on exertion and fatigue for 3 months. She reports no orthopnea or paroxysmal nocturnal dyspnea. Her medical history reveals 40 pack-years of smoking, poorly controlled chronic hypertension, and type 2 diabetes mellitus. She has a blood pressure of 180/100 mm Hg, a sustained apical impulse, bilateral rales, and moderate pretibial edema. Her complete blood cell count and basic chemistry results are normal. Her ECG shows left ventricular hypertrophy with strain. The chest radiograph reveals normal heart size and enlarged upper lobe vessels.

CASE 3 Your last case patient is a 58-year-old woman with idiopathic dilated cardiomyopathy. Cardiac catheterization showed normal coronary artery and a left ventricular ejection fraction (EF) of 35%. She has done well for a year but now complains of dyspnea on exertion despite treatment with diuretics, digoxin, and an angiotensin-converting enzyme (ACE) inhibitor. You find a displaced apical impulse, soft apical third heart sound, and clear lung fields. Her ECG and chest radiograph results are unchanged from before and show nonspecific ST changes and cardiomegaly, respectively.

WHY IS THE DIAGNOSIS IMPORTANT?

Clinicians seeing patients similar to case patient 1 must recognize that a reduced left ventricular EF can exist even when there is no fluid overload. The first patient, even if asymptomatic, should be treated with an ACE inhibitor if a previous infarction significantly reduced the EF.¹ A reduced EF also may suggest a need for coronary angiography to evaluate for possible revascularization.²

Case 2 presents a number of different diagnostic and therapeutic possibilities. The decision to pursue diagnostic testing for pulmonary, cardiac, or other causes of dyspnea rests with the clinician's ability to identify and interpret clinical findings. Knowledge of the accuracy of cardiac and pulmonary findings is essential. If an increased left ventricular filling pressure is detected, identifying the underlying pathophysiology is critical. Systolic and diastolic dysfunction have different causes that require different diagnostic considerations and treatment.^{3,4} Previous articles address pulmonary findings⁵ and

distinction of cardiac and pulmonary causes of dyspnea.⁶ We focus on cardiac findings.

Case 3 with known cardiomyopathy introduces another diagnostic dilemma. When is the left-sided heart filling pressure adequately decreased? The primary goals of treatment are improved survival and functional status. Clinical findings are not used to titrate therapy aimed at improved survival (ACE inhibitors); however, clinical examination is used to decide whether a patient needs more diuresis or afterload reduction to improve functional status.^{7,8} If clinical examination is inaccurate, the potential for undertreatment or over-treatment of patients with congestive symptoms exists.

PATOPHYSIOLOGY AND DEFINITIONS

The physiologic definition of heart failure seems precise: “the pathophysiological state in which an abnormality of cardiac function is responsible for failure of the heart to pump blood at a rate commensurate with the requirements of the metabolizing tissues, or to do so only from an elevated filling pressure.”⁹ In clinical practice, this definition includes a heterogeneous population of patients with varying underlying pathophysiologies for which there is no criterion standard (gold standard) test.

An alternative, clinically meaningful way to define left-sided heart failure is a decreased left ventricular EF or increased filling pressure. Patients with left-sided heart failure then fall into one of 3 groups: decreased EF with normal filling pressure, decreased EF with increased filling pressure, or normal EF with increased filling pressure. EF is easily measured, accurately identifies persons with systolic dysfunction, and has well-described treatment and prognostic implications. Filling pressure has diagnostic and therapeutic implications. As the failing heart adapts by increasing left ventricular filling pressure to augment cardiac output, increased filling pressure must indicate myocardial dysfunction. Thus, when the filling pressure is increased but the EF is normal, the patient has diastolic dysfunction.³ These relations hold if the clinician has excluded other causes of increased filling pressure such as intermittent ischemia, valvular and pericardial disease, and high output states.^{10,11} In addition, an increased filling pressure correlates with increased symptoms and edema,¹² even among patients with severe systolic dysfunction,¹³⁻¹⁵ that are reduced by diuretic or vasodilator therapy.^{7,8,16}

METHODS

Literature Search

We searched English-language medical literature regarding the clinical examination in heart failure, with 3 goals in mind: (1) to identify the most discriminating and useful clinical findings; (2) to estimate the utility of the overall clinical examination; and (3) to describe characteristics of patients or clinical settings when disease can be ruled out or confirmed. All studies we reviewed examined the ability of

clinical findings or the overall clinical examination to predict filling pressure or EF. Acceptable criterion standards for filling pressure were left ventricular end-diastolic pressure, left atrial pressure, pulmonary capillary wedge pressure, or pulmonary artery diastolic pressure. Finally, we sought studies that compared multiple clinical findings with a multivariate analysis.

To develop a structured search strategy, we used pertinent articles already in our files and 2 related critical reviews that had used extensive search methods.^{5,17} We then searched MEDLINE (English language) from January 1986 to November 1995 with the developed structured search strategy that required certain words in the title or abstract (strategy available on request). This search yielded 1254 articles, of which 28 met inclusion criteria (Table 16-1). We excluded 3 additional studies because the independent significance of cardiac findings was not assessed with a multivariate analysis.⁵¹⁻⁵³ Because only 2 articles addressed the distinction of systolic and diastolic dysfunction,^{40,41} we also included 9 studies of distinguishing systolic and diastolic dysfunction that met all inclusion criteria other than having a multivariate analysis.⁴²⁻⁵⁰ Excepting the lack of multivariate analysis, these studies have quality levels similar to the studies we reviewed of diagnosing a reduced EF or increased filling pressure.

Data Abstraction

Two of us (R.G.B. and C.R.L. or C.D.M.) independently reviewed all studies. We calculated sensitivities and specificities and tests of significance for studies that did not provide those results. If necessary, data were reconstructed from scattergrams and graphs. For studies of systolic function, we made calculations for an EF of less than 40% when possible. The quality level of evidence provided by each article was adapted from previous work.⁵ Levels 1 and 2 had independent comparison of clinical examination items with a suitable criterion standard among consecutive or random patients. Level 1 studies were larger and had at least 96 patients with and without a normal criterion standard (this number assures confidence intervals < +10%). Level 3 studies had independent comparison of findings to a criterion standard among patients who were not consecutively or randomly chosen. Level 4 studies did not have independent (or the use of blinding not stated) comparison of findings to a criterion standard.

To determine the utility of the clinical examination, studies were pooled with a random-effects model⁵⁴ for sensitivities, specificities, and likelihood ratios (LRs). When possible, we stratified the predicted probabilities of disease into 3 levels: low, intermediate, and high risk. For studies that compare a clinically predicted EF with a measured EF,^{29-32,37,38} we stratified risk of disease according to the predicted EF: low risk, predicted EF of 60% or greater; intermediate risk, predicted EF of 31% to 59%; and high risk, predicted EF of 30% or lower. For the study by McNamara et al,³⁷ low-probability patients had no abnormal findings, intermediate patients had 1 to 2 abnormal findings, and high-probability patients had 3 or 4 abnormal findings. We then pooled the studies to calculate multilevel LRs⁵⁵ and compare the true prevalences of disease in each risk stratum. We

Table 16-1 Summary of Studies Reviewed

Source, y	Population	Gold Standard
Studies of Increased Filling Pressure		
Harlan et al, ¹² 1977	1306 Patients (in validation group) with known coronary disease	LVEDP > 15 mm Hg
Carlson et al, ¹⁸ 1985	96 Patients who received elective right-sided heart catheterization	PCWP ≥ 12 mm Hg
Forrester et al, ^{19,20} 1977, 1976	188 Consecutive patients peri-infarction	PCWP > 15 mm Hg
Fein et al, ²¹ 1984	70 Consecutive ICU patients with pulmonary edema	PCWP > 18 mm Hg
Tuchschmidt et al, ²² 1987	35 ICU patients needing right-sided heart catheterization	PCWP ≥ 15 mm Hg
Eisenberg et al, ²³ 1984	97 ICU patients, without recent infarction, needing right-sided heart catheterization	PCWP > 15 mm Hg
Connors et al, ²⁴ 1983	62 ICU patients, without recent infarction, needing right-sided heart catheterization	PCWP > 12 mm Hg
Connors et al, ²⁵ 1990	502 ICU patients needing right-sided heart catheterization	PCWP > 18 mm Hg
Steingrub et al, ²⁶ 1991; Celoria et al, ²⁷ 1990 ^a	154 ICU patients needing right-sided heart catheterization	PCWP ≥ 18 mm Hg
Butman et al, ¹⁴ 1993	52 Patients with mean EF of <20% undergoing pretransplant evaluation	PCWP ≥ 18 mm Hg
Chakko et al, ¹³ 1991	52 Patients with mean EF of 19% undergoing pretransplantation evaluation	PCWP > 15 mm Hg
Stevenson and Perloff, ¹⁵ 1989	50 Patients with mean EF of 18% undergoing pretransplantation evaluation	PCWP ≥ 22 mm Hg
Studies of Systolic Dysfunction		
Rihal et al, ²⁸ 1995	14 507 Patients enrolled in the Coronary Artery Surgery Study	EF < 50%
Eagle et al, ²⁹ 1988	222 Patients in 2 groups electively referred for MUGA	EF < 50%
Mattleman et al, ³⁰ 1983	199 Elective referrals for MUGA in patients with coronary disease	EF < 50%, EF < 30%
Ostojic et al, ³¹ 1989	238 Patients in 2 groups who received cardiac catheterization	EF < 50%
Cease and Nicklas, ³² 1986	105 Patients in 2 groups referred for MUGA for various reasons	EF < 50%
Gadsbøll et al, ^{33,34} 1989	98 Patients who received MUGA 7-15 d after infarction	EF ≤ 52%
Jain et al, ³⁵ 1993	32 Patients who received echocardiogram 15-25 d after infarction	EF < 40%
Mangschau et al, ³⁶ 1986	477 Patients who received MUGA 8-12 d after infarction	EF < 50%
McNamara et al, ³⁷ 1988	760 Patients who received MUGA 6-24 d after infarction	EF ≤ 40%
Sanford et al, ³⁸ 1982	100 Patients who received MUGA after infarction	EF < 50%
Silver et al, ³⁹ 1994	304 Patients in 2 groups who received MUGA, echocardiogram, or catheterization 2-21 d after infarction	EF < 50%
Studies of Diastolic Dysfunction		
Ghali et al, ⁴⁰ 1991	82 Consecutive patients admitted for CHF	FS > 24%
McDermott et al, ⁴¹ 1995	298 Consecutive patients admitted for syndrome of CHF	EF ≥ 50%
Aguirre et al, ⁴² 1989	151 Patients with 2 signs of CHF who were referred for echocardiogram	EF ≥ 55%
Aronow et al, ⁴³ 1990	247 Elderly residents of a long-term care facility with clinical criteria of CHF	EF ≥ 50%
Bier et al, ⁴⁴ 1988	87 Consecutive inpatients with pulmonary edema	Normal wall motion by echocardiogram
Cocchi et al, ⁴⁵ 1991	118 Consecutive elderly patients on a geriatrics service with clinical criteria of CHF	EF ≥ 50%
Cohn et al, ⁴⁶ 1990	623 Male veterans who met criteria to be in the V-HeFT Study	EF ≥ 45%
Dougherty et al, ⁴⁷ 1984	72 Consecutive patients with clinical CHF referred for gated radionuclide ventriculography	EF ≤ 45%
Echeverria et al, ⁴⁸ 1983	50 Consecutive referrals for echocardiograms because of CHF	EF ≥ 50%
Takarada et al, ⁴⁹ 1992	172 Consecutive elderly patients admitted for CHF	FS ≥ 30%
Wong et al, ⁵⁰ 1989	54 Elderly patients admitted for CHF who were referred for echocardiogram	Normal wall motion by echocardiogram

Abbreviations: CHF, congestive heart failure; EF, ejection fraction; FS, fractional shortening; ICU, intensive care unit; LVEDP, left ventricular end diastolic pressure; MUGA, multi-gated angiography; PCWP, pulmonary capillary wedge pressure; V-HeFT, Vasodilator Heart Failure Trial.

^aThere is partial overlap among the patients in the studies by Steingrub et al²⁶ and Celoria et al.²⁷

excluded 1 study, although the published data allowed stratifying the predicted probabilities.³³ This study had outlying results and was the only study not to incorporate the chest radiograph into the clinical assessment.

To determine the best clinical findings, we tabulated how often a particular finding was studied and how often it had

univariate or multivariate significance. These tables are available on request. “Very helpful” findings have been studied at least twice and have either univariate or multivariate significance every time studied. “Somewhat helpful” findings for increased filling pressure or diastolic dysfunction are significant at least half the times they were studied. As many findings

are associated with systolic dysfunction, somewhat helpful findings are restricted to those significant more than half the times studied. Findings that are “helpful only when present” are those that are not usually statistically significant but are usually reported as having a specificity of at least 90%. We believe these findings are clinically significant when present.

Our last goal was to describe when disease could be ruled out or confirmed by the clinical examination. We used studies that successfully describe either low-probability or high-probability patients (positive or negative predictive value $\geq 90\%$). With these studies, we used their decision aids, prediction rules, or multivariate equations to estimate the number of abnormal clinical findings that would place a patient in each level of risk (Figures 16-1 and 16-2).

RESULTS

How to Detect Increased Left Ventricular Filling Pressure

Although clinicians routinely assess filling pressure in patients similar to those in cases 2 and 3, there is little literature on our ability to do so. Four studies¹²⁻¹⁵ assess whether

multiple clinical findings identify patients with invasively determined increased left ventricular filling pressure. Three of these studies¹³⁻¹⁵ involve patients with known severe systolic dysfunction (mean EF $< 20\%$) who are referred for pre-transplant evaluation (Table 16-1). These studies are biased by a high prevalence of increased filling pressure.

Unfortunately, isolated clinical findings alone have a limited role in diagnosis. Very helpful findings are radiographic redistribution and jugular venous distention (Table 16-2). These findings, when used alone, only help when they are abnormal and so can confirm the presence of increased filling pressure in patients with known severe systolic dysfunction. Among patients referred for consideration of cardiac transplant with a high (73%) prevalence of increased filling pressure,¹³⁻¹⁵ radiographic redistribution indicates an 80%¹³ to 90%¹⁴ probability and jugular venous distention, an 85%¹³ to 100%¹⁵ probability of increased filling pressure. The absence of either finding cannot rule out increased filling pressure. In patients with lesser probabilities of increased filling pressure, such as those without known severe systolic dysfunction, isolated findings may not be useful. Somewhat helpful findings include dyspnea and abnormal vital signs

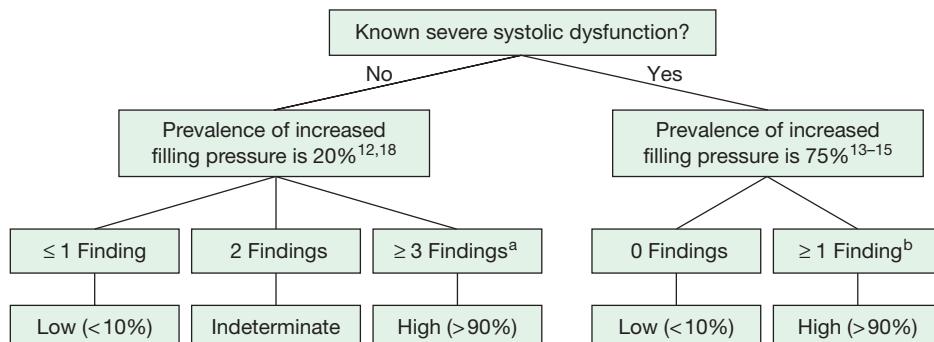


Figure 16-1 Suggested Algorithm to Determine the Risk of Increased Left Ventricular Filling Pressure Using Findings From Appropriate Column of Table 16-2

^aRecommendation extrapolated from Carlson et al,¹⁸ who found that the presence of 2 findings did not reliably indicate an increased filling pressure.

^bAt least 1 abnormal very helpful finding indicates an 80%-100% probability of increased filling pressure; the role of less helpful findings has not been quantified.

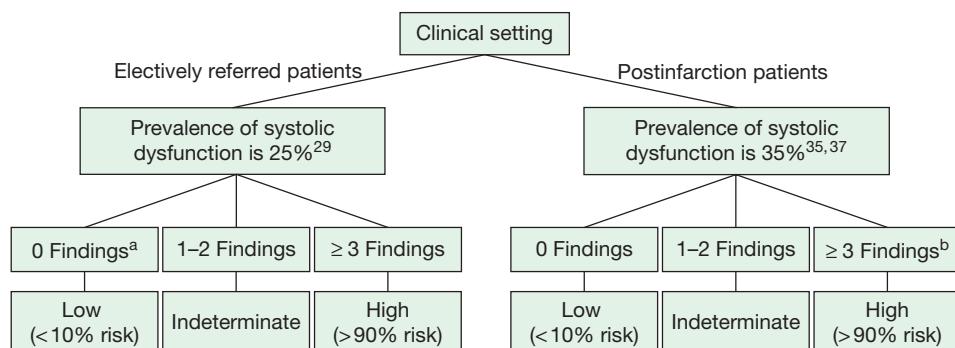


Figure 16-2 Suggested Algorithm to Determine Risk of Decreased Left Ventricular Ejection Fraction Using Findings From Appropriate Column of Table 16-2

^aNo abnormal findings in patients with a history of hypertension indicates low risk. In patients without hypertension, a decreased ejection fraction cannot be ruled out; however, other studies^{35,37} suggest that a history of hypertension does not affect the diagnosis. Unfortunately, these studies use regression equations that cannot be translated into number of abnormal findings.

^bMcNamara et al³⁷ report that at least 3 abnormal findings indicate only 83% probability of an ejection fraction less than 40%; however, the role of cardiomegaly was not studied.

Table 16-2 Helpful Clinical Findings for the Detection of Heart Failure

	Increased Filling Pressure	Ejection Fraction < 40%	Diastolic Dysfunction
Very helpful findings ^a	Radiographic redistribution, jugular venous distention	Radiographic cardiomegaly, ^b or redistribution, anterior Q waves, left bundle-branch block, abnormal apical impulse	Current hypertension
Somewhat helpful findings ^c	Dyspnea, ^d orthopnea, tachycardia, ^e low SBP, ^f PPP < 25%, S ₃ , rales, abnormal abdomino-jugular reflux, radiographic cardiomegaly ^f	Pulse > 90/min ^{29,37} or > 100/min, ^{35,36} SBP < 90 ²⁹ mmHg, PPP < 33%, ³⁵ S ₃ , rales, dyspnea, any previous infarction, CPK > 200 ³⁵ or > 1000 ³⁷ IU	Obesity, ^d no tachycardia, ^d elderly, ^d no smoking, no coronary disease
Findings helpful only when present ^g	Edema	Jugular venous distention, edema	Normal radiographic heart size

Abbreviations: CPK, creatinine phosphokinase in the postinfarction patient; PPP, proportional pulse pressure (pulse pressure/systolic pressure); SBP, systolic blood pressure.

^aVery helpful findings are significant in all studies and have been studied at least twice. Bolded findings are always independently significant.

^bCurrent hypertension is systolic pressure higher than 160 mm Hg⁴⁴ or diastolic pressure higher than 100 mm Hg⁴⁴ or higher than 105 mm Hg.⁴⁰

^cSomewhat helpful findings are significant in at least half of studies. For systolic dysfunction, only findings significant in more than half of studies are listed.

^dStudied only once.

^eNo cutoff to define abnormal.

^fCardiomegaly is somewhat helpful in initially detecting increased filling pressure; however, cardiomegaly may remain after reduction of the filling pressure.

^gFindings helpful only when present are consistently reported as highly specific, although they are not usually statistically significant.

(Table 16-2). Radiographic cardiomegaly is somewhat helpful¹²⁻¹⁴ but loses its specificity after the initial detection of increased filling pressure because it can be a permanent finding and not fluctuate with changes in filling pressure. Dependent edema is helpful only when present. Edema is highly specific for increased filling pressure, although it has poor sensitivity. Limiting our review to studies meeting our inclusion criteria, history of infarction¹³ was not helpful. Clinicians may use other findings not listed in Table 16-2; however, their independent role in the assessment of increased filling pressure is unknown.

If isolated findings alone are not helpful, can multiple findings in combination or the overall clinical examination rule out or confirm increased filling pressure? A total of 11 studies address this question (Table 16-3).^{12,18-27} Their pooled operating characteristics do not yield predictive values that reliably confirm or rule out an increased filling pressure in typical patients, such as those in the emergency department or hospital because of dyspnea.⁵⁶⁻⁵⁸

If the overall examination cannot successfully dichotomize patients into those with either normal or increased filling pressure, can the examination place patients into 3 groups, those whose filling pressure is increased, indeterminate, or normal?^{59,60} If so then clinicians could pursue alternative diagnoses in patients highly likely to have a normal filling pressure while initiating treatment in patients highly likely to have increased filling pressure. Although no study has formally evaluated the utility of this approach, we suggest that the probability of increased filling pressure is related to the number of findings that are detected on clinical examination.

The number of findings associated with low, intermediate, or high probability of increased filling pressure depends on the clinical setting (Figure 16-1). For example, patients without known severe systolic dysfunction have a low prevalence (22%) of increased filling pressure.^{12,18} Patients likely to have normal filling pressure in this setting are reported by Carlson et al¹⁸ to have no more than 1 finding (negative likelihood ratio

[LR-], 0.1). Extrapolating from the results of Carlson et al,¹⁸ in which only 73% of patients with at least 2 findings had increased filling pressure, patients highly likely to have increased filling pressure will have at least 3 abnormal findings. In patients with known severe systolic dysfunction, the prevalence of increased filling pressure is higher (73%)¹³⁻¹⁵ and easier to confirm but harder to rule out. Patients likely to have a normal filling pressure in this setting will have no abnormal findings.¹⁴ Increased filling pressure in patients with known severe systolic dysfunction is likely if there is a single very helpful finding such as redistribution or jugular venous distention.

In summary, in populations without known severe systolic dysfunction, patients with no more than 1 abnormal finding likely have a normal filling pressure, whereas those with at least 3 abnormal findings may have an increased filling pressure. Among populations with known severe systolic dysfunction, patients with no abnormal findings likely have a normal filling pressure, whereas those with 1 very helpful finding likely have increased filling pressure. Patients in either setting with an intermediate number of findings will have an indeterminate filling pressure. These conclusions are based on a limited number of small studies. Future research needs to confirm and refine these conclusions.

How to Detect Decreased EF

Although detection of decreased EF is better described than detection of increased filling pressure,²⁸⁻³⁵ isolated findings have an even smaller role in detecting a decreased EF. Five findings are very helpful in identifying patients with an EF of less than 40% (Table 16-2). Radiographic cardiomegaly consistently adds independent information in predicting decreased EF. However, its sensitivity and specificity (51% and 79%, respectively) are insufficient to help the clinician in most clinical settings.¹⁷ Abnormal apical impulse (especially sustained duration), radiographic redistribution, and anterior Q waves or left bundle-branch block on ECG are also consistent predictors, although they do not consistently add

Table 16-3 Performance of the Clinical Evaluation for Detecting an Increased Filling Pressure or a Decreased Ejection Fraction

Source	Level of Evidence	Prevalence of Disease, %	Sensitivity, %	Specificity, %	LR+	LR-
Increased Filling Pressure						
Patients referred for elective evaluation						
Harlan et al ¹²	1	21	52	85	3.5	0.6
Carlson et al ¹⁸	3	31	90	85	5.9	0.1
Postinfarction patients						
Forrester et al ¹⁹	2	64	85	85	5.7	0.2
ICU patients						
Fein et al ²¹	2	49	91	47	2.0	0.2
Eisenberg et al ²³	3	67	57	35	1.0	1.2
Tuchschmidt and Sharma ²²	3	55	53	94	8.8	0.5
Connors et al ²⁴	3	61	51	67	2.0	0.7
Connors et al ²⁵	3	45	50	78	2.0	0.6
Steingrub et al ²⁶	3	31	35	85	3.0	0.7
Summary			54	69	1.7	0.7
Patients referred for pretransplant evaluation						
Stevenson and Perloff ¹⁵	2	86	58	100	... ^a	0.4
Summary			62	76	2.6	0.5
Decreased Ejection Fraction						
Patients referred for elective evaluation						
Rihal et al ²⁸	1	23	68	74	2.6	0.4
Eagle et al ^{29b}	2	44	96	48	1.8	0.1
Eagle et al ^{29c}	2	45	90	58	2.1	0.2
Mattleman et al ^{30b}	2	38	97	66	2.9	0.1
Mattleman et al ^{30c}	2	42	88	74	3.4	0.2
Ostojic et al ^{31d}	2	45	92	68	5.7	0.1
Ostojic et al ^{31e}	2	23	83	46	2.0	0.4
Cease and Nicklas ^{32d}	4	42	96	65	2.6	0.2
Cease and Nicklas ^{32e}	4	39	100	56	2.5	0.1
Summary			91	62	2.4	0.1
Postinfarction patients						
Gadsbøll et al ³³	1	43	72	76	1.2	0.4
Mangschau et al ³⁶	1	24	67	83	8.5	0.7
McNamara et al ³⁷	1	35	67	77	3.8	0.4
Sanford et al ³⁸	3	25	100	42	1.7	0.3
Silver et al ^{39d}	3	31	98	57	1.7	0.1
Silver et al ^{39e}	3	25	97	65	1.5	0.1
Jain et al ³⁵	4	59	95	85	6.2	0.1
Summary			80	70	2.7	0.3
Summary			85	66	2.5	0.2

Abbreviations: ICU, intensive care unit; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

^aEllipses indicate date not available.

^bResults of a cardiologist or other physician.

^cResults of a prediction rule.

^dResults from a training set.

^eResults from a validation set.

independent information. Anterior Q waves and left bundle-branch block both have a specificity of almost 90% or higher.^{28,39} Other Q waves increase the sensitivity of the ECG but are less specific.^{20,40} A single study on predominantly is-

chemic patients reported that the presence of any electrocardiographic abnormality has a sensitivity of 90%.²⁸ This sensitivity may decline in other populations. Many other findings are somewhat helpful (Table 16-2). Two findings as-

sociated with increased filling pressure, edema and elevated jugular venous pressure, are helpful only when present. In 2 studies of postinfarction patients, these 2 findings are highly specific for decreased EF.^{33,37} However, this specificity will decrease in populations with less ischemic heart disease and a higher prevalence of increased filling pressure with normal EFs (diastolic dysfunction). Other findings examined that were not significant in a majority of studies are age,^{28,35,37,39} orthopnea,³⁸ left ventricular hypertrophy on ECG,²⁹ and a history of hypertension^{29,35,37,39} or congestive heart failure.^{29,39}

How should the clinician use this information? Unfortunately, the pooled sensitivity and specificity of the overall clinical examination do not yield high enough predictive values to reliably assess the EF (Table 16-3). However, the clinical examination can categorize patients into low, indeterminate, and high probability of systolic dysfunction. Eight studies describe patients who have either a very low ($\leq 10\%$) or a very high ($\geq 90\%$) probability of systolic dysfunction (Figure 16-2).^{29-32,35,37-39} Low-probability patients have none of the abnormal clinical findings associated with a decreased EF; high-probability patients have at least 3 and usually more findings, and indeterminate patients have an intermediate number of abnormal findings (1 to 2). (Although specific findings used in each study vary, we recommend that the clinician use those for ejection fraction $< 40\%$ in Table 16-2.) The probabilities of an EF of less than 40% in the low-probability, indeterminate, and high-probability categories according to clinical findings are 7% (range, 0%-10%), 34% (range, 23%-41%), and 89% (range, 86%-100%), respectively.^{33-36,39,40}

Typical outpatient populations probably have lower prevalences of decreased EFs than patients who have been included in the above-cited studies. Thus, outpatients categorized as low risk are likely to have less than a 7% probability of a low EF. Recent guidelines also suggest that diagnosing a noncardiac explanation for a patient's symptoms can help decrease the probability of systolic dysfunction.⁶¹

How to Distinguish Diastolic From Systolic Dysfunction

Studies addressing the distinction of diastolic from systolic dysfunction do so by predicting EF in patients (usually inpatients) with clinical evidence of increased filling pressure (patients with increased filling pressure and normal EF are assumed to have diastolic dysfunction). Only 2 studies^{40,41} use a multivariate analysis to report the independent information from each clinical finding (Table 16-1). Thus, we also reviewed studies that report the performance of multiple clinical findings but do not use a multivariate analysis to compare independent values of findings.⁴²⁻⁵⁰

The only very helpful finding is currently elevated blood pressure (Table 16-2). Its sensitivity ranges from 61%⁴⁰ to 66%⁴⁴ and its specificity is 59%⁴⁴ to 70%.⁴⁰ Thus, its value as an isolated finding for identifying the EF among patients with increased filling pressure is questionable. Somewhat helpful findings are obesity, the absence of tachycardia, older age (no cutoff age is available), and absence of smoking or coronary disease. A normal heart size on the chest radiograph is helpful only when present. A normal heart size is highly specific for diastolic dysfunction as the underlying cause of increased filling pressure. However, because 56% to

75% of patients with diastolic dysfunction have left ventricular hypertrophy^{42,44,47} that can cause radiographic cardiomegaly, a normal heart size is not a common (sensitive) finding among patients with diastolic dysfunction. Neither electrocardiographic evidence of left ventricular hypertrophy^{40,47} nor a history of hypertension^{40,42-45,48-50} discriminates diastolic from systolic dysfunction. In addition, the patient's sex and the presence of a third or fourth heart sound are not helpful.

Little information exists about whether the clinical examination or multiple findings in combination can distinguish diastolic from systolic dysfunction in patients with increased filling pressure. One study suggests multiple findings in combination have 76% accuracy.⁴¹ In practice, this accuracy may be higher because this study did not analyze the role of the current blood pressure. Until more research is available, we recommend all patients with evidence of increased filling pressure have objective assessment of their EF. Some patients with signs suggesting an increased filling pressure with a normal EF will have causes other than diastolic dysfunction.^{44,48} These causes include valvulopathy, right ventricular dysfunction from emphysema, iatrogenic volume overload, pulmonary fibrosis, and intermittent left ventricular ischemia. The clinician should consider these diagnoses before diagnosing diastolic dysfunction.

Precision of Clinical Findings

Much variability exists in reports of the precision of clinical findings. This reflects the subtle nature of findings and the varied abilities of clinicians. Most studies,^{31,62,63} but not all,⁶⁴ suggest this variability is partly attributable to subspecialty training or examiner experience.

Two studies report precision of multiple clinical findings in an overall bedside evaluation aimed at predicting EFs. (We report precision using standard qualitative descriptors of the κ statistic for interobserver agreement.⁶⁵) For the overall bedside estimate of EF, Gadsbøll et al³³ report "fair" precision ($\kappa = 0.28-0.37$). In comparing 3 examiners, Gadsbøll et al³³ find that the cardiologist tends to more accurately predict EF than the 2 resident physicians. When assessing specific clinical findings, precision is as follows: jugular venous distention, fair to substantial ($\kappa = 0.31-0.69$)^{34,14}; displaced apical impulse, moderate to substantial ($\kappa = 0.53-0.73$)³⁴; third heart sound, slight to moderate ($\kappa = 0.14-0.60$)^{34,14}; rales, slight to substantial ($\kappa = 0.12-0.65$)^{34,14}; and edema, fair to substantial ($\kappa = 0.27-0.64$).³⁴ For radiographic findings interpreted by radiologists, precision is as follows: cardiomegaly, moderate ($\kappa = 0.48$)¹⁴; redistribution, fair to moderate ($\kappa = 0.38-0.50$)^{14,17}; and interstitial edema, moderate to almost perfect ($\kappa = 0.56-0.83$).^{14,17} These results suggest that more experienced clinicians are more precise and, presumably, more accurate examiners.

THE ELICITATION OF SELECTED SIGNS OF HEART FAILURE

Vital Signs

Details on how to measure blood pressure are reviewed in another article.⁶⁶ A pulse rate faster than 90^{29,37} or 100/min^{35,38}

may indicate reduced EF. A systolic pressure lower than 90 mm Hg is associated with a reduced EF,²⁹ whereas a diastolic pressure higher than 105 mm Hg⁴⁰ or an overall blood pressure of 160/100 mm Hg⁴⁴ or higher may indicate diastolic dysfunction. Tachycardia and low systolic pressure are also associated with increased filling pressure¹²; however, no cut-offs are available. A proportional pulse pressure (the difference between systolic and diastolic pressures divided by the systolic pressure) less than 33% is associated with a decreased EF³⁵ and less than 25% is associated with a reduced cardiac index.¹⁵

Several studies suggest the bedside assessment of the blood pressure response to the Valsalva maneuver may be one of the best predictors of both systolic dysfunction and increased filling pressure.^{51,52} However, more research is needed because the independent contribution of this maneuver to the cardiac examination has not been assessed.

Jugular Veins

The jugular venous pulsation best correlates with right-sided heart filling pressure. Because the right atrial pressure can be elevated from left-sided heart disease, jugular venous distension also correlates with left-sided heart filling pressure¹²⁻¹⁵ and left ventricular EF.³³ The method of assessing the jugular venous pressure and the abdominojugular reflux is detailed in another article.⁶⁷

Apical Impulse

Abnormalities of the location, size, or duration of the apical impulse best correlate with increased left ventricular mass.^{53,68} Although both the location^{20,33} and the duration²⁹ of the apical impulse are significantly associated with a reduced EF, only a sustained impulse independently adds to predicting the EF.²⁹ The normal apical impulse is located in the fourth or fifth intercostal spaces and is a brief tap. It is palpable in less than half of supine patients.^{69,70} The 45-degree left lateral decubitus position increases the yield,^{53,69,70} as may palpating during expiration.⁷¹ Simultaneous auscultation allows quantification of the duration of the impulse, with a sustained impulse defined as lasting more than two-thirds of systole.⁶⁸

An abnormal heart size, as measured by precordial percussion, may be more sensitive than an abnormal apical impulse in detecting abnormal ventricular size.⁵³ However, the independent contribution of this finding to the overall clinical examination, especially compared with radiographic cardiomegaly, has not been assessed.

Third Heart Sound

The left ventricular third heart sound occurs during ventricular vibration with rapid diastolic filling. This vibration may occur when either filling pressure is increased or ventricular compliance is reduced. The third heart sound is low pitched and may be faint or intermittent. It should be sought with the bell of the stethoscope over the apical impulse. Listening while the patient is in the 45-degree left lateral decubitus position doubles the yield.⁷²

The third heart sound may be confused with other diastolic sounds such as an opening snap, an abnormally split second heart sound, or even a fourth heart sound if the patient is tachycardic. The third heart sound is, with rare exception, the only middiastolic sound. It occurs approximately 150 ms after the second heart sound, or 5 times longer than the normal split of the second heart sound.

Radiographic Cardiomegaly

We previously reviewed the role of the chest radiograph in assessing left ventricular dysfunction.¹⁷ The 2 most helpful findings are cardiac size and pulmonary vessels.

Radiographic cardiomegaly best correlates with total left ventricular size^{73,74} and can be caused by an enlarged left ventricular cavity or a hypertrophied ventricular wall. As decreased EF correlates with an enlarged left ventricle cavity, cardiomegaly is observed in persons with decreased EF. Because an increased filling pressure is due to a decreased EF in 52% to 72%,^{40,44} of patients with heart failure, cardiomegaly is also associated with an increased filling pressure.¹² Finally, because cardiomegaly can be caused by a hypertrophied ventricular wall, it may also be observed with diastolic dysfunction.^{41,46}

Cardiomegaly is most easily defined as an increased cardiothoracic ratio, usually more than 50%. The cardiothoracic ratio is the cardiac width divided by largest width of the thoracic cavity above the diaphragms.⁷⁵ False-positive interpretations of cardiomegaly may occur from an apical fat pad,⁷⁶ a transversely positioned heart,⁷⁷ a decrease in thoracic width,⁷⁷ or radiographs taken anteroposteriorly (supine) or during a poor inspiration.

Radiographic Redistribution

Redistribution, also called cephalization, flow shift, or pulmonary venous hypertension, best correlates with left ventricular filling pressure. Increased filling pressure is usually caused by systolic dysfunction. Thus, redistribution correlates with systolic dysfunction.^{20,29,33,35,37,38}

Less precision exists in the assessment of redistribution than cardiomegaly^{78,79} or signs of pulmonary interstitial edema.^{14,79,80} The easiest and best-studied definition of redistribution is simply upper lobe vessels larger than lower lobe vessels.⁸¹ Comparisons should be made at equal distances above and below the hilum. As when assessing cardiomegaly, supine or expiratory radiographs can cause false-positive interpretations.^{82,83}

How to Improve One's Skills

Several good audiotapes are available to assist in learning cardiac sounds.⁸⁴⁻⁸⁷ Assessment of the third heart sound and duration of the apical impulse can be assisted with visual feedback. A tongue blade may be pressed over the apical impulse with the examiner's fingernail or stethoscopic diaphragm. Alternatively, a cotton applicator can be wedged in the hole of a pediatric precordial suction electrode.⁸⁸ Both methods can visually demonstrate a sustained apical impulse or third heart sound.

THE BOTTOM LINE

1. Cardiac findings can be subtle and hard to elicit. Clinician experience contributes to the ability to detect findings.
2. Detecting an increased filling pressure suggests the need for diuretics. Detecting increased filling pressure when the EF is normal may signal the presence of diastolic dysfunction. For the detection of increased filling pressure:
 - Very helpful findings are radiographic redistribution and jugular venous distention.
 - Somewhat helpful findings are dyspnea, orthopnea, tachycardia, decreased systolic or pulse pressure, third heart sound, rales, and abdominojugular reflux.
 - Edema is helpful only when present.
 - Few studies address how to rule out or diagnose increased filling pressure, but the presence of systolic dysfunction affects the assessment. The clinician can probably exclude the diagnosis in patients without known systolic dysfunction when no more than 1 finding of increased filling pressure is present. In such patients, at least 3 abnormal findings suggest increased filling pressure. In patients with known severe systolic dysfunction, the absence of any abnormal findings of increased filling pressure probably rules out increased filling pressure, and the presence of at least 1 very helpful finding (radiographic redistribution or jugular venous distention) suggests increased filling pressure. It is unclear whether these results apply in the intensive care setting.
3. Detecting a decreased EF indicates the need for specific medical therapy and may influence the decision for coronary revascularization. For the detection of a decreased EF:
 - Very helpful findings are chest radiograph (especially cardiomegaly, but also redistribution), anterior Q waves or left bundle-branch block on ECG, and abnormal apical impulse (especially if sustained).
 - Somewhat helpful findings are tachycardia, decreased blood pressure or pulse pressure, third heart sound, rales, dyspnea, previous infarction other than anterior, and high peak creatine phosphokinase level (in the postinfarction patient).
 - Two signs of increased filling pressure, edema and increased jugular venous pressure, are helpful only when present. This is probably true only when diastolic dysfunction is unlikely as a cause of increased filling pressure.
 - The clinician can usually rule out the detection of a decreased EF when no abnormal findings, including no sign of increased filling pressure, are present (LR-, 0.1). At least 3, and frequently more, abnormal findings are needed to confirm the diagnosis (LR+, 14).
4. Among patients with increased filling pressure, distinguishing diastolic from systolic dysfunction determines further evaluation and treatment. In making the distinction:
 - The very helpful finding is elevated blood pressure during the episode of increased filling pressure.

- Somewhat helpful findings are obesity, lack of tachycardia, older age, and absence of smoking or coronary artery disease.
- Normal radiographic heart size is helpful only when present.
- Few studies address the distinction of systolic from diastolic dysfunction. Currently, the EF needs objective measurement in patients with increased filling pressure. In patients who appear to have increased filling pressure with a normal EF, the clinician should also consider cor pulmonale, valvular cardiac disease, pulmonary fibrosis, intermittent ischemia, and iatrogenic volume overload.

Since this manuscript was accepted for publication, an additional study⁸⁹ has been published that found that the highest combination of sensitivity and specificity in the detection of systolic dysfunction occurred when physical examination, ECG, and chest radiograph were combined.

Author Affiliations at the Time of the Original Publication

Department of Medicine, University of Texas Health Science Center at San Antonio and the Department of Medicine, Audie L. Murphy Memorial Veterans Hospital, San Antonio, (Drs Badgett and Mulrow); and Washington Hospital Center, Washington, DC (Dr Lucey).

Acknowledgments

The authors acknowledge and appreciate the statistical pooling of data in Table 16-3 with Meta-Test software by Joseph Lau, MD, the statistical help by Gil Ramírez, DrPH, and careful reviews of the manuscript by Matthew Gillman, MD, SM, and John Williams, MD, MHS.

This study was supported in part by the San Antonio Cochrane Center, Audie L. Murphy Memorial Veterans Affairs Hospital. Dr Mulrow is a Veterans Affairs Senior Research Associate.

REFERENCES

1. The SOLVD Investigators. Effect of enalapril on mortality and the development of heart failure in asymptomatic patients with reduced left ventricular ejection fractions. *N Engl J Med.* 1992;327(10):685-691.
2. The role of revascularization. In: Konstam M, Dracup K, Baker D, et al, eds. *Heart Failure: Evaluation and Care of Patients With Left-Ventricular Systolic Dysfunction.* Rockville, MD: Agency for Health Care Policy and Research, Public Health Service, US Dept of Health and Human Services; 1994:67-77. Clinical Practice Guideline 11.
3. Bonow RO, Udelson JE. Left ventricular diastolic dysfunction as a cause of congestive heart failure. *Ann Intern Med.* 1992;117(6):502-510.
4. Gaasch WH. Diagnosis and treatment of heart failure based on left ventricular systolic or diastolic dysfunction. *JAMA.* 1994;271(16):1276-1280.
5. Hollerman DR Jr, Simel DL. Does the clinical examination predict airflow limitation? *JAMA.* 1995;273(4):313-319.
6. Mulrow CD, Lucey CR, Farnett LE. Discriminating causes of dyspnea through the clinical examination. *J Gen Intern Med.* 1993;8(7):383-392.
7. Hutcheon D, Nemeth E, Quinlan D. The role of furosemide alone and in combination with digoxin in the relief of symptoms of congestive heart failure. *J Clin Pharmacol.* 1980;20(1):59-68.
8. Guyatt GH. The treatment of heart failure. *Drugs.* 1986;32(6):538-568.
9. Braunwald E. Pathophysiology of heart failure. In: Braunwald E, ed. *Heart Disease.* Philadelphia, PA: WB Saunders Co; 1992:393.
10. Eichna LW. Circulatory congestion and heart failure. *Circulation.* 1960; 22:864-886.

11. Braunwald E, Ross J. The ventricular end-diastolic pressure. *Am J Med*. 1963;34:147-150.
12. Harlan R, Oberman A, Grimm R, Rosati RA. Chronic congestive heart failure in coronary artery disease. *Ann Intern Med*. 1977;86(2):132-138.
13. Chakko S, Woska D, Martinez H, et al. Clinical, radiographic, and hemodynamic correlations in chronic congestive heart failure. *Am J Med*. 1991;90(3):353-359.
14. Butman SM, Ewy GA, Standen JR, Kern KB, Hahn E. Bedside cardiovascular examination in patients with severe chronic heart failure. *J Am Coll Cardiol*. 1993;22(4):968-974.
15. Stevenson LW, Perloff JK. The limited reliability of physical signs for estimating hemodynamics in chronic heart failure. *JAMA*. 1989;261(6):884-888.
16. Pharmacologic management. In: Konstam M, Dracup K, Baker D, et al, eds. *Heart Failure: Evaluation and Care of Patients With Left-Ventricular Systolic Dysfunction*. Rockville, MD: Agency for Health Care Policy and Research, Public Health Service, US Dept of Health and Human Services; 1994:49-66. Clinical Practice Guideline 11.
17. Badgett RG, Mulrow CD, Ramirez G, Otto PM. How well can the chest radiograph diagnose left ventricular dysfunction? *J Gen Intern Med*. 1996;11(10):625-634.
18. Carlson KJ, Lee DC-S, Goroll AH, Leahy M, Johnson RA. An analysis of physicians' reasons for prescribing long-term digitalis therapy in outpatients. *J Chronic Dis*. 1985;38(9):733-739.
19. Forrester JS, Diamond GA, Swan HJC. Correlative classification of clinical and hemodynamic function after myocardial infarction. *Am J Cardiol*. 1977;39(2):137-145.
20. Forrester JS, Diamond G, Chatterjee K, Swan HJC. Medical therapy of acute myocardial infarction by application of hemodynamic subsets. *N Engl J Med*. 1976;295(24):1356-1362.
21. Fein AM, Goldberg SK, Walkenstein MD, Dershaw B, Braitman L, Lippman ML. Is pulmonary artery catheterization necessary for the diagnosis of pulmonary edema? *Am Rev Respir Dis*. 1984;129(6):1006-1009.
22. Tuchschmidt J, Sharma OP. Impact of hemodynamic monitoring in a medical intensive care unit. *Crit Care Med*. 1987;15(9):840-843.
23. Eisenberg PR, Jaffe AS, Schuster DP. Clinical evaluation compared to pulmonary artery catheterization in the hemodynamic assessment of critically ill patients. *Crit Care Med*. 1984;12(7):549-553.
24. Connors AF, McMaffree DR, Gray BA. Evaluation of right-heart catheterization in the critically ill patient without acute myocardial infarction. *N Engl J Med*. 1983;308(5):263-267.
25. Connors AF, Dawson NV, Shaw PK, Montenegro HD, Nara AR, Martin L. Hemodynamic status in critically ill patients with and without acute heart disease. *Chest*. 1990;98(5):1200-1206.
26. Steingrub JS, Celoria G, Vickers-Lahti M, Teres D, Bria W. Therapeutic impact of pulmonary artery catheterization in a medical/surgical ICU. *Chest*. 1991;99(6):1451-1455.
27. Celoria G, Steingrub JS, Vickers-Lahti M, et al. Clinical assessment of hemodynamic values in two surgical intensive care units. *Arch Surg*. 1990;125(8):1036-1039.
28. Rihal CS, Davis KB, Kennedy JW, Gersh BJ. The utility of clinical, electrocardiographic, and roentgenographic variables in the prediction of left ventricular function. *Am J Cardiol*. 1995;75(4):220-223.
29. Eagle KA, Quertermous T, Singer DE, et al. Left ventricular ejection fraction: physician estimates compared with gated blood pool scan measurements. *Arch Intern Med*. 1988;148(4):882-885.
30. Mattleman SJ, Hakki A-H, Iskandrian AS, Segal BL, Kane SA. Reliability of bedside evaluation in determining left ventricular function. *J Am Coll Cardiol*. 1983;1(2 pt 1):417-420.
31. Ostoicic MC, Young JB, Hess KR. Prediction of left ventricular ejection fraction using a unique method of chest x-ray and ECG analysis. *Am Heart J*. 1989;117(3):590-598.
32. Cease KB, Nicklas JM. Prediction of left ventricular ejection fraction using simple quantitative clinical information. *Am J Med*. 1986;81(3):429-436.
33. Gadsbøll N, Høilund-Carlsen PF, Nielsen GG, Bernig J, Bruun NE, Hein E. Interobserver agreement and accuracy of bedside estimation of right and left ventricular ejection fraction in acute myocardial infarction. *Am J Cardiol*. 1989;63(18):1301-1307.
34. Gadsbøll N, Høilund-Carlsen PF, Nielsen GG, et al. Symptoms and signs of heart failure in patients with myocardial infarction. *Eur Heart J*. 1989;10(11):1017-1028.
35. Jain AP, Gupta OP, Kumar A, Trivedi SK. Evaluation of clinical variables as predictors of left ventricular function in acute myocardial infarction. *J Assoc Physicians India*. 1993;41(1):17-19.
36. Mangschau A, Jonsbu J, Karlson RL. Congestive heart failure and ejection fraction in acute myocardial infarction. *Acta Med Scand*. 1986;220(2):101-107.
37. McNamara RF, Carleen E, Moss AJ, and the Multicenter Postinfarction Research Group. Estimating left ventricular ejection fraction after myocardial infarction by various clinical parameters. *Am J Cardiol*. 1988;62(4):192-196.
38. Sanford CF, Corbett J, Nicod P, et al. Value of radionuclide ventriculography in the immediate characterization of patients with acute myocardial infarction. *Am J Cardiol*. 1982;49(4):637-644.
39. Silver MT, Rose GA, Paul SD, O'Donnell CJ, O'Gara PT, Eagle KA. A clinical rule to predict preserved left ventricular ejection fraction in patients after myocardial infarction. *Ann Intern Med*. 1994;121(10):750-756.
40. Ghali JK, Kadakia S, Cooper RS, Liao Y. Bedside diagnosis of preserved versus impaired systolic function in heart failure. *Am J Cardiol*. 1991;67(11):1002-1006.
41. McDermott MM, Feinglass J, Sy J, Gheorghiade M. Hospitalized congestive heart failure patients with preserved versus abnormal left ventricular systolic function. *Am J Med*. 1995;99(6):629-635.
42. Aguirre FV, Pearson AC, Lewen MK, McCluskey M, Labovitz AJ. Usefulness of Doppler echocardiography in the diagnosis of congestive heart failure. *Am J Cardiol*. 1989;63(15):1098-1102.
43. Aronow WS, Ahn C, Kronzon I. Prognosis of congestive heart failure in elderly patients with normal versus abnormal left ventricular systolic function associated with coronary artery disease. *Am J Cardiol*. 1990;63:1526-1528.
44. Bier AJ, Eichacker PQ, Sinoway LI, Terrible SM, Strom JA, Keefe DL. Acute cardiogenic pulmonary edema. *Angiology*. 1988;39(3 pt 1):211-218.
45. Cocchi A, Zuccala G, Del Sindaco D, et al. Cross-sectional echocardiography: a window on congestive heart failure in the elderly. *Aging*. 1991;3(3):257-262.
46. Cohn JN, Johnson G, and the Veterans Administration Cooperative Study Group. Heart failure with normal ejection fraction: the V-HeFT Study. *Circulation*. 1990;81(2 suppl):48-53.
47. Dougherty AH, Naccarelli GV, Gray EL, Hicks CH, Goldstein RA. Congestive heart failure with normal systolic function. *Am J Cardiol*. 1984;54(7):778-782.
48. Echeverria HH, Bilsker MS, Myerburg RJ, Kessler KM. Congestive heart failure: echocardiographic insights. *Am J Med*. 1983;75(5):750-755.
49. Takarada A, Kurogane H, Minamiji K, et al. Congestive heart failure in the elderly: echocardiographic insights. *Jpn Circ J*. 1992;56(6):527-534.
50. Wong WF, Gold S, Fukuyama O, Blanchette PL. Diastolic dysfunction in elderly patients with congestive heart failure. *Am J Cardiol*. 1989;63(20):1526-1528.
51. Zema MJ, Restivo B, Sos T, Sniderman KW, Kline S. Left ventricular dysfunction: bedside Valsalva manoeuvre. *Br Heart J*. 1980;44(5):560-569.
52. Zema MJ, Masters AP, Margouleff D. Dyspnea: the heart of the lungs? *Chest*. 1984;85(1):59-64.
53. Heckerling PS, Weiner SL, Wolfkiel CJ, et al. Accuracy and reproducibility of precordial percussion and palpation for detecting increased left ventricular end-diastolic volume and mass. *JAMA*. 1993;270(16):1943-1948.
54. Raudenbush SW. Random effects models. In: Cooper H, Hedges LV, eds. *The Handbook of Research Data Synthesis*. New York, NY: Russell Sage Foundation; 1994:301-316.
55. Sackett DL, Haynes RB, Guyatt GH, Tugwell P. *Clinical Epidemiology*. 2nd ed. Boston, MA: Little Brown & Co; 1991:120.
56. Mustchin CP, Tiwari I. Diagnosing the breathless patient. *Lancet*. 1982;1(8277):907-908.
57. Schmitt BP, Kushner MS, Weiner SL. The diagnostic usefulness of the history of the patient with dyspnea. *J Gen Intern Med*. 1986;1(6):386-393.
58. Fedullo AL, Swinborne AJ, McGuire-Dunn C. Complaints of breathlessness in the emergency department. *N Y State J Med*. 1986;86(1):4-6.
59. Pauker SG, Kassirer JP. The threshold approach to clinical decision making. *N Engl J Med*. 1980;302(20):1109-1117.
60. Simel DL, Feussner JR, DeLong ER, Matchar DB. Intermediate, indeterminate, and uninterpretable diagnostic test results. *Med Decis Making*. 1987;7(2):107-114.

61. Konstam M, Dracup K, Baker D, et al. *Heart Failure: Evaluation and Care of Patients With Left-Ventricular Systolic Dysfunction*. Rockville, MD: Agency for Health Care Policy and Research, Public Health Service, US Dept of Health and Human Services; 1994:2. Clinical Practice Guideline 11.
62. Westman EC, Matchar DB, Samsa GP, Mulrow CD, Waugh RA, Feussner JR. Accuracy and reliability of apical S3 gallop detection. *J Gen Intern Med*. 1995;10(8):455-457.
63. Mangione S, Nieman LZ, Gracely E, Kaye D. The teaching and practice of cardiac auscultation during internal medicine and cardiology training. *Ann Intern Med*. 1993;119(1):47-54.
64. St Clair EW, Oddone EZ, Waugh RA, Corey R, Feussner JR. Assessing housestaff diagnostic skills using a cardiology patient simulator. *Ann Intern Med*. 1992;117(9):751-756.
65. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33(1):159-174.
66. Reeves RA. Does this patient have hypertension? *JAMA*. 1995;273(15):1211-1218.
67. Cook DJ, Simel DL. Does this patient have abnormal central venous pressure? *JAMA*. 1996;275(8):630-634.
68. Conn RD, Cole JS. The cardiac apex impulse. *Ann Intern Med*. 1971;75(2):185-191.
69. Eilen SD, Crawford MH, O'Rourke RA. Accuracy of precordial palpation for detecting increased left ventricular volume. *Ann Intern Med*. 1983;99(5):628-630.
70. Badgett RG, Tanaka DJ, Hunt DK, et al. Can moderate chronic obstructive pulmonary disease be diagnosed by historical and physical findings alone? *Am J Med*. 1993;94(2):188-196.
71. Bates B, Bickley LS, Hoekelman RA. *A Guide to Physical Examination and History Taking*. 6th ed. Philadelphia, PA: JB Lippincott; 1995:286.
72. Bethel HJN, Nixon PGF. Examination of the heart in supine and left lateral positions. *Br Heart J*. 1973;35:902-907.
73. Glover L, Baxley WA. A quantitative evaluation of heart size measurements from chest roentgenograms. *Circulation*. 1973;47(6):1289-1296.
74. Chikos PM, Figley MM, Fisher L. Correlation between chest films and angiographic assessment of left ventricular size. *AJR Am J Roentgenol*. 1977;128(3):367-373.
75. Danzer CS. The cardiothoracic ratio. *Am J Med Sci*. 1919;157:513-521.
76. Feild BJ, Russell RO, Moraski RE, et al. Left ventricular size and function and heart size in the year following myocardial infarction. *Circulation*. 1974;50(2):331-339.
77. Comeau WJ, White PD. A critical analysis of standard methods of estimating heart size from roentgen measurements. *Am J Roentgenol Rad Ther*. 1942;47:665-677.
78. Milne ENC, Pistolesi ENC, Miniati M, Giuntini C. The radiologic distinction of cardiogenic and noncardiogenic edema. *AJR Am J Roentgenol*. 1985;144(5):879-894.
79. Miniati M, Pistolesi M, Paoletti P, et al. Objective radiographic criteria to differentiate cardiac, renal, and injury lung edema. *Invest Radiol*. 1988;23(6):433-440.
80. Norgaard TJ, Gjorup T, Brems-Dalgaard E, Hartelius H, Grun B. Inter-observer variation in the detection of pulmonary venous hypertension in chest radiographs. *Eur J Radiol*. 1990;11(3):203-206.
81. Simon M. The pulmonary vessels in incipient left ventricular decompensation. *Circulation*. 1961;24:185-190.
82. Burko H, Carwell G, Newman E. Size, location, and gravitational changes of normal upper lobe pulmonary veins. *AJR Am J Roentgenol*. 1971;111(4):687-689.
83. Steiner RM, Levin DC. Radiology of the heart. In: Braunwald E, ed. *Heart Disease*. Philadelphia, PA: WB Saunders Co; 1992:213.
84. Perloff JK, Silverman ME. *Physical Examination of the Heart and Circulation* [cassette tape]. Bethesda, MD: American College of Cardiology Extended Learning; 1984.
85. Tavel ME. *Heart Sounds and Murmurs* [cassette tape]. Chicago, IL: Year Book Medical Publishers Inc; 1973.
86. Tilkian AG, Conover MB. *Understanding Heart Sounds and Murmurs* [cassette tape/book]. 2nd ed. Philadelphia, PA: WB Saunders Co; 1984.
87. Harvey WP, Canfield DC. *Clinical Auscultation of the Cardiovascular System* [cassette tape]. Newton, NJ: Laennec Publishing Co; 1989.
88. Willis PW. Inspection and palpation of the precordium. In: Hurst JW, Schlant RC, Rackley CE, Sommenblick EH, Wenger NK, eds. *The Heart, Arteries, and Veins*. New York, NY: McGraw-Hill Book Co; 1990:165.
89. Gillespie ND, McNeil G, Pringle T, Ogston S, Struthers AD, Gillespie SD. Cross sectional study of contribution of clinical assessment and simple cardiac investigations in patients admitted with acute dyspnea. *BMJ*. 1997;314(7085):936-940.

This page intentionally left blank

CHAPTER 16

Does This Dyspneic Patient in the Emergency Department Have Congestive Heart Failure?

Charlie S. Wang, MD

J. Mark FitzGerald, MB, DM

Michael Schulzer, MD, PhD

Edwin Mak, BASc

Najib T. Ayas, MD, MPH

CLINICAL SCENARIOS

CASE 1 A 70-year-old woman with a history of a myocardial infarction and heart failure presents to the emergency department (ED) with a 2-day history of dyspnea at rest, orthopnea, and paroxysmal nocturnal dyspnea. Physical examination reveals an elevated jugular venous pressure, a third heart sound (ventricular filling gallop), bibasilar rales and wheezing, and bilateral lower extremity edema. The chest radiograph reveals cardiomegaly. An electrocardiogram (ECG) shows atrial fibrillation.

CASE 2 A 65-year-old previously healthy man with a 30 pack-year smoking history presents to the ED with a 3-week history of dyspnea on exertion and at rest, associated with productive cough and sputum. Physical examination reveals bilateral rales and wheezing. The chest radiograph reveals pulmonary venous congestion and a pattern of interstitial edema. An ECG shows lateral ST-segment depression.

CASE 3 A 60-year-old man with a history of chronic obstructive pulmonary disease (COPD) and myocardial infarction presents to the ED with a 2-week history of worsening dyspnea on exertion and cough. Physical examination reveals an elevated jugular venous pressure, bilateral wheezing, and bilateral lower extremity edema. The chest radiograph shows normal results. An ECG shows Q waves inferiorly.

WHY IS THIS QUESTION IMPORTANT?

Heart failure is a major public health concern. A heart failure epidemic affects more than 15 million people in North America and Europe, and an additional 1.5 million new cases are diagnosed every year.¹⁻⁵ It is the most costly cardiovascular disorder in western countries, accounting for an estimated total direct annual expenditure of more than \$24 billion in the United States in 2001.^{6,7} Failure to diagnose heart failure increases mortality, delays hospital discharge, and increases treatment costs.^{8,9}

Dyspnea, an uncomfortable sensation of breathing¹⁰ or an awareness of respiratory distress,¹¹ is the cause for more than 2.5 million clinician visits per year in the United States.¹² A number of disorders cause dyspnea, including congestive heart failure, COPD, asthma, deconditioning, metabolic acidosis, anxiety, upper airway obstruction, and neuromuscular weakness. Identifying patients with heart failure among the other causes allows early institution of appropriate symptomatic and evidence-based therapies.

It is not always possible (or feasible) to promptly evaluate every patient with dyspnea with tests of cardiac function (echocardiography, nuclear scans, or cardiac catheterization).

This challenges physicians who must identify heart failure according to medical history, physical examination, and rapidly available investigations (eg, chest radiograph, ECG, serum brain natriuretic peptide [BNP]). Therefore, the purpose of this review was to identify the most useful symptoms, signs, and tests in diagnosing the clinical syndrome of heart failure in dyspneic patients presenting to the ED. By the syndrome of heart failure, we mean an overall clinical diagnosis of heart failure as the cause of dyspnea (irrespective of etiology or systolic or diastolic dysfunction), using information from many sources, including medical history, physical examination, chest radiograph, ECG, serum chemistries, and 1 or more confirmatory tests of cardiac function.

Pathophysiology of Dyspnea in Heart Failure

Multiple pathophysiologic mechanisms have been hypothesized to modulate the sensation of dyspnea in patients with symptomatic heart failure (**Table 16-4**).

A previous Rational Clinical Examination article assessed the usefulness of the clinical examination in predicting decreased left ventricular ejection fraction (EF) or increased filling pressure.¹⁷ Our current review extends the previous report by focusing on the prediction of the clinical syndrome of heart failure in dyspneic patients. This clinical focus is useful because not every patient with left ventricular dysfunction or high filling pressures on objective cardiac testing will be subjectively dyspneic; furthermore, patients with a reduced EF may be dyspneic from causes other than heart failure.¹⁸⁻²⁰ Therefore, the use of the syndrome of heart failure

takes into account a patient's subjective sensation and findings on routine investigations, in addition to objective cardiac testing. One previous literature review has reported on the use of the clinical examination for discriminating causes of dyspnea; however, it was not restricted specifically to the syndrome of heart failure, and summary measures of sensitivity, specificity, and likelihood ratios (LRs) were not reported.²¹

We included serum BNP testing in this review because recent evidence suggests that it is useful in diagnosing heart failure.²² BNP is a neurohormone that is secreted almost exclusively from the ventricles in response to pressure and volume overload that produces natriuresis, diuresis, and smooth muscle relaxation.²³ There is also emerging evidence that BNP is useful in prognosticating cardiovascular mortality in both acute and chronic heart failure.²² Studies are currently ongoing regarding the use of serial BNP levels as an indicator of treatment response and for titrating therapy.²²

How to Elicit Symptoms and Signs

Appropriate history taking and physical examination of the cardiopulmonary system have been described in detail in previous Rational Clinical Examination articles,^{17,24-30} with the exception of the Valsalva maneuver. The Valsalva maneuver is performed by inflating and locking a blood pressure cuff to 15 mm Hg above the resting supine systolic pressure (Korotkoff sounds should not be audible), at which point the patient performs a sustained Valsalva (exhalation against a closed glottis) for at least 10 seconds. In a normal response, systolic blood pressure immediately increases 30 to 40 mm Hg above baseline for 1 to 3 seconds (phase 1, appearance of Korotkoff sounds). As venous return decreases, systolic blood pressure decreases sharply below baseline (phase 2, disappearance of Korotkoff sounds). When the Valsalva is released, there is a further decrease of systolic blood pressure below baseline (phase 3, continued absence of Korotkoff sounds). Between 3 and 15 seconds after release, systolic blood pressure increases 15 mm Hg or more above the baseline level (phase 4, reappearance of Korotkoff sounds).^{21,31-34} Two abnormal responses have been described in heart failure. In the absent overshoot response, phases 1 to 3 are normal, but Korotkoff sounds do not reappear in phase 4. In the square wave response, phase 1 is normal, but Korotkoff sounds are present in phases 2 and 3, followed by disappearance in phase 4.^{21,31,32,34}

Table 16-4 Physiological Categories and Mechanisms Causing Dyspnea in Heart Failure^a

Category	Mechanisms ^{b,10,13-16}
Increased respiratory drive	Increased left LVEDP → pulmonary venous congestion → stimulation of pulmonary J receptors (transmitted by vagal afferents to brain) Pulmonary venous congestion → ventilation/perfusion mismatch, shunt ^c → hypoxemia → stimulation of central and peripheral chemoreceptors
Increased work of breathing	Pulmonary venous congestion → reduced lung compliance → increased airways resistance → increased elastic and resistive work of breathing → mismatch between afferent information from upper airway, lower airway, chest wall mechanoreceptors, and efferent signals to respiratory muscles
Weakness of respiratory pump muscles	Activation of catabolic factors → myopathy (structural, biochemical, functional abnormalities of skeletal respiratory muscles) → reduced respiratory muscle efficiency and endurance → mismatch between afferent mechanoreceptors and efferent signals to respiratory muscles
Psychological	Anxiety, depression → altered central perception

Abbreviation: LVEDP, left ventricular end diastolic pressure.

^aAdapted from Murray,¹³ Braunwald,^{14,16} American Thoracic Society Dyspnea consensus statement,¹⁵ and Manning and Schwartzstein.¹⁰

^bArrows denote one mechanism leading to another.

^cOccurs when blood moves through the lung without coming into contact with oxygenated air.

METHODS

Search Strategy

We conducted a computerized search of MEDLINE from January 1966 to July 2005 concerning the precision and diagnostic accuracy of components of the clinical examination and simple investigations in diagnosing patients with dyspnea. Our strategy was deliberately broad to minimize the possibility of overlooking relevant articles. Multiple searches were performed with the first search using a similar strategy developed for The Rational

Clinical Examination series.³⁵ This strategy combined 4 exploded Medical Subject Headings (physical examination, medical history taking, professional competence, routine diagnostic tests) with 8 keyword categories ("physical exam," "medical history taking," "professional competence," "sensitivity and specificity," "reproducibility of results," "observer variation," "decision support techniques," "Bayes theorem") and 1 textword category ("sensitivity" and "specificity") and intersected with 1 exploded Medical Subject Heading ("dyspnea"). The search was limited to studies published in English about humans. Further MEDLINE searches were conducted combining the following Medical Subject Headings textword and keyword searches: "brain natriuretic peptide," "natriuretic peptide," "BNP," "Valsalva," "hepatojugular," "abdominojugular," and "breathlessness." These were intersected with the exploded medical subject heading "dyspnea" and the textword "dyspnoea."

The computerized search was supplemented with a manual search of reference lists of retrieved studies, review articles, and standard physical examination textbooks to identify additional articles not captured through the computerized search strategy.

Study Selection

One author (C.S.W.) screened the titles and abstracts of the computerized search to identify all potentially relevant articles. All retrieved articles were independently reviewed by 2 authors (C.S.W. and N.T.A.) for eligibility, assessment of methodologic quality, and data abstraction. Only studies that evaluated the diagnostic accuracy of some element of the medical history, physical examination, or readily available diagnostic tests in adult patients with undifferentiated dyspnea presenting to the ED, regardless of whether the patients had known cardiac or pulmonary diseases, were included. Data had to be presented so that 2×2 contingency tables could be extracted. Because there currently is no widely accepted criterion standard for diagnosing heart failure, and because the focus of this review was a syndrome of heart failure, we accepted as a reasonable reference standard a diagnosis agreed on by a panel of physicians after evaluating for appropriate symptoms and signs of heart failure and an appropriate measure of cardiac dysfunction.⁵

We included studies that evaluated common and rapidly available tests (chest radiograph, ECG, and serum BNP) because clinicians rely on these basic investigations in conjunction with their medical history and physical examination in bedside decision making.^{22,36} There are currently multiple BNP assays approved by the Food and Drug Administration for clinical use. To date, the largest published randomized clinical trials have been funded by industry and have reported using the BNP assay of a single manufacturer.

An *a priori* decision was made to exclude studies that investigated other cardiac neurohormones such as A-type natriuretic peptide or other forms of BNP (eg, N-terminal prohormone BNP). It was thought at the time of this review that there would be insufficient published data on these other neurohormones to draw significant conclusions. We also excluded studies that (1) were review articles with no

original data; (2) had no clinical examination performed or reported; (3) used only echocardiography, computed tomography scans, or invasive hemodynamic monitoring as the reference standard for heart failure without clinical correlation because the results from these tests serve as part of the reference standard for a clinical diagnosis; (4) were population based; (5) enrolled patients younger than 18 years; and (6) did not specifically include patients reporting dyspnea. We resolved disagreements between reviewers on study selection, assessment of quality, and abstraction of data by consensus.

Assessment of Study Quality

Study quality was assigned according to the grading scheme developed by Sackett et al³⁷ and previously used for this series.²⁴ Level 1 studies were primary prospective studies of the accuracy or precision of the clinical examination that involved comparisons of clinical findings (symptom or sign) with a reference standard of diagnosis among a large number (sufficient to have narrow confidence limits on the resulting sensitivity, specificity, or LRs) of consecutive or random patients with dyspnea. For precision studies, this required 2 or more independent blinded raters of symptoms or signs in a large number of patients. Level 2 studies were similar to level 1 but with smaller numbers of patients. Level 3 studies were comparisons of clinical findings with a reference standard of diagnosis among nonconsecutive or nonrandom patients with dyspnea. Studies of a retrospective nature were included as level 3. Level 4 studies were comparisons of clinical findings with a reference standard of diagnosis among convenience samples of patients who obviously have the target condition. Finally, level 5 studies were comparisons of clinical findings with a reference standard of unknown or uncertain validity among convenience samples of patients and, perhaps, healthy patients.

Statistical Methods

Two authors (C.S.W. and N.T.A.) independently extracted data for analysis. Published raw data were used to construct 2×2 contingency tables for each clinical variable. Where data for the same variable were available from 2 or more sources, meta-analytic techniques were applied to combine results across studies. When multiple articles from the same group were found, the studies were carefully reviewed to ensure no data were analyzed in duplicate. Summary positive and negative LRs and 95% confidence intervals (CIs) were calculated using random-effects models based on the delta method.³⁸ We display only the CIs of the LRs in the data tables because these values are most useful to clinicians and include the sensitivity and specificity in the calculation. The choice of random-effects measures decreases the risk of CIs that are too optimistically narrow.

Sensitivity is defined as the proportion of patients with heart failure who have a particular finding; specificity is the proportion of patients without heart failure who do not have the particular finding. The positive LR is the change in the odds of having heart failure when a particular finding is present, whereas the negative LR is the change in the odds of having heart failure when the particular finding is absent.

RESULTS

Search Results

A total of 815 citations were identified in our literature search. Of these, 682 were excluded after review of their titles and abstracts, with 133 studies remaining. These studies were reviewed in detail and we identified a total of 22 studies that evaluated the role of the clinical examination or basic routine investigation (chest radiograph, ECG, serum BNP) in patients

with undifferentiated dyspnea and that also met our inclusion criteria.^{12,31,32,36,39-56}

Study Characteristics

Only studies of sufficient quality (levels 1–3) were considered for the quantitative analysis. Of the 22 studies meeting inclusion criteria, 18 were included in the meta-analysis ([Table 16-5](#))^{12,31,36,39-48,52-56} whereas the remaining 4 studies^{32,49-51} were level 4 or 5 and were not included in the evidence tables.

Table 16-5 Summary of Studies in Emergency Department Patients

Source, y	Study Quality ^a	Study Design	Study Criteria	Total Men, No. (%)	Mean Age, y	Incidence of Heart Failure, %	Criterion Standard; Objective Measure
Mueller et al, ⁵⁶ 2005	1	Prospective	Inclusion: ED with dyspnea. Exclusion: acute myocardial infarction, trauma	251 (93)	73	55	Retrospective review by 1 physician; echocardiography
Lainchbury et al, ⁴² 2003	1	Prospective	Inclusion: ED with dyspnea, able to give blood within 8 h of arrival. Exclusion: n/a	205 (49)	70	34	Retrospective review by 2 independent cardiologists; echocardiography, RVG
Logeart et al, ⁴³ 2002	1	Prospective	Inclusion: ED with acute severe dyspnea. Exclusion: acute myocardial infarction, chest injury, recent surgery, therapy instituted >2 h before arrival in ED, emergency echocardiography not feasible	163 (67)	67	71	Retrospective review by 2 independent cardiologists and 1 pulmonologist; echocardiography, CC, RVG, PFT
Knudsen et al, ⁴⁴ 2004	2	Prospective	Inclusion: ED with dyspnea. Exclusion: chest pain, dyspnea clearly not secondary to heart failure	155 (45)	NA ^b	48	Retrospective review by 2 independent cardiologists; echocardiography, CC, RVG, PFT
Bayes-Genis et al, ⁴⁵ 2004	2	Prospective	Inclusion: ED with dyspnea, aged 40–88 y. Exclusion: NYHA classes I and II, dyspnea secondary to chest trauma or cardiac tamponade, acute coronary syndromes without dyspnea, severe renal insufficiency, liver cirrhosis	89 (60)	71	83	Retrospective review by 2 independent cardiologists; echocardiography, PFT
Villacorta et al, ⁴⁶ 2002	2	Prospective	Inclusion: ED with dyspnea. Exclusion: obvious diagnosis of dyspnea, acute coronary syndromes without dyspnea	70 (47)	72	51	Retrospective review by 1 cardiologist; echocardiography
Davis et al, ⁴⁷ 1994	2	Prospective	Inclusion: ED with dyspnea requiring admission. Exclusion: obvious cause of dyspnea, severe renal failure, acute chest pain	52 (40)	74	61	Retrospective review by committee of physicians and a radiologist; echocardiography, PFT
Marantz et al, ³¹ 1990	2	Prospective	Inclusion: ED with dyspnea, aged ≥ 40 y, English speaking, able to consent, presented during study hours. Exclusion: clinically unstable, non-English speaking, disoriented or unable to cooperate, refusal to consent, left against medical advice	51 (39)	64	45	Retrospective review by 1 physician; echocardiography
Alibay et al, ⁵⁴ 2005	3	Convenience sample	Inclusion: ED with dyspnea. Exclusion: n/a	160 (48)	80	38	Retrospective review by 2 independent cardiologists; echocardiography
Ray et al, ⁵⁵ 2004	3	Convenience sample	Inclusion: ED with dyspnea < 2 wk, aged ≥ 65 y, respiratory rate > 25/min or $\text{Pao}_2 < 70 \text{ mm Hg}$ or $\text{Paco}_2 > 45 \text{ mm Hg}$ or $\text{SpO}_2 < 92\%$. Exclusion: none	308 (49)	80	54	Retrospective review by 2 independent experts; echocardiography, high-resolution computed tomographic scan, PFT

(continued)

Table 16-5 Summary of Studies in Emergency Department Patients (Continued)

Source, y	Study Quality ^a	Study Design	Study Criteria	Total Men, No. (%)	Mean Age, y	Incidence of Heart Failure, %	Criterion Standard; Objective Measure
Springfield et al, ¹² 2004	3	Convenience sample	Inclusion: ED with dyspnea or respiratory rate > 20/min or $\text{PaO}_2 < 90$ mm Hg on room air. Exclusion: pregnancy, aged ≤ 18 y, trauma patients, unconscious or unable to speak, < 3 ft 11 in or > 7 ft 8 in tall, < 66 or > 341 lb	38 (42)	67	32	Retrospective review by 1 physician; echocardiography
Morrison et al, ³⁶ 2002	3	Convenience sample	Inclusion: ED with dyspnea. Exclusion: dyspnea clearly not secondary to heart failure, unstable angina/myocardial infarction without dyspnea	321 (95)	NA	42	Retrospective review by 2 independent cardiologists; echocardiography, CC, RVG, PFT
Maisel et al, ³⁹ 2002	3	Prospective	Inclusion: ED with dyspnea as prominent symptom. Exclusion: aged ≤ 18 y, dyspnea clearly not secondary to heart failure, acute myocardial infarction, unstable angina without dyspnea, renal failure on dialysis or creatinine clearance < 0.25 mL/s	1586 (56)	64	47	Retrospective review by 2 independent cardiologists; echocardiography, CC, RVG, PFT
McCullough et al, ⁴⁰ 2002	3	See Maisel et al ³⁹	Subgroup of Maisel et al ³⁹ with information recorded for ED physician assessment of probability of heart failure	1538 (56)	64	47	See Maisel et al ³⁹
Dao et al, ⁴⁸ 2001	3	Convenience sample	Inclusion: ED with dyspnea. Exclusion: dyspnea clearly not secondary to heart failure, acute coronary syndromes without dyspnea	250 (94)	63	39	Retrospective review by 2 independent cardiologists; echocardiography, CC, RCG, PFT

Abbreviations: CC, cardiac catheterization; ED, emergency department; n/a, not applicable; NYHA, New York Heart Association (classification of heart disease); PFT, pulmonary function test; RVG, radionuclide ventriculography; SpO_2 , peripheral oxygen saturation.

^aStudy quality was assigned according to the grading scheme developed by Sackett et al³⁷ and previously used for this series.²⁴ See also "Assessment of Study Quality" in the "Methods" section for more details.

^bNA denotes that the mean age was not published in the source article.

Precision of Clinical Examination and Investigations

Precision refers to the degree of variation between observers (interobserver variation) or within observers (intraobserver variation) for a particular finding. No study has specifically addressed the interobserver or intraobserver variability in the recording of findings in dyspneic patients ultimately diagnosed with the clinical syndrome of heart failure. However, analogous work has been done in other diagnoses, including pulmonary diseases and acute coronary syndromes, and in comparison with echocardiography, nuclear imaging, and cardiac catheterization.^{24,25,29,30,57-63} In general, there is much variability in the precision of clinical findings associated with heart failure, reflecting the potentially subtle nature of findings and variable examination skills of the clinician.

Accuracy of the Clinical Examination

Thirteen studies examined the accuracy of the clinical examination for predicting the presence of heart failure in dyspneic patients assessed in the ED. The sensitivities, specificities, and corresponding positive and negative LRs for the findings are shown in Table 16-6.

Overall Clinical Gestalt

The overall clinical gestalt of the initial treating ED physician was associated with a high LR+ (4.4; 95% CI, 1.8-10) for a final diagnosis of heart failure. When the emergency physician assessed the dyspneic patient as unlikely to have heart failure, the odds decreased by about half (LR, 0.45; 95% CI, 0.28-0.73).

Historical Items

The most useful historical features in confirming the presence of heart failure were congestive heart failure (LR, 5.8; 95% CI, 4.1-8.0), myocardial infarction (LR, 3.1; 95% CI, 2.0-4.9), or coronary artery disease (LR, 1.8; 95% CI, 1.1-2.8). Likewise, patients without a history of heart failure (LR, 0.45; 95% CI, 0.38-0.53), coronary artery disease (LR, 0.68; 95% CI, 0.48-0.96), or myocardial infarction (LR, 0.69; 95% CI, 0.58-0.82) were less likely to have their dyspnea explained by current heart failure. The results of other historical findings in Table 16-6 had LR CIs that included 1.

Symptoms

The presence of paroxysmal nocturnal dyspnea (LR, 2.6; 95% CI, 1.5-4.5), orthopnea (LR, 2.2; 95% CI, 1.2-3.9), or dyspnea

Table 16-6 Summary of Diagnostic Accuracy of Findings on History and Physical Examination in Emergency Department Patients

Finding	Pooled		Summary LR (95% CI) ^a	
	Sensitivity	Specificity	Positive	Negative
Initial clinical judgment ^{12,31,40,55}	0.61	0.86	4.4 (1.8-10.0)	0.45 (0.28-0.73)
History				
Heart failure ^{36,41,43,45,48,53,56}	0.60	0.90	5.8 (4.1-8.0)	0.45 (0.38-0.53)
Myocardial infarction ^{41,43-45,48,53}	0.40	0.87	3.1 (2.0-4.9)	0.69 (0.58-0.82)
Coronary artery disease ^{36,44,53,56}	0.52	0.70	1.8 (1.1-2.8)	0.68 (0.48-0.96)
Dyslipidemia ⁴⁵	0.23	0.87	1.7 (0.43-6.9)	0.89 (0.69-1.1)
Diabetes mellitus ^{43-45,48,56}	0.28	0.83	1.7 (1.0-2.7)	0.86 (0.73-1.0)
Hypertension ^{36,41,43-45,48,53,56}	0.60	0.56	1.4 (1.1-1.7)	0.71 (0.55-0.93)
Smoker ⁴⁵	0.62	0.27	0.84 (0.58-1.2)	1.4 (0.58-3.6)
Chronic obstructive pulmonary disease ^{36,45,48,53}	0.34	0.57	0.81 (0.60-1.1)	1.1 (0.95-1.4)
Symptoms				
Paroxysmal nocturnal dyspnea ^{36,45,48,53,56}	0.41	0.84	2.6 (1.5-4.5)	0.70 (0.54-0.91)
Orthopnea ^{36,41,43-45,48,53,56}	0.50	0.77	2.2 (1.2-3.9)	0.65 (0.45-0.92)
Edema ^{36,48,53}	0.51	0.76	2.1 (0.92-5.0)	0.64 (0.39-1.1)
Dyspnea on exertion ^{36,48}	0.84	0.34	1.3 (1.2-1.4)	0.48 (0.35-0.67)
Fatigue and weight gain ³⁶	0.31	0.70	1.0 (0.74-1.4)	0.99 (0.85-1.1)
Cough ^{36,45,48,53,56}	0.36	0.61	0.93 (0.70-1.2)	1.0 (0.87-1.3)
Physical Examination				
Third heart sound (ventricular filling gallop) ^{36,41,43-45,48,53,56}	0.13	0.99	11 (4.9-25)	0.88 (0.83-0.94)
Abdominojugular reflux ³¹	0.24	0.96	6.4 (0.81-51)	0.79 (0.62-1.0)
Jugular venous distention ^{36,41,43-45,48,53,56}	0.39	0.92	5.1 (3.2-7.9)	0.66 (0.57-0.77)
Rales ^{36,41,43-45,48,53,56}	0.60	0.78	2.8 (1.9-4.1)	0.51 (0.37-0.70)
Any murmur ^{36,44,48,53}	0.27	0.90	2.6 (1.7-4.1)	0.81 (0.73-0.90)
Lower extremity edema ^{41,43-45,53,56}	0.50	0.78	2.3 (1.5-3.7)	0.64 (0.47-0.87)
Valsalva maneuver ³¹	0.73	0.65	2.1 (1.0-4.2)	0.41 (0.17-1.0)
Systolic blood pressure < 100 mm Hg ⁴⁸	0.06	0.97	2.0 (0.60-6.6)	0.97 (0.91-1.0)
Fourth heart sound (atrial gallop) ^{36,48,53}	0.05	0.97	1.6 (0.47-5.5)	0.98 (0.93-1.0)
Systolic blood pressure ≥ 50 mm Hg ⁴⁸	0.28	0.73	1.0 (0.69-1.6)	0.99 (0.84-1.2)
Wheezing ^{36,44,45,48,53}	0.22	0.58	0.52 (0.38-0.71)	1.3 (1.1-1.7)
Ascites ⁴⁸	0.01	0.97	0.33 (0.04-2.9)	1.0 (0.99-1.1)

Abbreviations: CI, confidence interval; LR, likelihood ratio.

^aLRs are not independent of each other and should not be multiplied in series when multiple findings are considered.

on exertion (LR, 1.3; 95% CI, 1.2-1.4) increased the likelihood of heart failure. Likewise, the absence of dyspnea on exertion (LR, 0.48; 95% CI, 0.35-0.67), orthopnea (LR, 0.65; 95% CI, 0.45-0.92), or paroxysmal nocturnal dyspnea (LR, 0.70; 95% CI, 0.54-0.91) decreased the likelihood of heart failure. The results of other findings in Table 16-6 had CIs that included 1.

Physical Examination

The presence of a third heart sound (ventricular filling gallop) increased the likelihood of heart failure the most (LR, 11; 95% CI, 4.9-25). The presence of several other findings had CIs that excluded 1: jugular venous distention (LR, 5.1; 95% CI, 3.2-7.9), pulmonary rales (LR, 2.8; 95% CI, 1.9-4.1), any cardiac murmur (LR, 2.6; 95% CI, 1.7-4.1), and leg edema (LR, 2.3; 95% CI, 1.5-3.7). The presence of an abnormal

abdominojugular reflux response (LR, 6.4; 95% CI, 0.81-51) had a high LR, but its evaluation in only 1 study of 51 patients led to broad CIs.³¹ An abnormal response to the Valsalva maneuver in the same study had an LR of 2.1 but the lower limit of the 95% CI was 1.0. The presence of the other findings in Table 16-6 did not appear useful for assessing the likelihood of heart failure in dyspneic patients.

The absence of pulmonary rales (LR, 0.51; 95% CI, 0.37-0.70), leg edema (LR, 0.64; 95% CI, 0.47-0.87), or jugular venous distention (LR, 0.66; 95% CI, 0.57-0.77) was the most useful finding that decreased the likelihood of heart failure. Wheezing also decreased the likelihood that a dyspneic patient had heart failure (LR, 0.52; 95% CI, 0.38-0.71). The absence of a third heart sound or a murmur decreased the likelihood of heart failure but the point estimate of the LR of these findings approached 1. The absence of the other findings in Table 16-6 did not appear useful as the CI included 1. Diaphoresis as a sign

of heart failure was of uncertain validity, having been evaluated in only 2 studies that were each of level 4 quality.^{49,50}

Accuracy of Chest Radiographs

Seven studies examined the accuracy of various chest radiograph findings in the ED setting (Table 16-7). The presence of any of these findings (except for any edema) had high positive LRs with CIs exceeding 1 and therefore, increased the likelihood of heart failure in dyspneic patients. The presence of pulmonary venous congestion (distention of pulmonary veins and redistribution to the apices) (n = 4 studies; summary LR, 12; 95% CI, 6.8-21) and cardiomegaly (n = 6 studies; summary LR, 3.3; 95% CI, 2.4-4.7) increased the likelihood of heart failure and has undergone more extensive evaluation so that the results may be more reliable. The presence of interstitial edema also had a high LR (n = 2 studies; summary LR, 12; 95% CI, 5.2-27). The presence of pneumonia or hyperinflation decreased the likelihood of heart failure but was assessed in only 1 study.

The most extensively evaluated chest radiograph findings (pulmonary venous congestion and cardiomegaly) were also the findings that, when absent, had an LR that was appreciably different from 1. The absence of cardiomegaly was particularly useful (LR, 0.33; 95% CI, 0.23-0.48), with narrower CIs than the absence of pulmonary venous congestion (LR, 0.48; 95% CI, 0.28-0.83).

Accuracy of ECG

Seven studies examined the accuracy of various ECG findings in the ED setting (Table 16-7). The presence of atrial fibrillation in a dyspneic patient was the most important (LR, 3.8; 95% CI, 1.7-8.8) and evaluated in several studies (n = 5 studies).

The presence of new T-wave changes (LR, 3.0; 95% CI, 1.7-5.3) or abnormal ECG findings (LR, 2.2; 95% CI, 1.6-3.1) increased the likelihood of heart failure but was evaluated in fewer studies. A completely normal ECG result (LR, 0.64; 95% CI, 0.47-0.88) decreased the likelihood of heart failure and was the only normal finding that had a negative LR with a clinically meaningful difference from 1.

Accuracy of BNP

Eleven studies examined the operating characteristics of various cutoffs of serum BNP in the ED setting (Table 16-8). Eight of these reported pharmaceutical industry sponsorship, 2 did not disclose funding sources, and only 1 study reported no pharmaceutical relationship.

As the BNP cutoff increased, the positive LR generally increased. Thus, the higher the value of BNP, the more suggestive it was of heart failure. However, no BNP threshold indicated the presence of heart failure with certainty. At any BNP threshold up to 250 pg/mL, values lower than the threshold always made heart failure much less likely in comparison to those with values > 250 pg/mL. However, the serial LRs show that, overall, as the BNP increases, the likelihood of heart failure increases (Table 16-8). (LR, 0.06-0.15).

BNP levels must be interpreted differently for patients with renal insufficiency. According to an analysis of data from the Breathing Not Properly Multinational Study,^{39,64} no adjustment in the 100 pg/mL threshold appears necessary for patients with an estimated glomerular filtration rate of 60 to 89 mL/min/1.73 m², with an area under the receiver operating characteristic curve of 0.90 (a measure of overall accuracy). The loss of accuracy with worsening renal function can be minimized by using thresholds of 225 and 201 pg/mL, respectively, for patients with estimated

Table 16-7 Summary of Diagnostic Accuracy of Findings on Chest Radiograph and Electrocardiogram in Emergency Department Patients

Finding	Pooled		Summary LR (95% CI) ^a	
	Sensitivity	Specificity	Positive	Negative
Chest Radiograph				
Pulmonary venous congestion ^{36,41,45,48b}	0.54	0.96	12 (6.8-21)	0.48 (0.28-0.83)
Interstitial edema ^{41,53}	0.34	0.97	12 (5.2-27)	0.68 (0.54-0.85)
Alveolar edema ⁴¹	0.06	0.99	6.0 (2.2-16)	0.95 (0.93-0.97)
Cardiomegaly ^{36,41,43-45,48}	0.74	0.78	3.3 (2.4-4.7)	0.33 (0.23-0.48)
Pleural effusion(s) ^{36,41}	0.26	0.92	3.2 (2.4-4.3)	0.81 (0.77-0.85)
Any edema ^{43,44}	0.70	0.77	3.1 (0.60-16)	0.38 (0.11-1.3)
Pneumonia ⁴¹	0.04	0.92	0.50 (0.29-0.87)	1.0 (1.0-1.1)
Hyperinflation ⁴¹	0.03	0.92	0.38 (0.20-0.69)	1.1 (1.0-1.1)
ECG				
Atrial fibrillation ^{36,43,44,48,56}	0.26	0.93	3.8 (1.7-8.8)	0.79 (0.65-0.96)
New T-wave changes ³⁶	0.24	0.92	3.0 (1.7-5.3)	0.83 (0.74-0.92)
Any abnormal finding ^{41,53}	0.50	0.78	2.2 (1.6-3.1)	0.64 (0.47-0.88)
ST elevation ^{36,48}	0.05	0.97	1.8 (0.80-4.0)	0.98 (0.94-1.0)
ST depression ^{36,48}	0.11	0.94	1.7 (0.97-2.9)	0.95 (0.90-1.0)

Abbreviations: CI, confidence interval; ECG, electrocardiogram; LR, likelihood ratio.

^aLRs are not independent of each other and should not be multiplied in series when multiple findings are considered.

^bPulmonary venous congestion, manifest as distention of pulmonary veins and redistribution to the apices.

Table 16-8 Summary Operating Characteristics of Serum Brain Natriuretic Peptide in Emergency Department Patients

	Summary		
	Sensitivity	Specificity	Summary Serial LR (95% CI)
Clinical judgment or BNP $\geq 100 \text{ pg/mL}$ ^{40a}	0.94	0.70	3.1 (2.8-3.5)
BNP alone, pg/mL ^b			
≥ 250 ^{36,43,55}	0.89	0.81	4.6 (2.6-8.0)
≥ 200 ^{36,42-44,46,54,55}	0.92	0.75	3.7 (2.6-5.4)
≥ 150 ^{39,43,44,48,54-56}	0.89	0.71	3.1 (2.1-4.5)
≥ 100 ^{36,39,42-44,47,48,54-56}	0.93	0.66	2.7 (2.0-3.9)
≥ 80 ^{39,43,47,48}	0.96	0.71	3.3 (1.8-6.3)
≥ 50 ^{39,44,54}	0.97	0.44	1.7 (1.2-2.6)
<50 ^{39,44,54}	0.97	0.44	0.06 (0.03-0.12)

Abbreviations: BNP, brain natriuretic peptide; CI, confidence interval; LR, likelihood ratio.

^aEither an initial clinical probability of heart failure of at least 80% or BNP of at least 100 pg/mL was considered a positive result. A negative result was a clinical probability of heart failure lower than 80% and BNP lower than 100 pg/mL.

^bThe values shown represent the summary values at each specified threshold. For example, if the BNP is 85 pg/mL, then the LR for heart failure is 3.3. However, if the BNP is less than 50 pg/mL, the LR is 0.06.

glomerular filtration rates of 15 to 29 and 30 to 59 mL/min/1.73 m² (areas under receiver operating characteristic curves of 0.86 and 0.81, respectively). The utility of BNP levels in patients with advanced renal insufficiency (estimated glomerular filtration rate < 15 mL/min/1.73 m² or receiving dialysis) is unclear as these patients were not included in that study.

Accuracy of Findings in Patients With History of Pulmonary Disease

One study (Table 16-9) examined the accuracy of symptoms, signs, ECG, and serum BNP in diagnosing heart failure in dyspneic ED patients with a history of asthma or COPD.⁵² This study was a subgroup analysis of the Breathing Not Properly Multinational Study.³⁹

Initial Clinical Gestalt

A high initial clinical suspicion by the emergency physician ($\geq 80\%$ probability) was associated with a high likelihood for a final diagnosis of heart failure (LR, 9.9; 95% CI, 5.3-18), whereas an intermediate (21%-79%) or low ($\leq 20\%$) initial clinical suspicion decreased the likelihood of heart failure (LR, 0.65; 95% CI, 0.55-0.77) but did not exclude it. In fact, 32% of patients in the intermediate suspicion group and 9% of patients in the low clinical suspicion group were ultimately diagnosed with heart failure. Assigning a lower probability to the low suspicion group (eg, $\leq 5\%$) would likely have reduced misclassification in that study.

Historical Items

The presence of most historical findings in Table 16-9 increased the likelihood of heart failure with CIs excluding 1. A history of atrial fibrillation (LR, 4.1; 95% CI, 2.5-6.6) or coronary bypass surgery (LR, 2.8; 95% CI, 1.3-5.8) was the most useful finding that increased the likelihood of heart failure. The absence of relevant historical features did not result in clinically meaningful LRs less than 1, other than perhaps the absence of coronary artery disease (LR, 0.67; 95% CI, 0.54-0.84).

Symptoms

Only the absence of orthopnea (LR, 0.68; 95% CI, 0.48-0.95) had an LR that was appreciably different from 1. Thus, symptoms were not particularly useful among dyspneic patients with lung disease in determining who might also have heart failure.

Physical Examination

The presence of a third heart sound had a high diagnostic value for heart failure (LR, 57; 95% CI, 7.6-425). Other useful physical examination findings, when present, included jugular venous distention (LR, 4.3; 95% CI, 2.8-6.5), lower extremity edema (LR, 2.7; 95% CI, 2.2-3.5), pulmonary rales (LR, 2.6; 95% CI, 2.1-3.3), or hepatic congestion (LR, 2.4; 95% CI, 1.2-4.7). The absence of pulmonary rales (LR, 0.39; 95% CI, 0.28-0.55), lower extremity edema (LR, 0.41; 95% CI, 0.30-0.57), or jugular venous distention (LR, 0.65; 95% CI, 0.54-0.78) decreased the likelihood of heart failure.

Chest Radiograph

The presence of edema was the most useful radiographic finding for increasing the likelihood of heart failure (LR, 11; 95% CI, 5.8-22). Other very useful findings were cardiomegaly (LR, 7.1; 95% CI, 4.5-11) or pleural effusion(s) (LR, 4.6; 95% CI, 2.6-8.0). A normal chest radiograph result (LR, 0.11; 95% CI, 0.04-0.28), absence of cardiomegaly (LR, 0.54; 95% CI, 0.44-0.67), or absence of edema (LR, 0.68; 95% CI, 0.58-0.79) decreased the likelihood of heart failure.

Electrocardiogram

The presence of ECG findings of atrial fibrillation (LR, 6.0; 95% CI, 3.4-10), ischemic ST-T-wave changes (LR, 4.6; 95% CI, 2.4-8.7), or Q waves (LR, 3.1; 95% CI, 1.8-5.5) was helpful in suggesting a diagnosis of heart failure in the dyspneic ED patient with a history of pulmonary disease. No single ECG result had clinically useful outcomes for decreasing the likelihood of heart failure.

Table 16-9 Diagnostic Accuracy of History, Physical Examination, and Tests of Cardiac Function in Emergency Department Patients With History of Asthma or Chronic Obstructive Pulmonary Disease^a

Finding	Sensitivity	Specificity	Positive LR (95% CI) ^b	Negative LR (95% CI) ^b
Initial clinical judgment	0.37	0.96	9.9 (5.3-18)	0.65 (0.55-0.77)
History				
Atrial fibrillation	0.32	0.92	4.1 (2.5-6.6)	0.74 (0.63-0.85)
Coronary artery bypass grafting	0.13	0.95	2.8 (1.3-5.8)	0.92 (0.84-0.99)
Myocardial infarction	0.25	0.88	2.2 (1.4-3.5)	0.84 (0.74-0.96)
Diabetes mellitus	0.26	0.87	2.0 (1.3-3.2)	0.85 (0.74-0.97)
Coronary artery disease	0.49	0.75	2.0 (1.5-2.6)	0.67 (0.54-0.84)
Angina	0.21	0.88	1.7 (1.0-2.8)	0.90 (0.80-1.0)
Hypertension	0.54	0.55	1.2 (0.95-1.5)	0.84 (0.65-1.1)
Symptoms				
Orthopnea	0.70	0.44	1.3 (1.1-1.5)	0.68 (0.48-0.95)
Fatigue	0.74	0.34	1.1 (0.96-1.3)	0.79 (0.54-1.2)
Nocturnal cough	0.49	0.47	0.93 (0.73-1.2)	1.1 (0.85-1.4)
Physical Examination				
Third heart sound (ventricular filling gallop)	0.17	1.00	57 (7.6-425)	0.83 (0.75-0.91)
Jugular venous distention	0.41	0.90	4.3 (2.8-6.5)	0.65 (0.54-0.78)
Lower extremity edema	0.69	0.75	2.7 (2.2-3.5)	0.41 (0.30-0.57)
Rales	0.71	0.73	2.6 (2.1-3.3)	0.39 (0.28-0.55)
Hepatic congestion	0.14	0.94	2.4 (1.2-4.7)	0.91 (0.84-1.0)
Enlarged heart	0.03	0.98	1.6 (0.43-6.2)	0.99 (0.95-1.0)
Wheezing	0.42	0.50	0.85 (0.65-1.1)	1.2 (0.94-1.4)
Chest Radiograph				
Edema	0.34	0.97	11 (5.8-22)	0.68 (0.58-0.79)
Cardiomegaly	0.49	0.93	7.1 (4.5-11)	0.54 (0.44-0.67)
Pleural effusion(s)	0.26	0.94	4.6 (2.6-8.0)	0.78 (0.69-0.89)
Pneumonia	0.08	0.92	1.0 (0.46-2.3)	1.0 (0.93-1.1)
Hyperinflation	0.08	0.85	0.53 (0.25-1.1)	1.1 (1.0-1.2)
Normal	0.05	0.57	0.11 (0.04-0.28)	1.7 (1.5-1.8)
Electrocardiogram				
Atrial fibrillation	0.31	0.95	6.0 (3.4-10)	0.73 (0.63-0.84)
Ischemic ST-T waves	0.21	0.95	4.6 (2.4-8.7)	0.83 (0.74-0.93)
Q waves	0.22	0.93	3.1 (1.8-5.5)	0.84 (0.75-0.94)
BNP ≥ 100 pg/mL	0.93	0.77	4.1 (3.3-5.0)	0.09 (0.04-0.19)

Abbreviations: BNP, brain natriuretic peptide; CI, confidence interval; LR, likelihood ratio.

^aAdapted from McCullough et al.⁵²

^bLikelihood ratios are not independent of each other and should not be multiplied in series when multiple findings are considered.

Brain Natriuretic Peptide

BNP levels can increase in patients with chronic pulmonary diseases because of right ventricular strain. Nevertheless, BNP appears to still be useful in these patients. Studies have demonstrated that BNP levels are significantly higher in patients with a history of chronic lung disease but acute dyspnea from heart failure compared with those with a history of heart failure but acute dyspnea from lung disease.^{36,65}

Serum BNP for dyspneic patients with a history of asthma or COPD was useful for identifying heart failure (BNP ≥ 100 pg/mL: LR, 4.1; 95% CI, 3.3-5.0). However, it was more powerful for excluding heart failure when low (BNP < 100 pg/mL: LR, 0.09; 95% CI, 0.04-0.19). However, this was only 1 study, and thus,

the optimal cutoff for BNP to diagnose or exclude clinical heart failure in dyspneic patients with chronic lung diseases is unclear.

COMMENT

It is both important and difficult to rapidly differentiate among the common causes of dyspnea in ED patients. The syndrome of heart failure requires appropriate symptoms, along with objective measures of cardiac dysfunction.⁵ Although sophisticated and invasive tests such as Swan-Ganz catheterization can help to distinguish between cardiac and pulmonary causes of dyspnea, they are frequently unavailable in the acute setting, and thus, the diagnosis of heart failure and the decision to institute therapy on

an emergency basis rests on the bedside clinical assessment (chest radiograph, ECG, and recently, serum BNP). Relying purely on echocardiography to diagnose clinical heart failure is also problematic because it is often not easily accessible, requires specialized training,⁶⁶ and may not always truly reflect the current cause of dyspnea.⁶⁷ That is, not every patient presenting with heart failure will have a diminished left ventricular EF; patients with diastolic heart failure, for instance, may have elevated filling pressures and dyspnea in the presence of normal EF. The reverse is also true in that patients with a decreased left ventricular EF may be dyspneic from noncardiac causes such as COPD, and furthermore, the severity of impairment of EF does not always correlate with subjective severity of dyspnea.⁶⁸

In this systematic review, many features on clinical examination, chest radiograph, ECG, and serum BNP were useful in diagnosing heart failure in adult ED patients presenting with dyspnea in whom heart failure was suspected. Features listed in **Box 16-1** were assessed in more than 1 study and were useful when either present or absent. Other findings may prove useful when evaluated further.

CLINICIAN'S OVERALL ASSESSMENT

Our results are consistent with those of Marcus et al.⁶⁹ They recently studied patients undergoing elective left-sided heart

Box 16-1 Features Useful in Diagnosing Heart Failure in Adult Emergency Department Patients With Dyspnea

HISTORICAL FEATURES

- Heart failure
- Myocardial infarction
- Coronary artery disease

SYMPTOMS

- Paroxysmal nocturnal dyspnea
- Orthopnea
- Dyspnea on exertion

PHYSICAL EXAMINATION

- Listening for a third heart sound (ventricular filling gallop)
- Jugular venous pressure assessment
- Auscultating for rales and wheezing
- Auscultating for a murmur
- Assessing the legs for edema

CHEST RADIOGRAPH

- Pulmonary venous congestion
- Interstitial edema
- Cardiomegaly
- Pleural effusion(s)

ELECTROCARDIOGRAM FINDINGS

- Atrial fibrillation
- An abnormal result

BRAIN Natriuretic Peptide

- Most useful when < 100 pg/mL for decreasing the likelihood of heart failure

catheterization, comparing the test characteristics of third and fourth heart sounds with objective measures of left ventricular dysfunction. Although the patient population and reference standard for heart failure were different in our review compared with theirs (eg, ventricular dysfunction vs a clinical diagnosis of heart failure), both studies found that third and fourth heart sounds had greater specificity than sensitivity and that a third heart sound had a better specificity than a fourth heart sound for the diagnosis of heart failure.

We did not find any studies examining combinations of historical and physical examination findings in making a diagnosis of heart failure. However, our analysis suggests that the initial clinical gestalt of the physician according to available information (history, physical examination, chest radiograph, ECG) is valuable. Because the overall clinical gestalt had LRs that approximate some of the individual findings, along with a lack of consistent multivariate models, we do not know whether all the symptoms and signs are independently useful. When clinicians are not confident in their clinical gestalt, they should preferentially rely on the results of the few findings that have LR estimates most different from 1.

A high initial clinical suspicion alone (LR, 4.4; 95% CI, 1.8-10) (Table 16-6) had a greater positive LR than a composite of high clinical suspicion, BNP level greater than or equal to 100 pg/mL, or both, which had a combined positive LR of 3.1 (95% CI, 2.8-3.5) (Table 16-8). This suggests that BNP may not contribute much more in patients for whom the initial clinical suspicion of heart failure was already high. However, in patients for whom the initial clinical suspicion of heart failure was not high, BNP at a threshold value of 100 pg/mL was useful, especially for excluding heart failure in this group of patients. To apply these results correctly, it is necessary that clinicians first quantify and acknowledge their clinical suspicion (eg, formulate a pretest probability). If the physician waits until the BNP results are available before establishing clinical suspicion, these tests are no longer independent and the clinical suspicion becomes biased by the BNP. The results of our BNP analysis add support to recent European guidelines for diagnosing heart failure, which state that BNP may be a clinically useful test to rule out heart failure because of its high negative predictive values.⁵ Clinicians should be aware that factors other than heart failure can affect serum BNP levels (**Box 16-2**). Algorithms for the use of the BNP test have been proposed⁷⁰ but not extensively validated.

Limitations

The results of our meta-analysis should be interpreted in the context of study limitations. One limitation of this review is the reference standard for heart failure (adjudication by a panel of physicians). Given the subjectivity and potential bias of such a standard, many of the studies had disagreement (up to 10%) among the adjudicators of whether heart failure was the contributing cause of dyspnea. However, in the absence of a true criterion standard for this clinical syndrome, the reference standard, although imperfect, is likely the best available and

consistent with the clinical focus of this review. Another limitation that arises from using a clinical reference standard is that the final diagnosis of heart failure may not have been made independently of the individual findings of interest. That is, the panel of physicians may have used some of the clinical findings in deciding whether patients ultimately had heart failure as the cause of their dyspnea. As such, this may overestimate sensitivities and specificities. Although this is a valid concern, we believe the effects on each finding would be small because the final diagnosis relied on a combination of information from many diverse sources, including any or all of the following: medical history, physical examination, routine laboratory tests, chest radiograph, ECG, heart failure scores, objective measures of cardiac function (eg, echocardiography, radionuclide ventriculography, radionuclide angiography, and left ventriculography at cardiac catheterization), pulmonary function tests, response to treatment, hospitalization course, and follow-up records.

Our data are derived from studies on patients presenting to the ED with dyspnea. Therefore, these results may not generalize to inpatients, outpatients in clinic settings who may have more chronic dyspnea, or patients without dyspnea. The 18 studies included in this meta-analysis represent diverse and heterogeneous populations with various comorbidities. The majority of the studies excluded patients with acute coronary syndromes and in whom an obvious cause of dyspnea (eg, pneumothorax, trauma) was present. All the studies of BNP excluded patients in whom dyspnea was clearly not secondary to heart failure. Therefore, the usefulness of BNP from our analysis can be applied only to patients in whom the diagnosis of heart failure is a consideration. In patients in whom the suspicion of heart failure is low (after taking a careful history and performing the physical examination, chest radiograph, and ECG), a BNP level is unlikely to affect diagnosis or management (eg, an obvious pulmonary etiology of dyspnea).

Other limitations include the inherent subjectivity of clinical findings on medical history, physical examination, chest radiograph, and ECG. It is impossible to confirm the accuracy of individual findings presented in each study, and no formal definitions were given. For example, we do not have standardized information on the technique used for each chest radiograph performed (anteroposterior, posteroanterior, portable).

The Bottom Line

The features evaluated in more than 1 study with the highest LRs ($LR > 3.5$) for diagnosing heart failure were the following: the overall clinical judgment, history of heart failure, a third heart sound, jugular venous distention, radiographic pulmonary venous congestion or interstitial edema, and electrocardiographic atrial fibrillation.

The features evaluated in more than 1 study with the lowest LRs ($LR < 0.60$) for diagnosing heart failure were the following: the overall clinical judgment, no history of heart failure, no dyspnea on exertion, the absence of rales, and the absence of radiographic pulmonary venous congestion, or cardiomegaly. The single finding that decreased the likelihood of heart failure the most was a BNP of less than 100 pg/mL (for patients with an estimated glomerular filtration rate of 15–60 mL/min/1.73 m², a

Box 16-2 Factors That Can Affect BNP Levels^a

FACTORS (OTHER THAN HEART FAILURE) THAT CAUSE ELEVATED BNP LEVELS

- Advanced age
- Renal failure
- Acute coronary syndromes
- Lung disease with cor pulmonale
- Acute large pulmonary embolism
- High-output cardiac states

FACTORS THAT DECREASE BNP IN THE SETTING OF HEART FAILURE

- Acute pulmonary edema
- Stable New York Heart Association class I patients with low EF
- Acute mitral regurgitation
- Mitral stenosis
- Atrial myxoma

^aAdapted from Maisei.⁷⁰

threshold of 201 pg/mL can be used). However, the clinician must always remember to first quantify and acknowledge his or her clinical suspicion according to the clinical examination before interpreting the BNP result.

In the subgroup of ED patients with a history of asthma or COPD, the features that strongly suggested a diagnosis of heart failure were the overall clinical assessment, a third heart sound, radiographic edema or cardiomegaly, and electrocardiographic atrial fibrillation. The features that suggested the diagnosis was not heart failure were normal chest radiograph result and a low serum BNP level (<100 pg/mL). However, these results are from a subgroup analysis in 1 study and require confirmation.

Although the findings of this study are useful when dyspneic patients suspected of having heart failure were assessed, no individual feature is sufficiently powerful in isolation to rule heart failure in or out. Therefore, an overall clinical impression based on all available information is best. If the appropriate constellation of findings with high LRs for heart failure is present, that may be sufficient to warrant empirical treatment without further urgent investigations. Conversely, if the clinical suspicion of heart failure is low (eg, pulmonary disease), the physician should investigate and treat other causes of dyspnea.

CLINICAL SCENARIOS—RESOLUTION

CASE 1 The patient has many features that raise the suspicion of heart failure, such as previous myocardial infarction (LR, 3.1), previous heart failure (LR, 5.8), orthopnea (LR, 2.2), paroxysmal nocturnal dyspnea (LR, 2.6), elevated jugular venous pressure (LR, 5.1), a third heart sound (LR, 11.0), rales (LR, 2.8), extremity edema (LR, 2.3), cardiomegaly (LR, 3.3), and atrial fibrillation (LR, 3.8) and only the single feature of wheezing (LR, 0.52) that decreases the suspicion slightly. The overall constellation of symptoms and signs is so suggestive of heart failure that additional testing is not needed to make the diagnosis.

CASE 2 Both heart failure and obstructive airways disease are considerations. The symptoms of dyspnea on exertion and cough are not helpful in making a diagnosis of heart failure because their LRs are close to 1. Rales (LR, 2.8) and ECG showing ST depression (LR, 1.7) both increase the likelihood of heart failure, but more important, the findings of pulmonary venous congestion and interstitial edema are both associated with large LRs (>10) that significantly increase the suspicion for heart failure. Wheezing reduces the likelihood somewhat (LR, 0.52). According to the information available, the patient likely has acute heart failure and should be treated without waiting for further tests. An ECG should be ordered nonurgently. Furthermore, the patient may also be having a superimposed COPD or asthma exacerbation. The physician should consider ordering pulmonary function tests to confirm a diagnosis of obstructive airways disease.

CASE 3 There are some features that increase the likelihood of heart failure, such as history of myocardial infarction (LR, 2.2), elevated jugular venous pressure (LR, 4.3), lower extremity edema (LR, 2.7), and Q waves (LR, 3.1), whereas other features decrease the likelihood (wheezing, LR, 0.85; and normal chest radiograph, (LR, 0.11). According to these LRs, there are insufficient data to make or rule out a diagnosis of heart failure. In this case, a BNP level could be helpful. If it were less than 100 pg/mL, heart failure would be unlikely (LR, 0.09). If it were elevated, the probability of heart failure is higher but not diagnostic. More urgent echocardiogram and pulmonary function studies would be appropriate next steps.

Author Affiliations at the Time of the Original Publication

Department of Medicine (Drs Wang, FitzGerald, and Ayas) and Division of Respiratory Medicine, Vancouver Hospital and Health Science Centre (Drs FitzGerald and Ayas), University of British Columbia, Vancouver, Canada; Centre for Clinical Epidemiology and Evaluation (Drs Wang, FitzGerald, Schulzer, and Ayas), Vancouver Coastal Health Research Institute, Vancouver, British Columbia, Canada; and Pacific Parkinson Research Centre, University of British Columbia (Mr Mak), Vancouver, Canada.

Acknowledgments

None of the authors of this review has any professional or financial affiliations with manufacturers of BNP assays.

Dr Ayas is a recipient of a New Investigator Award from the Canadian Institute for Health Research and BC Lung Association, a Departmental Scholar Award from the University of British Columbia Department of Medicine, and a Scholar Award from the Michael Smith Foundation for Health Research. Dr FitzGerald is a recipient of a Canadian Institute for Health Research and BC Lung Association Scientist Award and a Michael Smith Distinguished Scholar Award.

The authors thank David Simel, MD, for his valuable guidance during this study. We are also grateful to Robert Badgett, MD, University of Texas Health Science Center at San Antonio; Michael Cuffe, MD, Duke University Health System, and Karen Welty-Wolf, MD, Division of Pulmonary

and Critical Care Medicine, Duke University Medical Center, Durham VA Medical Center, Durham, North Carolina, for their expert advice and helpful reviews of earlier versions of the manuscript.

REFERENCES

- Braunwald E. Shattuck lecture—cardiovascular medicine at the turn of the millennium: triumphs, concerns, and opportunities. *N Engl J Med.* 1997;337(19):1360-1369.
- Hunt SA, Baker DW, Chin MH, et al; American College of Cardiology/American Heart Association Task Force on Practice Guidelines (Committee to Revise the 1995 Guidelines for the Evaluation and Management of Heart Failure); International Society for Heart and Lung Transplantation; Heart Failure Society of America. ACC/AHA guidelines for the evaluation and management of chronic heart failure in the adult: executive summary: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines (Committee to Revise the 1995 Guidelines for the Evaluation and Management of Heart Failure). *Circulation.* 2001; 104(24):2996-3007.
- Redfield MM. Heart failure: an epidemic of uncertain proportions. *N Engl J Med.* 2002;347(18):1442-1444.
- McCullough PA, Philbin EF, Spertus JA, Kaatz S, Sandberg KR, Weaver WD. Confirmation of a heart failure epidemic: findings from the Resource Utilization Among Congestive Heart Failure (REACH) study. *J Am Coll Cardiol.* 2002;39(1):60-69.
- Remme WJ, Swedberg K. Guidelines for the diagnosis and treatment of chronic heart failure. *Eur Heart J.* 2001;22(17):1527-1560.
- Mueller C, Scholer A, Laule-Kilian K, et al. Use of B-type natriuretic peptide in the evaluation and management of acute dyspnea. *N Engl J Med.* 2004;350(7):647-654.
- Rich MW, Nease RF. Cost-effectiveness analysis in clinical practice: the case of heart failure. *Arch Intern Med.* 1999;159(15):1690-1700.
- Bales AC, Sorrentino MJ. Causes of congestive heart failure: prompt diagnosis may affect prognosis. *Postgrad Med.* 1997;101(1):44-46.
- Wuerz RC, Meador SA. Effects of prehospital medications on mortality and length of stay in congestive heart failure. *Ann Emerg Med.* 1992;21(6):669-674.
- Manning HL, Schwartzstein RM. Pathophysiology of dyspnea. *N Engl J Med.* 1995;333(23):1547-1553.
- Wasserman K, Casaburi R. Dyspnea: physiological and pathophysiological mechanisms. *Annu Rev Med.* 1988;39:503-515.
- Springfield CL, Sebat F, Johnson D, Lengle S, Sebat C. Utility of impedance cardiography to determine cardiac vs. noncardiac cause of dyspnea in the emergency department. *Congest Heart Fail.* 2004;10(2 suppl 2):14-16.
- Murray JF. History and physical examination. Murray JF, Nadel JA, eds. *Textbook of Respiratory Medicine.* 3rd ed. Philadelphia, PA: WB Saunders Co; 2000:587.
- Braunwald E. Heart failure. In: Braunwald E, Fauci AS, Kasper DL, et al, eds. *Harrison's Principles of Internal Medicine.* 15th ed. New York, NY: McGraw-Hill; 2001:1318-1329.
- Dyspnea: mechanisms, assessment, and management: a consensus statement: American Thoracic Society. *Am J Respir Crit Care Med.* 1999;159(1): 321-340.
- Braunwald E. Pathophysiology of heart failure. In: Braunwald E, ed. *Heart Disease.* Philadelphia, PA: WB Saunders Co; 1992:393.
- Badgett RG, Lucey CR, Mulrow CD. Can the clinical examination diagnose left-sided heart failure in adults? *JAMA.* 1997;277(21):1712-1719.
- Poole-Wilson PA. Relation of pathophysiologic mechanisms to outcome in heart failure. *J Am Coll Cardiol.* 1993;22(4 suppl A):22A-29A.
- Coats AJ, Adamopoulos S, Meyer TE, Conway J, Sleight P. Effects of physical training in chronic heart failure. *Lancet.* 1990;335(8681):63-66.
- Wilson JR, Mancini DM, Dunkman WB. Exertional fatigue due to skeletal muscle dysfunction in patients with heart failure. *Circulation.* 1993;87(2):470-475.
- Mulrow CD, Lucey CR, Farnett LE. Discriminating causes of dyspnea through clinical examination. *J Gen Intern Med.* 1993;8(7):383-392.
- Cowie MR, Jourdain P, Maisel A, et al. Clinical applications of B-type natriuretic peptide (BNP) testing. *Eur Heart J.* 2003;24(19):1710-1718.

23. Baughman KL. B-type natriuretic peptide: a window to the heart. *N Engl J Med.* 2002;347(3):158-159.
24. Holleman DR Jr, Simel DL. Does the clinical examination predict airflow limitation? *JAMA.* 1995;273(4):313-319.
25. Metlay JP, Kapoor WN, Fine MJ. Does this patient have community-acquired pneumonia? diagnosing pneumonia by history and physical examination. *JAMA.* 1997;278(17):1440-1445.
26. Chunilal SD, Eikelboom JW, Attia J, et al. Does this patient have pulmonary embolism? *JAMA.* 2003;290(21):2849-2858.
27. Reeves RA. Does this patient have hypertension? how to measure blood pressure. *JAMA.* 1995;273(15):1211-1218.
28. Cook DJ, Simel DL. Does this patient have abnormal central venous pressure? *JAMA.* 1996;275(8):630-634.
29. Etchells E, Bell C, Robb K. Does this patient have an abnormal systolic murmur? *JAMA.* 1997;277(7):564-571.
30. Choudhry NK, Etchells EE. Does this patient have aortic regurgitation? *JAMA.* 1999;281(23):2231-2238.
31. Marantz PR, Kaplan MC, Alderman MH. Clinical diagnosis of congestive heart failure in patients with acute dyspnea. *Chest.* 1990;97(4):776-781.
32. Ard RW, Twinning RH. Evaluation of Valsalva test in bedside diagnosis of dyspnea. *Am J Med Sci.* 1957;234(4):403-412.
33. Knowles JH, Gorlin R, Storey CF. Clinical test for pulmonary congestion with use of the Valsalva maneuver. *JAMA.* 1956;160(1):44-48.
34. Engstrom JW, Martin JB. Disorders of the autonomic nervous system. In: Braunwald E, Fauci AS, Kasper DL, et al, eds. *Harrison's Principles and Practice of Internal Medicine.* 15th ed. New York, NY: McGraw-Hill Publishers; 2001:2416-2421.
35. Lederle FA, Simel DL. Does this patient have abdominal aortic aneurysm? *JAMA.* 1999;281(1):77-82.
36. Morrison LK, Harrison A, Krishnasamy P, Kazanegra R, Clopton P, Maisel A. Utility of a rapid B-natriuretic peptide assay in differentiating congestive heart failure from lung disease in patients presenting with dyspnea. *J Am Coll Cardiol.* 2002;39(2):202-209.
37. Sackett DL, Haynes RB, Guyatt GH, Tugwell P. *Clinical Epidemiology: A Basic Science for Clinical Medicine.* 2nd ed. Boston, MA: Little Brown; 1991:52.
38. Rao CR. *Linear Statistical Inference and Its Applications.* New York, NY: John Wiley & Sons; 1968.
39. Maisel AS, Krishnasamy P, Nowak RM, et al. Rapid measurement of B-type natriuretic peptide in the emergency diagnosis of heart failure. *N Engl J Med.* 2002;347(3):161-167.
40. McCullough PA, Nowak RM, McCord J, et al. B-type natriuretic peptide and clinical judgment in emergency diagnosis of heart failure: analysis from Breathing Not Properly (BNP) Multinational Study. *Circulation.* 2002;106(4):416-422.
41. Knudsen CW, Omland T, Clopton P, et al. Diagnostic value of B-type natriuretic peptide and chest radiographic findings in patients with acute dyspnea. *Am J Med.* 2004;116(6):363-368.
42. Lainchbury JG, Campbell E, Frampton CM, Yandle TG, Nicholls MG, Richards AM. Brain natriuretic peptide and N-terminal brain natriuretic peptide in the diagnosis of heart failure in patients with acute shortness of breath. *J Am Coll Cardiol.* 2003;42(4):728-735.
43. Logeart D, Saudubray C, Beyne P, et al. Comparative value of Doppler echocardiography and B-type natriuretic peptide assay in the etiologic diagnosis of acute dyspnea. *J Am Coll Cardiol.* 2002;40(10):1794-1800.
44. Knudsen CW, Riis JS, Finsen AV, et al. Diagnostic value of a rapid test for B-type natriuretic peptide in patients presenting with acute dyspnoea: effect of age and gender. *Eur J Heart Fail.* 2004;6(1):55-62.
45. Bayes-Genis A, Santalo-Bel M, Zapico-Muniz E, et al. N-terminal pro-brain natriuretic peptide (NT-proBNP) in the emergency diagnosis and in-hospital monitoring of patients with dyspnoea and ventricular dysfunction. *Eur J Heart Fail.* 2004;6(3):301-308.
46. Villacorta H, Duarte A, Duarte NM, et al. The role of B-type natriuretic peptide in the diagnosis of congestive heart failure in patients presenting to an emergency department with dyspnea. *Arq Bras Cardiol.* 2002;79(6): 569-572.
47. Davis M, Espiner E, Richards G, et al. Plasma brain natriuretic peptide in assessment of acute dyspnoea. *Lancet.* 1994;343(8895):440-444.
48. Dao Q, Krishnasamy P, Kazanegra R, et al. Utility of B-type natriuretic peptide in the diagnosis of congestive heart failure in an urgent-care setting. *J Am Coll Cardiol.* 2001;37(2):379-385.
49. Brown LH, Gough JE, Seim RH. Can quantitative capnometry differentiate between cardiac and obstructive causes of respiratory distress? *Chest.* 1998;113(2):323-326.
50. McNamara RM, Cionni DJ. Utility of the peak expiratory flow rate in the differentiation of acute dyspnea: cardiac vs pulmonary origin. *Chest.* 1992;101(1):129-132.
51. Fleischer D, Espiner EA, Yandle TG, et al. Rapid assay of plasma brain natriuretic peptide in the assessment of acute dyspnoea. *N Z Med J.* 1997;110(1039):71-74.
52. McCullough PA, Hollander JE, Nowak RM, et al. Uncovering heart failure in patients with a history of pulmonary disease: rationale for the early use of B-type natriuretic peptide in the emergency department. *Acad Emerg Med.* 2003;10(3):198-204.
53. Januzzi JL Jr, Camargo CA, Anwaruddin A, et al. The N-terminal Pro-BNP Investigation of Dyspnea in the Emergency Department (PRIDE) study. *Am J Cardiol.* 2005;95(8):948-954.
54. Alibay Y, Beauchet A, El Mahmoud R, et al. Plasma N-terminal pro-brain natriuretic peptide and brain natriuretic peptide in assessment of acute dyspnea. *Biomed Pharmacother.* 2005;59(1-2):20-24.
55. Ray P, Arthaud M, Lefort Y, Birolleau S, Beigelman C, Riou B. Usefulness of B-type natriuretic peptide in elderly patients with acute dyspnea. *Intensive Care Med.* 2004;30(12):2230-2236.
56. Mueller T, Gegenhuber A, Poelz W, Haltmayer M. Diagnostic accuracy of B type natriuretic peptide and amino terminal proBNP in the emergency diagnosis of heart failure. *Heart.* 2005;91(5):606-612.
57. Gadsboll N, Hoilund-Carlsen PF, Nielsen GG, et al. Symptoms and signs of heart failure in patients with myocardial infarction: reproducibility and relationship to chest X-ray, radionuclide ventriculography and right heart catheterization. *Eur Heart J.* 1989;10(11):1017-1028.
58. Ishmail AA, Wing S, Ferguson J, Hutchinson TA, Magder S, Flegel KM. Interobserver agreement by auscultation in the presence of a third heart sound in patients with congestive heart failure. *Chest.* 1987;91(6):870-873.
59. Cook DJ. Clinical assessment of central venous pressure in the critically ill. *Am J Med Sci.* 1990;299(3):175-178.
60. Butman SM, Ewy GA, Standen JR, Kern KB, Hahn E. Bedside cardiovascular examination in patients with severe chronic heart failure: importance of rest or inducible jugular venous distension. *J Am Coll Cardiol.* 1993;22(4):968-974.
61. Badgett RG, Mulrow CD, Otto PM, Ramirez G. How well can the chest radiograph diagnose left ventricular dysfunction? *J Gen Intern Med.* 1996;11(10):625-634.
62. Panju AA, Hemmelgarn BR, Guyatt GH, Simel DL. Is this patient having a myocardial infarction? *JAMA.* 1998;280(14):1256-1263.
63. Holleman DR Jr, Simel DL, Goldberg JS. Diagnosis of obstructive airways disease from the clinical examination. *J Gen Intern Med.* 1993;8(2):63-68.
64. McCullough PA, Duc P, Omland T, et al. B-type natriuretic peptide and renal function in the diagnosis of heart failure: an analysis from the Breathing Not Properly Multinational Study. *Am J Kidney Dis.* 2003;41(3):571-579.
65. Cabanes L, Richaud-Thiriez B, Fulla Y, et al. Brain natriuretic peptide blood levels in the differential diagnosis of dyspnea. *Chest.* 2001;120(6):2047-2050.
66. Seward JB, Douglas PS, Erbel R, et al. Hand-carried cardiac ultrasound (HCU) device: recommendations regarding new technology: a report from the Echocardiography Task Force on New Technology of the Nomenclature and Standards Committee of the American Society of Echocardiography. *J Am Soc Echocardiogr.* 2002;15(4):369-373.
67. Devereux RB, Liebson PR, Horan MJ. Recommendations concerning use of echocardiography in hypertension and general population research. *Hypertension.* 1987;9(2 pt 2):II97-II104.
68. Wenger NK. Left ventricular dysfunction, exercise capacity and activity recommendations. *Eur Heart J.* 1988;9(suppl F):S63-S66.
69. Marcus GM, Gerber IL, McKeown BH, et al. Association between phonocardiographic third and fourth heart sounds and objective measures of left ventricular function. *JAMA.* 2005;293(18):2238-2244.
70. Maisel A. B-type natriuretic peptide measurements in diagnosing congestive heart failure in the dyspeptic emergency department patient. *Rev Cardiovasc Med.* 2002;3(suppl 4):S10-S17.

This page intentionally left blank

UPDATE : Congestive Heart Failure

Prepared by Robert G. Badgett, MD,
and Catherine R. Lucey, MD

Reviewed by Najib Ayas, MD, and Sheri Keitz, MD

CLINICAL SCENARIO

A 68-year-old patient with a history of smoking, dyslipidemia, and claudication presents for a follow-up appointment. He has no cardiac symptoms. Although he has no history of a myocardial infarction, he has anterior Q waves on an electrocardiogram (ECG). You believe him to be at high risk for asymptomatic left ventricular dysfunction and that a change in his medications might decrease the risk for symptomatic heart failure. Does he have a reduced ejection fraction?

UPDATED SUMMARY ON LEFT VENTRICULAR DYSFUNCTION

Original Reviews

Badgett RG, Lucey CR, Mulrow CD. Can the clinical examination diagnose left-sided heart failure in adults? *JAMA*. 1997;277(21):1712-1719.

UPDATED REVIEW

Wang CS, FitzGerald JM, Schulzer M, Mak E, Ayas NT. Does this dyspneic patient in the emergency department have congestive heart failure? *JAMA*. 2005;294(15):1944-1956.

UPDATED LITERATURE SEARCH

Our literature search used the search strategy for The Rational Clinical Examination articles in OVID MEDLINE, combined with the terms “exp congestive heart failure/” or “heart failure.tw.” or “exp ventricular dysfunction/” or “ventricular dysfunction.tw.” limited to original human and English-language diagnostic articles published from January 1997 to November 2004. Though the original publication included studies of the clinical evaluation of elevated left ventricular diastolic pressure (as measured by the pulmonary capillary wedge pressure), the update focused only on identifying patients with a low ejection fraction (EF) (systolic dysfunction) or distinguishing patients with systolic dysfunction from those with diastolic dysfunction. We included studies that evaluated outpatients or inpatients while excluding studies of emergency department patients with acute dyspnea. The patient with dyspnea who

presents to the emergency department is reviewed in the recent Rational Clinical Examination article noted above.

We reviewed 1005 citations and retrieved 15 promising citations to see whether they had sensitivity, specificity, likelihood ratio (LR) data, or a multivariable analysis of the clinical examination for systolic or diastolic dysfunction compared with a reference standard test. Of 9 articles with appropriate data, only 5 were of new original data and of high enough quality for inclusion.

NEW FINDINGS

- In all settings, new studies confirm that symptoms, signs, and risk factors for left ventricular systolic dysfunction should be interpreted with electrocardiogram (ECG) and chest radiograph results.
- After a myocardial infarction, the clinical evaluation cannot rule out systolic dysfunction and all of these patients should have their EF measured. The presence of a third heart sound, rales, pulmonary venous congestion on chest radiograph, or an anterior Q wave identifies patients most likely to have an EF of 40%.
- For nonemergency outpatients, the absence of clinical findings (orthopnea, paroxysmal nocturnal dyspnea, rales, third heart sound, jugular venous distention), chest radiograph findings, and ECG abnormalities makes the diagnosis of systolic dysfunction unlikely. When these normal results are combined with a brain natriuretic peptide (BNP) level lower than 37 pg/mL, systolic dysfunction is even less likely.
- Because of verification bias in existing studies, the presence of increasing numbers of symptoms and signs may be better than previously thought for identifying patients with an increasing likelihood of systolic dysfunction. The addition of a BNP assay to increasing numbers of abnormal symptoms and signs does not add much to the clinical evaluation.

Details of the Update

Since our initial review in 1997, the goals have changed for the clinical evaluation of patients with heart failure. Recent US and European guidelines emphasize the use of symptoms alone to titrate many medicines for heart failure. However, initiation of optimal pharmacologic management to minimize morbidity and extend life is predicated on identifying patients with a low EF.

Diagnosis of Systolic Dysfunction

In our 1997 review, we found that the clinical evaluation had a positive likelihood ratio ($LR+$) of 2.5 and a negative likelihood ratio ($LR-$) of 0.2 for detecting decreased EF. Three new studies highlight the clinical considerations and methodologic difficulties when studying the role of the clinical examination.

An important methodologic problem presents itself with the selection of patients and how the clinical examination is reported. Most studies accept that individual physical examination findings (eg, rales, a third heart sound, or jugular venous distention) are used best in combination. Thus, studies typically evaluate the overall clinical diagnosis of heart failure, allowing the clinician to consider all the patient's symptoms and signs, along with ECG and chest radiograph abnormalities. Alternatively, studies may evaluate explicit clinical criteria (from a list of findings) or the performance of a multivariable model derived from the clinical data. Another problem is the varying choice in the cut point for combining findings; some approaches seek to maximize accuracy, and other approaches optimize the $LR-$ to ensure that few patients with left ventricular dysfunction escape detection.

Postmyocardial Infarction

One new study evaluated patients after a myocardial infarction.¹ The study had distinctly different results when 2 US centers, where the clinical findings were recorded on admission to a cardiac care unit and then twice daily (incidence of left ventricular systolic dysfunction, 39%), were compared with a Scottish center, where the clinical assessment was done just once in the morning after admission (incidence of systolic dysfunction, 60%). A clinical diagnosis of systolic dysfunction after myocardial infarction should be based on twice-daily examinations because the evaluation made the morning after admission to the hospital had no diagnostic value ($LR+$, 0.86; $LR-$, 1.0). On the other hand, twice-daily assessments using predefined criteria for heart failure had an $LR+$ of 3.1 (95% confidence interval [CI], 1.7-5.8) and an $LR-$ of 0.62 (95% CI, 0.46-0.83). The clinical criteria were pulmonary rales one-third up the lung fields in the absence of chronic lung disease, a third heart sound, or radiographic evidence of pulmonary venous congestion. Anterior Q waves do not occur in most infarctions, but their presence is important (LR for anterior Q waves, 5.0; 95% CI, 2.2-11). The inability to "rule out" left ventricular systolic dysfunction by clinical diagnosis alone supports the role of routine objective assessment of the EF after myocardial infarction. However, when an examination that does not support heart failure is combined with the absence of anterior Q waves, no previous myocardial infarction, and a peak creatinine kinase level less than 1000 U/L in the absence of thrombolytic therapy, the LR for systolic dysfunction decreases to 0.11 (95% CI, 0.04-0.29).

Patients Referred for Echocardiograms

Two studies evaluated consecutive patients referred to an echocardiography laboratory solely for the determination of the systolic EF. One study² assessed the clinical diagnosis in inpatients, whereas the other study assessed referred outpatients.³ The patients in these studies represent those who might be referred

by general internists, but they do not present the entire spectrum because the study population excludes the patients for whom the physician used the clinical findings and medical history to determine that an echocardiogram was not needed. This has 2 important implications for clinicians. First, the results of these studies provide an incomplete picture of how patients should be prospectively identified for referral to echocardiography. However, the results could be used to help decide the likelihood of an abnormal result once the physician refers the patient for an objective assessment of systolic function. For example, an echocardiographer might use the information to identify patients for whom the likelihood of systolic dysfunction is so low that an echocardiogram could be deferred.

Second, these studies are affected by verification bias in which not all patients who were considered as possibly having systolic dysfunction were evaluated. Typically, verification bias leads to an underestimation of the $LR+$ (ie, the clinical findings are actually much better at detecting affected patients than the investigators report) while overestimating the efficiency of the $LR-$ (ie, the clinical findings are not as good at identifying the patients with a normal EF result).

The first study was conducted on inpatients who had a 41% prevalence of systolic dysfunction. The overall clinical diagnosis for inpatients was based on a combination of clinical findings, the ECG, and the chest radiograph. Using all these features, a clinical diagnosis of heart failure had an LR of 2.0 (95% CI, 1.6-2.5) for systolic dysfunction vs an LR of 0.41 (95% CI, 0.30-0.56) when the clinician believed the patient did not have heart failure. Although the physical examination findings of rales, a third heart sound, and jugular venous distention were useful, a multivariate model showed that none of them had independent utility when the patient's sex, chest radiograph, and ECG were considered. Once clinicians have decided to refer a patient for measurement of the systolic EF, they should be aware that those with a normal ECG result are unlikely to have an EF lower than 45% (LR , 0.03; 95% CI, 0.01-0.10).

A second study evaluated consecutive outpatients referred specifically for assessment of systolic function. The prevalence of systolic dysfunction was 11%. The study combined symptomatic patients with asymptomatic patients who had risk factors. According to the characteristics of the referred patients, the investigators established a clinical score that required the presence of at least 1 abnormality: a history of myocardial infarction, previous diagnosis of heart failure, orthopnea or paroxysmal nocturnal dyspnea, Q-wave or intraventricular conduction delay, or a chest radiograph abnormality (cardiomegaly, pulmonary venous hypertension, or edema). The presence of at least 1 abnormality ($LR+$, 2.5; 95% CI, 2.2-2.9) worked as well as the clinical diagnosis for identifying hospitalized patients with an EF of less than 45%. Patients with none of the findings were unlikely to have a low EF (LR , 0.09; 95% CI, 0.03-0.28). The authors also assessed the effect of a normal BNP at cut point of 37 pg/mL (the mean value in healthy persons). At this threshold, the BNP performed similarly to the clinical score and was not independently useful. A BNP of 37 pg/mL or higher combined with an abnormal clinical score increases the likelihood of a low EF (LR , 3.9; 95% CI, 3.0-5.0). The only important utility of the BNP at this cut

point, when combined with the clinical score, was in identifying patients unlikely to have an abnormal EF result. A normal clinical score and a normal BNP level decrease the likelihood of systolic dysfunction, with an LR of 0.04 (95% CI, 0.01-0.30). Patients with a normal score still had a decreased likelihood of a low EF even when the BNP level was elevated (LR, 0.23; 95% CI, 0.06-0.92). The investigators did not assess the utility of the BNP at various cut points, but a higher BNP threshold would improve the LR+ at the expense of the LR-, making “normality” more difficult to identify.

Distinguishing Diastolic From Systolic Dysfunction Among Patients With Known Heart Failure

Two new studies support our original recommendation that the presence of hypertension has some utility in distinguishing heart failure patients with systolic dysfunction from those with a normal EF result (diastolic dysfunction). However, none of the individual findings either increases the likelihood of diastolic dysfunction more than 2-fold or decreases the likelihood of systolic dysfunction by more than one-half. Female sex and hypertension (systolic blood pressure >160 mm Hg) make diastolic dysfunction more likely, whereas tachycardia (heart rate > 100/min) and left atrial abnormality on the ECG make systolic dysfunction more likely.⁴ Unfortunately, multivariate assessment with a large number of candidate variables creates a complicated regression model (18 variables) that does not appreciably improve on these few individual findings when validated independently.⁵

IMPROVEMENTS IN THE DATA PRESENTED IN THE ORIGINAL PUBLICATION

The results for the overall clinical evaluation for systolic dysfunction reported in this update should supplant the summary estimates given in Table 16-6 of our original review. Although we showed summary measures for the performance of the clinical examination in the postinfarction patient, the reported data came from studies with various enrollment methods, timing of the EF assessment, and variable thresholds accepted as indicating systolic dysfunction. The data presented in this update confirm the utility of a clinical diagnosis of heart failure but highlight the inability of the clinical evaluation to confirm that the postinfarction patient has a normal EF result. For patients electively referred for echocardiography, temporal changes in management of the heart failure patient coupled with an increasing awareness of the limits of clinical diagnosis have likely changed the spectrum of patients undergoing determination of systolic function. Although the results of this update do not yield dramatically different LRs, the results are nonetheless more contemporary and applicable to the current care of heart failure patients.

CHANGES IN THE REFERENCE STANDARD

No changes in measurement of the EF of left ventricular filling pressure have been advocated. However, 2 recent studies suggest the EF does not identify early ventricular dysfunction and that

changes in the reference standard may be needed. First, in an analysis of the Framingham subjects without heart failure, echocardiographic evidence of increased left ventricular volume, independent of the fractional shortening, predicted subsequent clinical episodes of heart failure.⁶ Likewise, the BNP was found to improve on the EF and clinical signs of heart failure in predicting mortality among patients with coronary disease.⁷

RESULTS OF LITERATURE REVIEW

See Tables 16-10 and 16-11.

EVIDENCE FROM GUIDELINES

The American College of Cardiology advocates that clinicians consider heart failure a syndrome that progresses from an asymptomatic state among patients with risk factors to symptoms of heart failure. Patients with symptoms should receive a physical examination, chest radiography,

Table 16-10 Diagnosing Left Ventricular Dysfunction

Setting (EF)	LR+ (95% CI)	LR- (95% CI)
Clinical diagnosis ^a After MI (EF ≤ 40%)	3.1 (1.7-5.8)	0.62 (0.46-0.83)
Clinical diagnosis Inpatients (EF < 45%)	2.0 (1.6-2.5)	0.41 (0.30-0.56)
Clinical score ≥ 1 ^b Outpatients (EF < 45%)	2.5 (2.2-2.9)	0.09 (0.03-0.28)
Clinical score Outpatients (EF < 45%) and BNP		
Score ≥ 1 + BNP > 37 pg/mL	3.9 (3.0-5.0)	
Score ≥ 1 + normal BNP	1.1 (0.61-2.0)	
Score < 1 + BNP > 37 pg/mL	0.23 (0.06-0.92)	
Score < 1 + normal BNP	0.04 (0.01-0.30)	

Abbreviations: BNP, brain natriuretic peptide; CI, confidence interval; EF, ejection fraction; LR+, positive likelihood ratio; LR-, negative likelihood ratio; MI, myocardial infarction.

^aTwice-daily examinations after an MI. Clinical diagnosis “positive” required radiographic pulmonary venous congestion with edema, rales one-third up the lung fields in the absence of chronic pulmonary disease, or a third heart sound.

^bScore of 1 or more when any of the following 5 findings are present: a history of myocardial infarction, previous diagnosis of heart failure, orthopnea or paroxysmal nocturnal dyspnea, Q-wave or intraventricular conduction delay, or a chest radiograph abnormality (cardiomegaly, pulmonary venous hypertension, or edema).

Table 16-11 Distinguishing Diastolic Dysfunction From Systolic Dysfunction

Finding	LR for Diastolic Dysfunction	LR for Systolic Dysfunction (EF < 45%)
Favors Normal Systolic Function		
Female sex	1.6 (1.2-2.2)	0.62 (0.46-0.84)
Systolic blood pressure ≥ 160 mm Hg	1.8 (1.3-2.6)	0.55 (0.39-0.78)
Favors Systolic Dysfunction		
Heart rate ≥ 100/min	0.43 (0.28-0.65)	2.3 (1.5-3.5)
Left atrial ECG abnormality	0.42 (0.26-0.63)	2.4 (1.6-3.6)

Abbreviations: ECG, electrocardiogram; EF, ejection fraction; LR, likelihood ratio.

and electrocardiography and have ongoing clinical assessment for volume status (level of evidence: C). However, the echocardiogram is the single most useful test and is required for identifying patients with systolic dysfunction.⁸

The European Society of Cardiology suggests a slightly different approach, advocating the addition of BNP testing to medical history, physical examination, ECG, and the chest radiograph.⁹ When any of these results are abnormal, echocardiography is recommended. Although they acknowledged that heart failure is unlikely with a completely normal ECG result, the Society also notes the poor relationship between symptoms, signs, and the actual EF that makes the echocardiogram a necessary test.

CLINICAL SCENARIO—RESOLUTION

The patient has no symptoms of left ventricular dysfunction, but he does have increased risk of cardiovascular disease, as evidenced by his claudicatory symptoms. The presence of anterior Q waves on the ECG suggests he may have had a silent myocardial infarction, putting him at risk for the evolution of heart failure. Anterior Q waves have a sufficiently high LR for identifying patients with a low EF that an echocardiogram should be obtained.

REFERENCES FOR THE UPDATE

- Darbar D, Gillespie N, Choy AM, et al. Diagnosing left ventricular dysfunction after myocardial infarction: the Dundee algorithm. *Q J Med.* 1997;90(11):677-683.^a
- Talreja D, Gruver C, Sklenar J, Dent J, Kaul S. Efficient utilization of echocardiography for the assessment of left ventricular systolic function. *Am Heart J.* 2000;139(3):394-398.^a
- Yamamoto K, Burnett JC, Bermudez EA, Jougasaki M, Bailey KR, Redfield MM. Clinical criteria and biochemical markers for the detection of systolic dysfunction. *J Card Fail.* 2000;6(3):194-200.^a
- Thomas JT, Kelly RF, Thomas SJ, et al. Utility of history, physical examination, electrocardiogram, and chest radiograph for differentiating normal from decreased systolic function in patients with heart failure. *Am J Med.* 2002;112(6):437-445.^a
- Philbin EF, Hunsberger S, Garg R, et al; Digitalis Investigation Group. Usefulness of clinical information to distinguish patients with normal from those with low ejection fractions in heart failure. *Am J Cardiol.* 2002;89(10):1218-1221.^a
- Vasan RS, Larson MG, Benjamin EJ, Evans JC, Levy D. Left ventricular dilatation and the risk of congestive heart failure in people without myocardial infarction. *N Engl J Med.* 1997;336(19):1350-1355.
- Kraglund C, Gronning B, Kober L, Hildebrandt P, Steffensen R. N-terminal pro-B-type natriuretic peptide and long-term mortality in stable coronary heart disease. *N Engl J Med.* 2005;352(7):666-675.
- Hunt SA, Baker DW, Chin MH, et al. ACC/AHA guidelines for the evaluation and management of chronic heart failure in the adult: executive summary (Committee to Revise the 1995 Guidelines for the Evaluation and Management of Heart Failure). *J Am Coll Cardiol.* 2001;38(7):2101-2113.
- Remme WJ, Swedberg K; Task Force for the Diagnosis and Treatment of Chronic Heart Failure; European Society of Cardiology. Guidelines for the diagnosis and treatment of chronic heart failure. *Eur Heart J.* 2001;22(17):1527-1560.

^aFor the Evidence to Support the Update for this topic, see <http://www.JAMAevidence.com>.

LEFT VENTRICULAR DYSFUNCTION—MAKE THE DIAGNOSIS

PRIOR PROBABILITY

A broad range of prior probabilities (10%-40%) is required for clinical decisions, with outpatients who have suggestive symptoms at the lower end of the range and inpatients toward the upper end. The patient with dyspnea who presents to the emergency department without an obvious cause for dyspnea has about a 50% probability of left ventricular dysfunction (range, 34%-83%).

POPULATION FOR WHOM LEFT VENTRICULAR DYSFUNCTION SHOULD BE CONSIDERED

- Any patient with compatible symptoms, especially orthopnea and paroxysmal nocturnal dyspnea
- Coronary artery disease, especially patients who have had a myocardial infarction
- Hypertension
- Diabetes mellitus
- Patient receiving cardiotoxic medications
- Family history of cardiomyopathy

DETECTING THE LIKELIHOOD OF LEFT VENTRICULAR DYSFUNCTION

Patients with symptoms of heart failure and those with risk factors should be examined for pulmonary rales, jugular venous distention, a third heart sound, and peripheral edema and should have an ECG and chest radiograph (see Table 16-12).

REFERENCE STANDARD TESTS

An objective measure of systolic dysfunction, typically echocardiography but including nuclear cardiology or cardiac catheterization.

Table 16-12 Likelihood Ratios for Diagnosis of Left Ventricular Dysfunction

Medical Inpatients, Including Postmyocardial Infarction	LR+ or Range (95% CI)	LR- or Range (95% CI)
Clinical diagnosis ^a	2.0-3.1	0.41-0.62
ECG abnormal ^b	2.8 (2.3-3.4)	0.03 (0.01-0.10)
Outpatients		
Clinical score ^c with a BNP > 37 pg/mL		
Score ≥ 1 + elevated BNP	3.9 (3.0-5.0)	
Score ≥ 1 + normal BNP	1.1 (0.61-2.0)	
Score < 1 + elevated BNP	0.23 (0.06-0.92)	
Score < 1 + normal BNP	0.04 (0.01-0.30)	
The Breathless Emergency Patient	LR + (95% CI)	LR - (95% CI)
Patient History		
Heart failure	5.8 (4.1-8.0)	0.45 (0.38-0.53)
Myocardial infarction	3.1 (2.0-4.9)	0.69 (0.58-0.82)
Physical Examination		
Third heart sound	11 (4.9-25)	0.88 (0.83-0.94)
Abdominojugular reflux	6.4 (0.81-51)	0.79 (0.62-1.0)
Jugular venous distention	5.1 (3.2-7.9)	0.66 (0.57-0.77)
Rales	2.8 (1.9-4.1)	0.51 (0.37-0.70)
Chest Radiograph		
Pulmonary venous congestion	12 (6.8-21)	0.48 (0.28-0.83)
Interstitial edema	12 (5.2-27)	0.68 (0.54-0.85)
Alveolar edema	6.0 (2.2-16)	0.95 (0.93-0.97)
Cardiomegaly	3.3 (2.4-4.7)	0.33 (0.23-0.48)
Pleural effusion(s)	3.2 (2.4-4.3)	0.81 (0.77-0.85)
Electrocardiogram		
Atrial fibrillation	3.8 (1.7-8.8)	0.79 (0.65-0.96)
New T-wave changes	3.0 (1.7-5.3)	0.83 (0.74-0.92)
Any abnormal finding	2.2 (1.6-3.1)	0.64 (0.47-0.88)
B-Type Natriuretic Peptide, pg/mL^d		
≥250	4.6 (2.6-8.0)	
≥100	2.7 (2.0-3.9)	
≥50	1.7 (1.2-2.6)	
<50	0.06 (0.03-0.12)	
Overall Clinical Impression		
Initial clinical judgment that the patient is in heart failure	4.4 (1.8-10)	0.45 (0.28-0.73)

Abbreviations: BNP, brain natriuretic peptide; CI, confidence interval; ECG, electrocardiogram; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

^aClinical symptoms, physical examination, ECG, and chest radiograph.

^bVentricular hypertrophy, ST-segment or T-wave changes, left bundle-branch block, or a paced rhythm. All other findings were considered "normal," for which the LR- applies.

^cScore of 1 or higher when any of the following 5 findings are present: a history of myocardial infarction, previous diagnosis of heart failure, orthopnea or paroxysmal nocturnal dyspnea, Q-wave or intraventricular conduction delay, or a chest radiograph abnormality (cardiomegaly, pulmonary venous hypertension, or edema).

^dThe likelihood ratios (LRs) represent serial LRs, where the LRs are associated with a series of ordered BNP thresholds rather than just a single BNP threshold. In this table, a patient with a BNP of 110 pg/mL would have an LR of 2.7, whereas a value of 33 pg/mL confers an LR of 0.06.

This page intentionally left blank

EVIDENCE TO SUPPORT THE UPDATE:

16

Congestive Heart Failure

TITLE Diagnosing Left Ventricular Dysfunction After Myocardial Infarction: The Dundee Algorithm.

AUTHORS Darbar D, Gillespie N, Choy AM, et al.

CITATION *Q J Med.* 1997;90(11):677-683.

QUESTION How well can a clinical algorithm identify patients, after an acute myocardial infarction, with a left ventricular ejection fraction of 40% or less?

DESIGN Prospective data collection, but no information on whether the patients were consecutively enrolled.

SETTING A university hospital and its Veterans Affairs affiliate medical center (Nashville, Tennessee) and a Scottish University hospital (Dundee).

PATIENTS Patients were evaluated while hospitalized for acute myocardial infarction. Exclusion criteria included patients who had a history of congestive heart failure, were already taking angiotensin-converting enzyme inhibitors at admission, or underwent primary angioplasty within 24 hours of their current infarction.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

The diagnostic algorithm required at least 1 of the following: (1) clinical findings of heart failure, (2) new infarction with anterior Q waves, or (3) previous infarction, with the current infarction exhibiting a peak creatinine kinase (CK) level of more than 1000 U/L without thrombolysis.

The heart failure clinical diagnosis used the criteria from a previous randomized controlled trial and required any of the following: (1) pulmonary venous congestion with edema on at least 1 chest radiograph, (2) rales extending at least one-third up the lung fields in the absence of chronic pulmonary disease, or (3) a third heart sound with persistent tachycardia.¹

In the US centers, the clinical diagnosis could have been established at admission of the patient or on twice-daily rounds with a cardiology fellow or attending physician. The Scottish center used the clinical assessment by a consultant cardiologist on the morning after admission.

An investigator interpreted the electrocardiograms (ECGs) independent of the clinical diagnosis. However, the clinicians establishing the clinical diagnosis had access to the ECG result and the cardiac enzyme levels.

MAIN OUTCOME MEASURES

The UK site used transthoracic echocardiography only. The US site used transthoracic echocardiography, contrast ventriculography, or radionuclide ventriculography. Left ventricular systolic dysfunction was defined by an ejection fraction (EF) of 40% or less.

MAIN RESULTS

A total of 46 (39%) US patients had left ventricular systolic dysfunction vs 56 (60%) Scottish patients (see Table 16-13).

Table 16-13 Likelihood Ratios for the Overall Algorithm and Its Components

Finding (No. With the Finding Present)	LR+ (95% CI)	LR- (95% CI)
Clinical Diagnosis		
US centers (33)	3.1 (1.7-5.8)	0.62 (0.46-0.83)
UK center (25)	0.86 (0.44-1.7)	1.0 (0.82-1.4)
Anterior Q Wave		
US center (13)	8.6 (2.0-37)	0.78 (0.66-0.92)
UK center (27)	3.9 (1.5-10)	0.66 (0.52-0.84)
Summary	5.0 (2.2-11)	0.74 (0.64-0.85)
CK > 1000 U/L		
US center (52)	0.72 (0.51-1.0)	1.5 (1.0-2.1)
UK center ^a (52)	0.73 (0.51-1.0)	1.5 (0.91-2.5)
Summary	0.72 (0.57-0.92)	1.5 (1.1-2.0)
Diagnostic "Algorithm" ^a Result Positive		
US center (58)	4.1 (2.6-6.4)	0.11 (0.04-0.29)
UK center ^b (57)	2.8 (1.7-4.7)	0.25 (0.14-0.46)

Abbreviations: CI, confidence interval; CK, creatinine kinase; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

^aAny of the following present: (1) clinical findings of heart failure, (2) new infarction has anterior Q waves, or (3) previous infarction and the current infarction exhibited a peak CK of more than 1000 U/L without thrombolysis.

^bData corrected from that originally reported in this study.

CONCLUSIONS

LEVEL OF EVIDENCE Level 3 (uncertainty regarding consecutive enrollment).

STRENGTHS The diagnostic criteria were evaluated in 2 sites and highlighted the potential variability in performance of the clinical examination. However, an important feature was the explicit clinical criteria for a heart failure diagnosis that had been used in previous clinical trials.

LIMITATIONS Although the electrocardiograph and CK criteria were collected independently of the clinical evaluation, the clinicians had access to the ECG result and CK enzyme levels. There was a difference in prevalence of left ventricular systolic dysfunction, but more important, there was a distinct difference in the timing and frequency of the clinical evaluations. The US clinicians had more opportunities to detect heart failure than the Scottish physicians.

Although it seems most likely that the difference in the clinical diagnosis could be attributed to the timing and frequency of clinical assessments, it is not possible to rule out a difference in patients (eg, more anterior Q-wave infarctions in Scotland) or an actual difference in examiners' skills. It does seem clear that the CK criterion cannot be used to identify patients with left ventricular systolic dysfunction. An anterior Q wave might be useful, but the confidence intervals are broad.

The proposed "algorithm" is actually a heart failure diagnosis that required the presence of one of 3 findings. In the US center, the combination of findings was more efficient than the clinical diagnosis for identifying patients less likely to have heart failure. In the Scottish center, the criteria added to the clinical diagnosis resulted in a much more accurate finding. If these data are reliable, then we could conclude that patients undergoing twice-daily clinical assessments (with chest radiographs as part of the criteria) without evidence of heart failure after their first nonanterior Q-wave myocardial infarction are much less likely to have a low EF. The nature of the study design requires that this conclusion undergo validation, but even with validation the clinical performance would have to be much better to forgo echocardiographs when the clinician wants to know whether the EF is 40% or lower.

REFERENCE FOR THE EVIDENCE

1. Acute Infarction Ramipril Efficacy (AIRE) Study Investigators. Effect of ramipril on mortality and morbidity of survivors of acute myocardial infarction with clinical evidence of heart failure. *Lancet*. 1993;342(8875):821-828.

Reviewed by Robert G. Badgett, MD,
and Catherine R. Lucey, MD

TITLE Usefulness of Clinical Information to Distinguish Patients With Normal From Those With Low Ejection Fractions in Heart Failure.

AUTHORS Philbin EF, Hunsberger S, Garg R, et al; for the Digitalis Investigation Group.

CITATION *Am J Cardiol*. 2002;89(10):1218-1221.

QUESTION Can the clinical evaluation diagnose patients with diastolic dysfunction among patient with chronic heart failure?

DESIGN Prospective with random allocation to derivation and validation groups.

SETTING A total of 302 clinical centers in the United States and Canada participating in the Digitalis Investigation Group Trial.

PATIENTS A total of 7534 patients with stable symptoms caused by ischemic, hypertensive, idiopathic, or alcohol-related chronic heart failure who were in sinus rhythm. Ten percent of patients had missing values for any predictor variables and were excluded from multivariate analyses. Patients were randomly assigned to either a derivation group or validation group.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

Candidate findings were from the medical history, physical examination, and radiographic evaluation. Ejection fraction (EF) was measured with angiography, radionuclide ventriculography, or 2-dimensional echocardiography.

MAIN OUTCOME MEASURES

A multivariate model was used to calculate the predicted EF in a derivation set of patients ($n = 3768$). The model was then applied to a separate validation set of patients ($n = 3766$) to estimate the accuracy of the prediction and the ability to detect patients with a normal EF ($>45\%$) vs a low EF ($\leq 45\%$).

MAIN RESULTS

Eighteen findings were independently significant, including previously identified findings of female patient, older age, and having a smaller cardiothoracic ratio on chest radiograph. The findings included symptom functional class, rales, jugular venous distention, peripheral edema, a third heart sound, heart rate, and blood pressure, along with several historical features and medical use. Despite the large sample size and large number of variables, the model predicted an EF that was within 5% of the actual result in only 45% of patients.

The investigators evaluated the ability of the model to classify patients correctly at various EF thresholds. Identifying patients with systolic dysfunction improved when a cut point for the predictive model was 30%. In other words, when the model predicted an EF of less than 30%, the likelihood ratio (LR) that the patient has systolic dysfunction ($\leq 45\%$) is 3.9 (95% confidence interval [CI], 2.9-5.3). Although patients with an estimated EF of 30% or higher were less likely to have systolic dysfunction, the condition was not ruled out (LR, 0.65; 95% CI, 0.62-0.69).

CONCLUSIONS

LEVEL OF EVIDENCE Level 1.

STRENGTHS Large study that analyzed many variables to create a predictive model.

LIMITATIONS Electrocardiographic findings were not studied. Sensitivity and specificity of individual findings were not reported.

Some of the findings that entered the final model were those advocated when assessing the patient with heart failure. However, a complicated model with a large number of findings was too inaccurate for clinical prediction of the actual EF. In addition, the model was insufficient for sorting out patients with systolic vs diastolic dysfunction.

Reviewed by Robert G. Badgett, MD,
and Catherine R. Lucey, MD

TITLE Efficient Utilization of Echocardiography for the Assessment of Left Ventricular Systolic Function.

AUTHORS Talreja D, Gruver C, Sklenar J, Dent J, Kaul S.

CITATION *Am Heart J.* 2000;139(3):394-398.

QUESTION Can a combination of clinical findings predict ejection fraction (EF)?

DESIGN Consecutive patients referred for echocardiography. Data were recorded prospectively.

SETTING Inpatient echocardiography laboratory.

PATIENTS A total of 330 inpatients referred to echocardiography specifically for evaluation of left ventricular systolic function. Thirty patients who did not have a required electrocardiogram (ECG) were excluded. The majority (91.5%) of patients had not had an anterior myocardial infarction.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

The physical examination data and clinician's diagnosis of heart failure were collected during a chart review. The chest

radiograph information was abstracted from the radiologist's report, with a focus on the presence of cardiomegaly or vascular congestion. The ECG was interpreted independently of the clinical data. A positive ECG result had to contain a Q wave in 2 or more contiguous leads, poor R-wave progression, left ventricular hypertrophy, ST-segment or T-wave changes, left bundle-branch block, or a paced rhythm.

An echocardiogram, done by an independent observer without access to the clinical data, determined whether there was left ventricular systolic dysfunction that was defined as an EF of less than 0.45.

MAIN OUTCOME MEASURES

Sensitivity and specificity for the overall clinical examination, chest radiograph, and ECG findings.

MAIN RESULTS

One hundred twenty-four (41%) patients had left ventricular systolic dysfunction.

The clinical findings, ECG, and radiograph were used in the decision to refer for echocardiography. However, no individual finding occurred in more than 50% of patients (rales, third heart sound, jugular venous distention, peripheral edema, or a positive overall clinical diagnosis). See Tables 16-14 and 16-15.

The patient's sex (male) was the only other finding that was significant in the logistic model.

$$\text{Systolic dysfunction score} = -127 + 130 \text{ (male patient)} + 80 \text{ (cardiomegaly)} + 190 \text{ (left bundle-branch block)} - 340 \text{ ("normal" ECG result)}$$

Table 16-14 Likelihood Ratio of Overall Clinical Impression Among Patients Referred for Echocardiography

Finding	LR+ (95% CI)	LR- (95% CI)
Overall clinical impression	2.0 (1.6-2.5)	0.41 (0.30-0.56)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

Table 16-15 Useful Predictors for a Low Ejection Fraction Identified in a Multivariable Model

Finding	LR+ (95% CI)	LR- (95% CI)
Radiographic cardiomegaly	1.9 (1.5-2.6)	0.64 (0.52-0.79)
Left bundle-branch block	6.7 (2.3-19)	0.87 (0.80-0.94)
ECG result abnormal ^a	2.8 (2.3-3.4)	0.03 (0.01-0.10)

Abbreviations: CI, confidence interval; ECG, electrocardiogram; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

^aVentricular hypertrophy, ST-segment or T-wave changes, left bundle-branch block, or a paced rhythm was considered abnormal, so that the LR+ applies to these patients. Absence of all of these findings was considered as a "normal" result so that the LR- applies to these patients.

(If finding is present, substitute 1; if absent, substitute 0)

$$\text{Estimated probability} = [\exp^{(\text{score}/100)}]/[1 + \exp^{(\text{score}/100)}]$$

CONCLUSION

LEVEL OF EVIDENCE

Level 3.

STRENGTHS Consecutive patients referred for an echocardiogram to measure the EF.

LIMITATIONS Verification bias exists in that the clinical findings (none of which was independently important) were used to select patients for echocardiography. This tends to overestimate sensitivity and underestimate specificity. Because the individual clinical findings were not collected in a structured way, we did not have confidence in their reliability and therefore did not include them in the evidence table. The systolic dysfunction score (calculated from the logistic model) *cannot* be used to identify patients who should be appropriately referred for echocardiography because it was derived from a group of patients whose physicians had already decided to determine the EF. Despite the data collection method, it is striking that rales, third heart sounds, jugular venous distention, and peripheral edema had no independent utility among patients referred to the echocardiography laboratory.

These data are most useful to the echocardiographer, who might choose to forgo testing for patients with a low probability of a reduced EF. For example, a man with no cardiomegaly and none of the abnormal ECG findings who is referred for EF determination has only a 3% probability of an EF lower than 0.45 (a woman has a probability of only 1%).

Given the selection bias in how patients were identified for echocardiography, the likelihood ratio (LR) for the clinical diagnosis of heart failure (LR, 2.0) could actually be much higher. This phenomenon might be even more striking for the individual physical examination findings if patients with abnormal results were preferentially referred. When clinicians do not clinically diagnose heart failure, but refer the patient for EF determination, the LR is not likely to be as low as the LR of 0.41 found in this study.

Reviewed by Catherine R. Lucey, MD,
and Robert Badgett, MD

TITLE Utility of History, Physical Examination, ECG, and Chest Radiograph for Differentiating Normal From Decreased Systolic Function in Patients With Heart Failure.

AUTHORS Thomas JT, Kelly RF, Thomas SJ, et al.

CITATION *Am J Med.* 2002;112(6):437-445.

QUESTION Can clinical findings differentiate normal vs decreased systolic left ventricular function in patients with heart failure?

DESIGN Consecutive patients, without primary valvular disease, admitted with the primary diagnosis of congestive heart failure.

SETTING Cook County Hospital, Chicago, Illinois.

PATIENTS A total of 225 patients, of whom 46% had diastolic dysfunction. An additional 43 were excluded because their ejection fraction (EF) was not assessed during echocardiography.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

Clinical findings were extracted by a chart review of the attending physician's notes from hospital day 1 or 2. Some attending physicians may have been aware of a patient's echocardiographic results, making the symptoms and subjective physical examination unsuitable for review. However, the vital signs were not biased by awareness of the systolic function. The electrocardiogram (ECG) and chest radiographs were collected and interpreted without awareness of the echocardiogram.

Left ventricular systolic function was determined by echocardiography performed by experienced cardiologists blinded to clinical findings. Normal systolic function was defined as an EF of more than 45% as assessed by visual inspection.

MAIN OUTCOME MEASURES

Sensitivity and specificity from univariate analysis and adjusted odds ratios from multivariate analysis.

MAIN RESULTS

One hundred four patients had normal systolic function and 121 had systolic dysfunction (EF < 45%). A multivariate analysis of 34 findings identified only a few significant variables from the clinical examination, none of which was physical examination findings other than vital signs. See Table 16-16.

The chest radiograph findings (cardiomegaly, cephalization, pulmonary edema, pleural effusion) did not distinguish between individuals with normal systolic function and with systolic dysfunction.

Table 16-16 Useful Predictors for Systolic Dysfunction Identified in a Multivariable Model

Finding	LR for Normal Systolic Function (95% CI)	LR for Systolic Dysfunction, EF < 45% (95% CI)
Favors Normal Systolic Function		
Female sex	1.6 (1.2-2.2)	0.62 (0.46-0.84)
Systolic blood pressure ≥ 160 mm Hg	1.8 (1.3-2.6)	0.55 (0.39-0.78)
Favors Systolic Dysfunction		
Heart rate ≥ 100/min	0.43 (0.28-0.65)	2.3 (1.5-3.5)
Left atrial ECG abnormality	0.42 (0.26-0.63)	2.4 (1.6-3.6)

Abbreviations: CI, confidence interval; ECG, electrocardiogram; EF, ejection fraction; LR, likelihood ratio.

CONCLUSIONS

LEVEL OF EVIDENCE Level 1.

STRENGTHS Large series of consecutive patients admitted with heart failure.

LIMITATIONS Many of the symptoms and some of the physical findings could have been affected by expectation bias, created when the physician was aware of the diagnosis. Despite expectation bias, none of the clinical findings (jugular venous distention, rales, third and fourth heart sounds) or the physician's interpretation of the patient's symptoms were independently useful for distinguishing heart failure patients with normal systolic function from those with systolic dysfunction.

These results support the clinical need for an objective measure of systolic function to distinguish diastolic dysfunction (a normal or elevated EF) from systolic dysfunction. The few variables with independent significance for identifying diastolic dysfunction did not have likelihood ratios so different from 1 that they would obviate the need for echocardiography in the patient with heart failure. Hypertension approximately doubles the likelihood of a normal EF, whereas tachycardia and left atrial ECG abnormalities approximately double the likelihood of systolic dysfunction.

Reviewed by Catherine R. Lucey, MD,
and Robert Badgett, MD

TITLE Clinical Criteria and Biochemical Markers for the Detection of Systolic Dysfunction.

AUTHORS Yamamoto K, Burnett JC, Bermudez EA, Jougasaki M, Bailey KR, Redfield MM.

CITATION *J Card Fail.* 2000;6(3):194-200.

QUESTION Does a clinical score with items from the medical history, electrocardiogram (ECG), and chest radiograph predict left ventricular systolic dysfunction better when added to the information from a brain natriuretic peptide (BNP) assay?

DESIGN Prospective, consecutive patients referred for echocardiography. The echocardiographers were blinded to the study questions.

SETTING University hospital echocardiography laboratory.

PATIENTS Four hundred sixty-six consecutive outpatients referred for echocardiography who were classified further as either having symptoms of heart failure or risk factors for systolic dysfunction. Patients with known systolic dysfunction, or those referred for characterization of a murmur in the absence of any cardiac symptoms, were excluded.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

A score assigned to each patient referred for echocardiography was based on 5 possible risk factors (maximum score = 5): (1) history of myocardial infarction, (2) previous diagnosis of congestive heart failure, (3) current orthopnea or paroxysmal nocturnal dyspnea, (4) presence of pathologic Q waves or an intraventricular conduction defect on ECG, or (5) cardiomegaly, pulmonary venous hypertension, or interstitial edema on chest radiograph. An abnormal score was 1 or higher.

The ECG and chest radiograph results were obtained from the clinical record.

A BNP level higher than 37 pg/mL was defined prospectively as abnormal.

An investigator who extracted the clinical findings, but who was blinded to the results of the BNP and echocardiograph results, reviewed the clinical record.

Left ventricular systolic dysfunction was defined by an ejection fraction (EF) of less than 45%. The echocardiographers were unaware of the study questions.

MAIN OUTCOME MEASURES

Sensitivity and specificity of the clinical score and BNP alone, vs in combination. Likelihood ratios were calculated from data provided in the article.

MAIN RESULTS

Of 466 patients, 201 had an abnormal score of 1 or higher. The prevalence of heart failure symptoms was 33%, but only 11% of all patients had an EF less than 45%. See Tables 16-17 and 16-18.

Table 16-17 Likelihood Ratios for Clinical Score and Brain Natriuretic Peptide

Test	LR+ (95% CI)	LR- (95% CI)
Clinical score $\geq 1^a$	2.5 (2.2-2.9)	0.09 (0.03-0.28)
BNP assay ($> 37 \text{ pg/mL}$)	2.2 (1.8-2.6)	0.34 (0.19-0.54)

Abbreviations: BNP, brain natriuretic peptide; CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

^aClinical score was number for: a history of myocardial infarction, previous diagnosis of heart failure, orthopnea or paroxysmal nocturnal dyspnea, Q-wave or intraventricular conduction delay on an electrocardiogram, or a chest radiograph with cardiomegaly, pulmonary venous hypertension, or edema.

Table 16-18 Serial Likelihood Ratios for the Combination of Clinical Score and Brain Natriuretic Peptide

Clinical Score ^a	BNP	LR (95% CI)
+	+	3.9 (3.0-5.0)
+	-	1.1 (0.61-2.0)
-	+	0.23 (0.06-0.92)
-	-	0.04 (0.01-0.30)

Abbreviations: BNP, brain natriuretic peptide; CI, confidence interval; LR, likelihood ratio.

^aClinical score was positive for a history of myocardial infarction, previous diagnosis of heart failure, orthopnea or paroxysmal nocturnal dyspnea, Q-wave or intraventricular conduction delay on an electrocardiogram, or a chest radiograph with cardiomegaly, pulmonary venous hypertension, or edema.

CONCLUSIONS

LEVEL OF EVIDENCE

Level 3.

STRENGTHS These consecutively enrolled patients had symptoms suggesting heart failure or else were asymptomatic but had risk factors for a low EF. By including asymptomatic but higher-risk patients, the investigators assembled a population of patients

that reflects those for whom an outpatient physician would appropriately evaluate for left ventricular systolic dysfunction.

LIMITATIONS The clinical data were determined from a chart review, but the person who abstracted the data was blinded appropriately to the BNP and echocardiograph results. Relatively few patients had systolic dysfunction, despite their heart failure symptoms or risk factors. As in other studies conducted in echocardiography laboratories, the investigators evaluated patients through the referral filter imposed by the clinician. Thus, patients with obvious systolic dysfunction or those who were obviously healthy were not enrolled. Although this is a pragmatic approach, the results might differ for physicians who have different clinical thresholds for referring patients for echocardiograms.

The clinical score, by itself, is far superior to the BNP primarily because the clinical score is so efficient at identifying patients with an EF of 45% or higher. Remarkably, the clinical score had no physical examination data (eg, rales, a third heart sound, peripheral edema) and only 1 symptom (orthopnea or paroxysmal nocturnal dyspnea). An abnormal clinical score with a normal BNP result (likelihood ratio, ≈ 1) adds no information to the pretest likelihood.

When patients have a normal clinical score, the probability of a low EF is greatly reduced. At a prior probability of 40% for a low EF (much higher than the prevalence in this study), a normal clinical score decreases the probability to approximately 6%. For clinicians who are inclined to measure the EF at that probability level, a BNP of 37 pg/mL or lower would decrease the probability to 2.5% (clinicians should confirm their laboratory's BNP value for healthy patients). Patients who start with a prior probability of less than 40% would have a diminishingly low likelihood of systolic dysfunction with a normal clinical score and normal BNP result.

Just over one-fourth of patients with a normal score (27%) had an abnormal BNP result, but most of these patients prove to have a normal EF result (70/72). The utility of a BNP in patients with a normal clinical score depends entirely on the pretest probability and the physician's and patient's need for certainty.

Reviewed by Robert G. Badgett, MD,
and Catherine Lucey, MD

CHAPTER 17

Is This Patient Dead, Vegetative, or Severely Neurologically Impaired?

Assessing Outcome for Comatose Survivors of Cardiac Arrest

Christopher M. Booth, MD

Robert H. Boone, MD, MSc

George Tomlinson, PhD

Allan S. Detsky, MD, PhD, FRCPC

CLINICAL SCENARIOS

CASE 1 A 65-year-old man experienced a witnessed ventricular fibrillation cardiac arrest at home 24 hours ago. A neighbor had performed cardiopulmonary resuscitation for 5 minutes until the paramedics arrived and performed successful defibrillation. His electrocardiogram revealed a large anterior myocardial infarction, for which he underwent urgent coronary angioplasty. Although still unresponsive, he withdraws from a painful stimulus and his pupillary and corneal reflexes are present. The family asks you about his chance of meaningful recovery.

CASE 2 A 26-year-old woman presented to the emergency department with severe pleuritic chest pain and dyspnea. While waiting for a computed tomographic scan in the radiology department, she had an asystolic cardiac arrest. The resuscitation lasted 20 minutes, after which she was found to have reactive pupils. Three days later, the family is considering withdrawing care because she is still comatose. On examination, her pupils are now unreactive and she has no motor response or brainstem reflexes. The nurse reports that the patient had myoclonus 12 hours ago.

WHY IS THE CLINICAL EXAMINATION IMPORTANT?

With the development of closed-chest cardiac massage in 1960 and the creation of intensive care units shortly thereafter, it became possible to survive cardiac arrest. Half a century later, cardiovascular disease is the leading cause of death in North America and Europe, accounting for approximately half of all deaths in the United States. At least 225 000 people die annually in the United States from cardiovascular disease before they reach a hospital. Twice as many will have cardiac arrest and attempted resuscitation during hospitalization. Survival rates for prehospital cardiac arrest range from 2% to 33%, and reported inpatient survival rates range between 0% and 29%.^{1,2} Most survivors of cardiac arrest ($\approx 80\%$) are comatose after resuscitation. After trauma and drug overdose, cardiac arrest is now the third most common cause of coma.^{3,4} With increasing public education in basic life support and with the use of automated defibrillators in public places, such as in airports and shopping malls, postcardiac arrest coma has become a common and important clinical syndrome.

With the increased success of resuscitation from cardiac arrest comes a multitude of medical, ethical, and economic questions. Once spontaneous circulation has been restored, recovery is far from certain. Possible outcomes range from complete neurologic recovery to death to the persistent vegetative state. In admitted patients who survive the initial car-

diac arrest, rates of meaningful neurologic recovery range from 10% to 30%.⁵ This uncertainty furthers the emotional distress of a grieving and anxious family. Accordingly, it is important for families and physicians to have an understanding of a patient's chance of meaningful recovery.

Unfortunately, the result of the gold standard test for prognosis in this population can be determined only when the true outcome of each patient is known, rather than at presentation. Recent interest has developed in the potential role of neurophysiologic testing.⁶⁻⁸ A recent systematic review found somatosensory-evoked potentials useful in predicting "wakening" of comatose patients.⁷ Other research suggests that elevated serum levels of neuron-specific enolase may predict poor outcome in comatose survivors of cardiac arrest.⁸ Although these results are promising, it will take some time before the precise operating characteristics of these tests are fully understood and before the technology is widely available in clinical practice.

The physical examination has the potential to be extremely useful in this common clinical scenario because of its universal availability and ease of performance. From a compassionate standpoint, the clinical evaluation yields the first information that is relayed to family members desperate for information. Thus, it is crucial for physicians to understand the precision and accuracy of the clinical examination in determining prognosis in hypoxic-ischemic coma.

Pathophysiology

Unlike traumatic or focal ischemic causes of coma, cardiac arrest presents a global ischemic insult to the brain. The

extent of cerebral damage is largely influenced by the duration of interrupted cerebral blood flow. Accordingly, minimizing both the arrest (no-flow) time and cardiopulmonary resuscitation (low-flow) time is critical. With the return of spontaneous circulation comes a transient period of cerebral hyperemia, which is followed by vasospasm and protracted global and multifocal hypoperfusion. Cerebral oxygen stores and consciousness are lost within 20 seconds of the onset of cardiac arrest, whereas glucose and adenosine triphosphate stores are lost by 5 minutes. A cascade of complex chemical derangements ensues, which leads to neuronal death and culminates in the postcardiac arrest coma.⁹

How to Examine a Comatose Patient

Glasgow Coma Scale

Before 1974, the clinical assessment of coma relied on qualitative, descriptive terminology and the presence or absence of brainstem reflexes. Plum and Posner¹⁰ described the widely used definition of coma as "a state of unarouseable unresponsiveness." In 1974, Teasdale and Jennett¹¹ published the first description of the Glasgow Coma Scale (GCS; **Table 17-1**), which has since been used worldwide as a means of classifying coma. Although originally described for traumatic coma, it is equally applicable to the assessment of nontraumatic coma. This ordinal scale is calculated from the sum of 3 components: motor response, verbal response, and eye opening. In assessment of the motor response, it is important to apply central pain because spinal reflexes may occur with peripheral stimulation and do not represent a true motor response. A painful stimulus may be applied to the supraorbital region (deep pinching of the skin) or the sternum (firm twisting pressure applied with the examiner's knuckles). The minimum GCS score is 3 and maximum is 15.

Physical Examination Maneuvers

In addition to the GCS, various brainstem reflexes are used in the physical examination of comatose patients.^{10,12} The pupillary reflex involves cranial nerves II and III. Shining a penlight into one eye and then the other tests the patient's pupillary light response; the examiner observes the direct and consensual response (constriction of the opposite eye). The corneal reflex involves cranial nerves V and VII. Touching the cornea with a piece of cotton or tissue should cause both eyes to blink. The gag and cough reflexes test cranial nerves IX and X. To elicit a gag, apply a tongue depressor to the posterior pharynx. The soft palate should rise symmetrically. In patients who are intubated, assess the cough (or carinal) reflex by applying deep suction through the endotracheal tube to the carina. The suction will produce a gasp, followed by several rapid coughs.

Vestibular signs are also commonly examined in the comatose patient. The oculocephalic (or "doll's eye") reflex involves observing the patient's eyes during passive rotation of the skull. In a comatose patient with intact midbrain and vestibular reflexes, the eyes will move in a direction opposite to that in which the head is moved. If this reflex is lost, the globes will remain fixed within the head and the eyes will continue to stare in whatever direction the head is pointed. This reflex should not be tested in cases of suspected cervical trauma. Cold water caloric testing

Table 17-1 Glasgow Coma Scale^a

Best Motor Response	
Obeying commands	6
Localizing to pain	5
Withdrawning to pain	4
Abnormal flexion (decorticate)	3
Extensor response (decerebrate)	2
None	1
Best Verbal Response ^b	
Oriented	5
Confused conversation	4
Inappropriate words	3
Incomprehensible sounds	2
None	1
Eye Opening	
Spontaneous	4
To speech	3
To pain	2
None	1

^aThe score for the scale is summed across the 3 components and ranges from 3 to 15. A lower score indicates more severe neurologic deficits. Original Glasgow Coma Scale in Teasdale and Jennett.¹¹

^bIntubated patients cannot be given a score for the verbal component, so their total scores accordingly range from 2 to 10.

(oculovestibular reflex) also tests the vestibular and oculomotor systems. To perform the test, first examine the tympanic membrane to ensure there is no perforation or impacted cerumen. With the head 30 degrees higher than the horizontal, irrigate up to 120 mL of ice cold water into the auditory canal. In the unconscious patient with intact brainstem function, there will be slow tonic deviation of eyes toward the irrigated ear.

It is also important to observe the presence of seizures or myoclonus when examining the comatose patient because some clinicians believe they may be useful in prognosis of comatose survivors of cardiac arrest. Seizures may be generalized or focal. Myoclonus refers to isolated sudden muscular contractions and may be either focal or generalized contractions of axial and limb musculature. In patients with seizures, the physical examination should be repeated after the postictal period.

Finally, mechanically ventilated patients are frequently sedated or paralyzed. Accordingly, when a detailed neurologic examination is performed, it is crucial that these medications be at least temporarily discontinued.

Outcomes of Interest

The neurologic outcome of comatose patients is most often described with the Cerebral Performance Categories (CPCs) 1-5, as shown in Box 17-1.¹³

METHODS

Search Strategy and Quality Review

We conducted a computerized bibliographic search of MEDLINE and EMBASE for 1966-2003 to determine the precision and accuracy of components of the clinical examination in prognosis of hypoxic-ischemic coma. Search terms included "coma," "cardiac arrest," "prognosis," "physical examination," "sensitivity and specificity," and "observer variation." The search was conducted by using a previously published search strategy for The Rational Clinical Examination series.¹⁴ We checked the reference lists of all review articles and primary studies for additional articles that were not identified on the computerized search. Standard physical examination textbooks and personal communications with the authors of primary studies provided additional citations. Finally, we manually reviewed published abstracts from the annual scientific meetings of the American Neurological Association, the American Academy of Neurology, the Society of Critical Care Medicine, and the European Society for Intensive Care Medicine for 1997-2003.

One of the authors (C.M.B.) initially screened the titles and abstracts of the search results and classified them as primary studies, review articles, or not relevant. Because we were interested in both the precision and accuracy of the clinical examination in postcardiac arrest coma, we included primary studies of each type. A preliminary review of the literature revealed few precision studies, so the inclusion criteria for this type of study were broadened. Precision studies were included if they assessed the interobserver agreement in the neurologic examination of comatose adult patients. We included both traumatic and non-traumatic forms of coma.

Primary studies of accuracy were independently reviewed by 2 of us (C.M.B. and R.H.B.) and included if they assessed the accuracy of the clinical examination in prognosis of hypoxic-ischemic coma in patients older than 10 years. Other criteria for study selection were the presentation of outcome data for individual clinical variables measured at discrete intervals. Selected studies also presented neurologic outcome data as defined by the CPCs or in such a manner that an equivalent CPC score could be determined (Box 17-1). Studies were excluded if they involved patients with coma from other medical conditions or trauma.

According to our findings in a preliminary literature search, we realized there were 2 types of accuracy studies in the literature. The majority of studies dichotomized patient outcome as good or poor. Unfortunately, there is not a uniform definition of what constitutes a good vs a poor outcome. Most studies combined outcome data for severe neurologic disability, vegetative state, and death (ie, CPC 3-5) as a poor outcome and normal or moderate disability (ie, CPC 1-2) as a good outcome. However, there were 6 studies that included severe neurologic disability (ie, CPC 3) as a good outcome, 4 of which included fewer than 65 patients.¹⁵⁻²⁰ We included studies from which combined outcome data for severe neurologic disability, vegetative state, and death (ie, CPC 3-5) could be extracted. We did this because most primary studies presented outcome data in this fashion and because we could not combine studies that had differing definitions of good vs poor outcomes. Furthermore, we thought

Box 17-1 Glasgow-Pittsburgh Cerebral Performance Categories^a

1. GOOD CEREBRAL PERFORMANCE

Conscious. Alert, able to work and lead a normal life. May have minor psychological or neurologic deficits (mild dysphasia, nonincapacitating hemiparesis, or minor cranial nerve abnormalities).

2. MODERATE CEREBRAL DISABILITY

Conscious. Sufficient cerebral function for part-time work in sheltered environment or independent activities of daily life (dressing, traveling by public transportation, and preparing food). May have hemiplegia, seizures, ataxia, dysarthria, dysphasia, or permanent memory or mental changes.

3. SEVERE CEREBRAL DISABILITY

Conscious. Dependent on others for daily support because of impaired brain function (in an institution or at home with exceptional family effort). At least limited cognition. Includes a wide range of cerebral abnormalities, from ambulatory with severe memory disturbance or dementia precluding independent existence to paralytic and able to communicate only with eyes, as in the locked-in syndrome.

4. COMA, VEGETATIVE STATE

Not conscious. Unaware of surroundings, no cognition. No verbal or psychological interactions with environment.

5. DEATH

Certified brain dead or dead by traditional criteria.

^aAdapted from Cummings et al.¹³

it was reasonable to assume that most clinicians, patients, and families would not consider severe neurologic disability (defined as CPC 3) a desirable outcome.

The methodologic quality of each primary study was assessed in duplicate with modified criteria previously developed for The Rational Clinical Examination series (see Table 1-7).²¹ Because this study assessed prognosis and not diagnosis, investigators were considered blinded if the study was prospective and clinical variables were assessed before a patient's outcome was known. Level 1 studies were prospective studies with 100 or more consecutive unselected patients. Level 2 studies were similar but involved fewer than 100 patients. Level 3 studies were retrospective chart reviews, and level 4 studies included selected (ie, nonconsecutive) patients.

Statistical Methods

Two authors (C.M.B. and R.H.B.) independently extracted data for analysis; we resolved disagreement by consensus. When data were missing or unclear, we contacted the primary investigators requesting further information. Published raw data were used to calculate positive and negative likelihood ratios (LRs) for specific clinical variables. To create 2×2 evidence tables, we dichotomized CPC 1 and 2 as good outcome and CPC 3 through 5 as poor outcome. Sensitivity was defined as the proportion of patients with a poor neurologic outcome who had a particular physical finding; specificity was the proportion of patients who had a good neurologic outcome and did not have the particular finding.

When 3 or more studies examined the same clinical variable at the same time after cardiac arrest, we calculated summary LRs and 95% confidence intervals (CIs) using bayesian

random-effects meta-analyses. We also present the strongest LRs for individual clinical variables at various times after cardiac arrest. LRs were modeled using a method described by Warn et al²² for relative risks, also using the prior distributions used therein. Posttest probabilities were computed from the estimated pretest probability and LRs.²³ All analyses were done using the WinBUGS software package (Version 1.4, 2003; MRC Biostatistics Unit, Cambridge, England).²⁴

Likelihood Ratios

LRs are a method of converting pretest information (ie, probability, or more precisely, odds) into posttest information.²⁵ The pretest information is the probability of a poor outcome among all comatose survivors of cardiac arrest. The results of the clinical examination, reflected in the LRs for the findings, are combined with the pretest information to estimate the posttest probability of a poor outcome. For clinicians, the easiest way to interpret LRs is to keep in mind that when an abnormal clinical finding is present in a comatose survivor (eg, absent pupillary response), the likelihood of a poor outcome increases and the LR will be greater than 1. Similarly, if the finding does not indicate a poor prognosis (eg, present pupillary response), an LR of less than 1 will occur.

RESULTS

Search Results and Quality of the Evidence

Our search yielded 5 studies of precision that met our inclusion criteria (Table 17-2).²⁶⁻³⁰ Two other studies of precision

Table 17-2 Interobserver Agreement of Clinical Examination for Coma

Source, y	No. of Observers	Observers' Level of Experience	Variable Assessed	Agreement, κ Statistic
Braakman et al, ²⁶ 1977	12	Neurosurgeons and residents	GCS motor	0.72
	20		GCS motor	0.75
Teasdale et al, ²⁷ 1978	7	Neurosurgeons	GCS eye	DR = 14%
			GCS verbal	DR = 5.4%
			GCS motor	DR = 11%
			Pupil response	DR = 4.3%
			Oculocephalic response	0.49
van den Berge et al, ²⁸ 1979	6	Neurosurgeons	Spontaneous eye movement	0.46
			Pupil response	0.65
			Oculocephalic response	0.49
Minderhoud et al, ²⁹ 1982	4	Physicians	GCS verbal	0.59
			GCS motor	0.68
			Pupil response	0.79
			Oculocephalic response	0.74
			GCS eye	0.62
Born et al, ³⁰ 1987	6	Neurosurgeons	Brainstem score	0.69
			Pupil response	0.70
			GCS motor	0.65
			Brainstem reflexes	0.42
			Brainstem reflexes	0.36
			Brainstem reflexes	0.42

Abbreviations: DR, reported disagreement rate; GCS, Glasgow Coma Scale.

Table 17-3 Studies on the Accuracy of the Clinical Examination in Prognosis of Hypoxic-Ischemic Coma^a

Source, y	Level of Evidence	Study Population	Site of Arrest	Mean Age, y ^b	No. of Patients	Neurologic Outcomes ^c		Outcome Assessment
						Good	Poor	
Berek et al, ³³ 1997	2	Postcardiac arrest coma	PH	68	42	13	29	At discharge
Chen et al, ³⁴ 1996	4	Patients in hypoxic-ischemic coma at 24 h	PH or IH	58	34	7	27	3 mo ^d
Earnest et al, ³⁵ 1979	1	Postcardiac arrest coma	PH	62	100	30	70	At discharge
Edgren et al, ³⁶ 1987	4	Postcardiac arrest coma	PH or IH	71	32	11	21	6 mo ^d
Edgren et al, ³⁷ 1994 ^e	1	Postcardiac arrest coma	PH or IH	58	262	89	173	12 mo ^d
Krumholz et al, ³⁸ 1988	1	Patients in postcardiac arrest coma at 24 h	PH or IH	67	114	21	93	At discharge
Levy et al, ³⁹ 1985	1	Hypoxic-ischemic coma	PH or IH	61	210	26	184	12 mo ^d
Madl et al, ⁴⁰ 2000	1	Patients in postcardiac arrest coma at 24 h	PH or IH	57	209	49	160	6 mo ^d
Madl et al, ⁴¹ 1993	2	Postcardiac arrest coma	PH or IH	58	66	17	49	At discharge ^d
Sasser, ⁴² 1999 ^e	1	Patients in postcardiac arrest coma at 12 h	PH or IH	63	937	230	707	6 mo ^d
Snyder et al, ⁴³⁻⁴⁵ 1980-1981	2	Postcardiac arrest coma	PH	64	63	25	38	6 mo ^d
Widjiks et al, ⁴⁶ 1994	3	Postcardiac arrest coma	PH	63	107	15	92	6 mo

Abbreviations: IH, in-hospital cardiac arrest; PH, prehospital cardiac arrest.

^aThe 14 sources represent 11 studies.

^bWhen the mean age was not provided, the median age of the study population is listed.

^cGood neurologic outcome refers to cerebral performance categories (CPCs) 1 and 2. Poor outcome includes CPCs 3 through 5. See Box 17-1 for a definition of CPCs.

^dOutcome refers to best ever CPC in specified period.

^eThis article includes patients from the first Brain Resuscitation Clinical Trial (BRCT), also included in Sasser's⁴² dissertation, which involves all 3 BRCTs.

were excluded because neither rates of agreement (κ) nor raw data were presented.^{31,32} Fourteen accuracy articles describing 11 studies met our inclusion criteria (Table 17-3).³³⁻⁴⁶ We had 100% agreement on the inclusion of studies for the systematic review. Reasons for excluding relevant studies included studies that did not present neurologic outcomes as CPC 1 and 2 as a good outcome and CPC 3-5 as a poor outcome,¹⁵⁻²⁰ studies in which patients were not comatose,⁴⁷⁻⁵² studies that included only patients in a persistent vegetative state,^{53,54} studies that included other forms of medical coma,^{55,56} and studies that presented the same data set.^{3,57} One study was a systematic review of clinical and neurophysiologic variables.⁶

We reached 100% agreement on the methodologic quality scores. Of the 11 accuracy studies, 5 were classified as level 1, 3 as level 2, 1 as level 3, and 2 as level 4. The studies and methodologic quality scores are summarized in Table 17-3.

Precision of the Clinical Examination of Coma

Five studies have reported the precision of the examination of comatose patients (Table 17-2). Heterogeneity in study methodology, patient population, and variables assessed precluded a quantitative synthesis of results; thus, these studies were reviewed qualitatively. As presented in Table 17-2, interobserver agreement was moderate to substantial in each of the studies. Three studies found no difference in interobserver agreement among experienced nurses, residents, and physicians.²⁶⁻²⁸ One study did find precision to be diminished in groups of less experienced examiners.³⁰ No study examined only patients with nontraumatic causes of coma. In summary, there was reasonable consistency among studies, and the precision of the clinical examination of coma

(including components of the GCS and brainstem reflexes) has been found to be moderate to substantial.

Accuracy of the Clinical Examination of Coma

Fourteen articles involving 11 studies of the accuracy of the clinical examination were included (Table 17-3). These studies provided a sample size of 1914 comatose survivors of cardiac arrest. The proportion of individuals dying or having a poor neurologic outcome was calculated by pooling the outcome data from the 11 studies and was used as an estimate of the pretest probability of poor outcome (Table 17-3). The random-effects estimate of poor outcome was 77% (95% CI, 72%-80%). This value represents an estimate of the pretest probability of death or a poor outcome for the entire population of comatose survivors of cardiac arrest, and it is combined with the LRs for various clinical findings to revise the estimated probability of a poor clinical outcome.

Motor Response and Brainstem Reflexes

Six studies examined the association between motor and brainstem function and the recovery of comatose survivors of cardiac arrest. Data for specific clinical findings were pooled if they were assessed in at least 3 studies. Table 17-4 shows potentially useful clinical findings from individual studies. Summary measures for pooled variables are shown in Table 17-5.

In 1987, Edgren et al³⁶ reported motor and brainstem function in 32 comatose patients at 24 and 48 hours after cardiac arrest. Patients were weaned from intensive care at 72 hours if they did not respond to pain and had no evidence of brainstem reflexes. Chen et al³⁴ examined similar clinical variables

Table 17-4 Useful Clinical Findings in the Prognosis of Postcardiac Arrest Coma Organized by Time After Onset of Coma (Not Pooled)^a

Clinical Finding	Study	LR of Poor Neurologic Outcome (95% CI) ^b	
		Positive	Negative
At Onset of Coma			
Absent pupillary reflex	Earnest et al ³⁵	7.2 (1.9-28.0)	0.5 (0.4-0.6)
Absent motor response	Levy et al ³⁹	3.5 (1.4-8.6)	0.6 (0.4-0.7)
Absent corneal reflex	Levy et al ³⁹	3.2 (1.1-9.5)	0.7 (0.6-0.8)
Absent oculocephalic reflex	Earnest et al ³⁵	2.5 (1.3-4.8)	0.4 (0.3-0.6)
Absent spontaneous eye movement	Levy et al ³⁹	2.2 (1.3-4.0)	0.4 (0.3-0.6)
ICS < 4	Berek et al ³³	2.2 (1.1-4.5)	0.2 (0.1-0.6)
GCS < 5	Madl et al ⁴⁰	1.4 (1.1-1.6)	0.3 (0.2-0.5)
Absent verbal effort	Levy et al ³⁹	1.2 (0.9-1.6)	0.1 (0.0-0.7)
At 12 h			
Absent cough reflex	Sasser ⁴²	13.4 (4.4-40.3)	0.3 (0.2-0.4)
Absent corneal reflex	Sasser ⁴²	9.1 (3.9-21.1)	0.3 (0.2-0.4)
Absent gag reflex	Sasser ⁴²	8.7 (4.0-18.9)	0.4 (0.4-0.5)
Absent pupillary reflex	Sasser ⁴²	4.0 (2.5-6.6)	0.5 (0.5-0.6)
GCS < 5	Sasser ⁴²	3.5 (2.4-5.2)	0.4 (0.3-0.4)
Absent motor response	Sasser ⁴²	3.2 (2.2-4.6)	0.4 (0.3-0.5)
Absent withdrawal to pain	Sasser ⁴²	2.4 (1.9-3.1)	0.2 (0.1-0.2)
Absent verbal effort	Sasser ⁴²	1.6 (1.4-1.9)	0.1 (0.0-0.1)
At 24 h			
Absent cough reflex	Sasser ⁴²	84.6 (5.3-1342.0)	0.4 (0.3-0.5)
Absent gag reflex	Sasser ⁴²	24.9 (6.3-98.3)	0.5 (0.4-0.5)
GCS < 5	Sasser ⁴²	8.8 (5.1-15.1)	0.4 (0.3-0.4)
Absent eye opening to pain	Sasser ⁴²	5.9 (3.9-9.0)	0.3 (0.3-0.4)
Absent spontaneous eye movement	Levy et al ³⁹	3.5 (1.4-8.8)	0.5 (0.4-0.7)
Absent eye opening to pain	Levy et al ³⁹	3.0 (1.5-6.2)	0.4 (0.3-0.5)
Absent oculocephalic reflex	Sasser ⁴²	2.9 (1.8-4.6)	0.5 (0.5-0.6)
Absent spontaneous eye movement	Sasser ⁴²	2.7 (2.1-3.4)	0.3 (0.2-0.3)
Absent verbal effort	Sasser ⁴²	2.4 (2.0-2.9)	0.1 (0.0-0.1)
At 48 h			
GCS < 6	Madl et al ⁴¹	2.8 (1.3-5.9)	0.3 (0.1-0.5)
GCS < 10	Madl et al ⁴¹	1.3 (1.0-1.7)	0.0 (0.0-0.7)
At 72 h			
Absent withdrawal to pain	Levy et al ³⁹	36.5 (2.3-569.9)	0.3 (0.2-0.4)
Absent spontaneous eye movement	Levy et al ³⁹	11.5 (1.7-79.0)	0.6 (0.5-0.7)
Absent verbal effort	Levy et al ³⁹	7.4 (2.0-28.0)	0.3 (0.2-0.5)
Absent eye opening to pain	Levy et al ³⁹	6.9 (1.8-27.0)	0.5 (0.4-0.6)
At 7 d			
Absent withdrawal to pain	Levy et al ³⁹	29.7 (1.9-466.0)	0.4 (0.3-0.6)
Absent verbal effort	Levy et al ³⁹	14.1 (2.0-97.7)	0.4 (0.2-0.6)

Abbreviations: CI, confidence interval; GCS, Glasgow Come Scale; ICS, Innsbruck Coma Scale³³; LR, likelihood ratio.

^aClinical findings that have a positive LR greater than 2 and lower CI boundary greater than 1 are presented with the corresponding negative LR.

^bThe positive LR indicates that the abnormal clinical finding shown in the left column was present. The negative LR indicates that the patient had a normal result for the clinical finding; thus, the negative LR in the first row is the value associated with the presence of normal pupillary reflexes.

in a study of 34 comatose patients. As in the study by Edgren et al,³⁶ patients with absent brainstem reflexes at 24 hours were excluded from this study.

The Brain Resuscitation Clinical Trials (BRCTs) were a series of 3 large prospective, randomized, multicenter studies

of pharmacologic interventions in cardiac arrest. In BRCT I⁵⁸ (1979-1984), 262 comatose survivors of cardiac arrest were assessed for the use of a barbiturate (thiopental). In BRCT II⁵⁹ (1984-1989), 516 comatose patients were randomly assigned to receive placebo or a calcium-channel blocker

(lidoflazine) after cardiac arrest. In BRCT III⁶⁰ (1989-1992), 2915 patients were randomly assigned to receive standard- or high-dose epinephrine during cardiac arrest. All 3 BRCT studies reported negative results; there was no difference found in survival or neurologic outcome among treatment groups. Two articles described the association between clinical neurologic signs and outcome in the BRCT study population. In 1994, Edgren et al³⁷ reported the neurologic examination and outcomes of the 109 individuals in BRCT I who had survived to 72 hours. In an analysis of all 3 BRCT studies, Sasser⁴² assessed the prognostic utility of motor response and brainstem reflexes at 12 and 24 hours after cardiac arrest. As in all studies of cardiac arrest, there was a high degree of early mortality. Accordingly, only 1450 patients of the original 3693 studied in all 3 BRCTs survived to 12 hours. Of this group, 506 patients were sedated or anesthetized at the neurologic examination and therefore were not included in Sasser's⁴² review. Of the remaining 944 patients, outcome data were available for 937. This is the largest population of comatose survivors of cardiac arrest reported to date.

Summary measures for clinical variables that were assessed in at least 3 studies are presented in Table 17-5. Five pooled variables were found to have a 95% CI lying entirely above 1. The clinical signs at 24 hours with the highest LRs were absent corneal reflexes (LR, 13; 95% CI, 2.0-69), absent pupillary reflexes (LR, 10; 95% CI, 1.8-49), absent motor response (LR, 4.9; 95% CI, 1.6-13), and absent withdrawal to pain (LR, 4.7; 95% CI, 2.2-9.8). At 72 hours after cardiac arrest, absent motor response was found to accurately predict death or poor neurologic outcome (LR, 9.2; 95% CI, 2.1-49). No clinical findings were found to accurately predict good neurologic outcome (ie, no useful negative LRs).

Coma Scales

Four studies assessed composite coma scores as prognostic indicators in postcardiac arrest coma. Madl et al⁴¹ reported 2 studies that assessed the role of the GCS in predicting neurologic recovery. In 1993, this group reported on a series of 66 comatose patients who survived cardiac arrest.⁴¹ The GCS at 48 hours was compared with survival and functional recovery. A second study of 209 patients measured GCS on admission to the intensive care unit after cardiac arrest.⁴⁰ In the BRCT reports, GCS scores at 12, 24, and 72 hours were compared with neurologic recovery.^{37,42} In 1997, Berek et al³³ examined the utility of the Innsbruck Coma Scale (ICS) in 42 comatose patients who survived prehospital arrest. The ICS includes an assessment of the GCS components in addition to various brainstem reflexes. A score from 0 to 23 is assigned. A lower score indicates more severe neurologic deficits.

Although the composite coma scores did predict poor neurologic outcome, they were not as predictive as the individual motor and brainstem reflex components. This is demonstrated in Table 17-4.

Seizures

Four studies have examined whether seizures in the post-arrest period accurately predict outcome. In 1988, Krumholz et al³⁸ described 114 comatose survivors of cardiac arrest,

Table 17-5 Pooled Clinical Signs in the Prognosis of Postcardiac Arrest Coma

Source	LR of Poor Neurologic Outcome (95% CI)	
	Positive	Negative
At Time of Coma Onset ^a		
Absent withdrawal to pain		
Summary LR	1.7 (0.7-4.2)	0.4 (0.1-1.1)
Earnest et al ³⁵	3.7 (1.6-8.2)	0.4 (0.3-0.6)
Levy et al ³⁹	1.4 (1.0-1.9)	0.4 (0.2-0.7)
Snyder et al ⁴³	1.4 (0.9-2.1)	0.5 (0.2-1.2)
At 24 h		
Absent withdrawal to pain		
Summary LR	4.7 (2.2-9.8)	0.2 (0.1-0.6)
Edgren et al ³⁶	3.9 (1.1-14)	0.4 (0.2-0.8)
Levy et al ³⁹	6.8 (2.3-20)	0.2 (0.2-0.3)
Sasser ⁴²	5.1 (3.6-7.3)	0.2 (0.1-0.2)
Snyder et al ⁴³	6.5 (1.0-42)	0.3 (0.1-0.7)
Absent pupil response		
Summary LR	10 (1.8-49)	0.8 (0.4-1.4)
Chen et al ³⁴	0.9 (0.0-19)	1.0 (0.8-1.2)
Edgren et al ³⁶	5.6 (0.3-95)	0.8 (0.6-1.1)
Levy et al ³⁹	11 (0.7-170)	0.8 (0.7-0.9)
Sasser ⁴²	39 (5.6-277)	0.6 (0.6-0.7)
Absent motor response		
Summary LR	4.9 (1.6-13)	0.6 (0.3-1.3)
Chen et al ³⁴	3.7 (0.2-59)	0.8 (0.6-1.1)
Levy et al ³⁹	5.5 (1.4-21)	0.6 (0.5-0.8)
Sasser ⁴²	7.6 (4.6-13)	0.4 (0.3-0.4)
Snyder et al ⁴³	3.5 (0.5-24)	0.7 (0.5-1.1)
Absent corneal reflex		
Summary LR	13 (2.0-69)	0.6 (0.2-1.9)
Edgren et al ³⁶	1.8 (0.2-15.4)	0.9 (0.7-1.2)
Levy et al ³⁹	15 (0.9-233)	0.7 (0.7-0.8)
Sasser ⁴²	91 (5.7-1443)	0.4 (0.4-0.5)
At 72 h		
Absent pupil response		
Summary LR	3.4 (0.5-24)	0.9 (0.4-2.1)
Chen et al ³⁴	0.9 (0.0-19)	1.0 (0.8-1.2)
Edgren et al ³⁷	5.3 (0.3-84)	0.8 (0.7-1.0)
Levy et al ³⁹	5.8 (0.4-94)	0.9 (0.8-1.0)
Absent motor response		
Summary LR	9.2 (2.1-49)	0.7 (0.3-1.3)
Chen et al ³⁴	2.0 (0.1-35)	0.9 (0.7-1.2)
Edgren et al ³⁷	13 (0.8-193)	0.6 (0.5-0.7)
Levy et al ³⁹	16 (1.1-261)	0.7 (0.6-0.8)
Snyder et al ⁴³	3.0 (0.2-39)	0.6 (0.3-1.1)
Seizure or myoclonus^b		
Summary LR	1.4 (0.5-3.9)	0.8 (0.3-2.1)
Krumholz et al ³⁸	1.7 (0.8-3.4)	0.7 (0.5-1.0)
Levy et al ³⁹	1.1 (0.5-2.3)	1.0 (0.8-1.2)
Snyder et al ⁴⁴	1.7 (0.7-4.2)	0.8 (0.6-1.1)

Abbreviations: CI, confidence interval; LR, likelihood ratio.

^aTimes reflect number of hours since cardiac arrest.

^bThese figures refer to the presence of seizures or myoclonus at any time after cardiac arrest.

Nearly half (44%) of the patients had some seizure activity. In a study conducted by Snyder et al⁴⁴ on 63 patients, 19 (30%) had seizures or myoclonus. In 1994, Widjiks et al⁴⁶ described the prevalence of myoclonus status in a group of 107 patients. Forty (37%) of 107 patients had myoclonus status within 24 hours. In the study conducted by Levy et al³⁹ on 210 patients, 53 (25%) had seizure or myoclonic activity. Most clinicians infer that seizures portend a poor prognosis in comatose survivors of cardiac arrest. However, none of the individual studies or the summary measures established that seizures accurately predict outcome (Table 17-5).

CLINICAL SCENARIOS—RESOLUTIONS

In both cases, an estimate of the pretest probability (derived from our overall study population) of poor neurologic outcome is 77%. This figure will vary according to comorbid disease, duration of cardiopulmonary resuscitation, and other clinical variables. The 65-year-old man who withdraws to pain and has intact brainstem reflexes 24 hours after cardiac arrest has none of the clinical findings associated with poor neurologic outcome. In discussing this with the family, it is important to explain that although there are no signs suggestive of poor outcome, the physical examination is much less useful in predicting good outcome. Consequently, his probability of poor neurologic outcome remains unchanged (ie, 77%).

In the second case, the young woman has no brainstem reflexes or response to painful stimuli at 3 days. Unfortunately, these findings suggest an extremely poor chance of meaningful neurologic recovery. The most powerful of these indicators elevates her posttest probability of poor neurologic outcome to 97%. Although the existing literature does not examine the combined effects of different physical examination predictors, because she has multiple poor prognostic findings her prognosis may be even worse. You should recognize that the observation of reactive pupils immediately after cardiac arrest and the presence of myoclonus are not useful in determining her neurologic prognosis.

THE BOTTOM LINE

In this systematic review we found that the precision of the neurologic examination in comatose patients is moderate to substantial. According to our results, we suggest that in patients who lack pupillary and corneal reflexes at 24 hours and have no motor response at 72 hours, the chance of meaningful neurologic recovery is small. This meta-analysis includes almost 2000 patients and is the largest such review to date. In addition to providing other information, it corroborates the findings of the oft-quoted study by Levy et al,³⁹ in which none of the 210 patients who had any of these 3 clinical findings ever regained an independent lifestyle.

In our study population, the random-effects estimate of poor outcome was 77% (95% CI, 72%-80%). The highest LR increases the pretest probability of 77% to a posttest probability of 97% (95% CI, 87%-100%). Immediately after car-

diac arrest, no clinical signs accurately predict the patient's outcome. Finally, no clinical findings were found to have LRs that strongly predicted good neurologic outcome.

The results of our meta-analysis should be interpreted in the context of study limitations. To calculate LRs from 2 × 2 tables, there must be a delineation between what constitutes a good vs a poor neurologic outcome. We chose to define poor outcome as death, vegetative state, or severe neurologic impairment (precluding independent living). We made this decision because that is where most primary studies dichotomize outcome. Furthermore, we believe most patients, families, and physicians would not consider severe neurologic impairment a desirable outcome. However, in applying the results of this study to individual patients, physicians must realize that some families and patients may have different perceptions of what constitutes an acceptable neurologic outcome. It was not the purpose of this study to provide an ethical framework for treatment decisions in the management of comatose survivors of cardiac arrest; rather, we attempted to summarize the existing literature to provide guidance to clinicians and families about prognostic probabilities.

Any study of prognosis in the critically ill is potentially influenced by the tendency for poor prognoses to be self-fulfilling. It is difficult to determine whether poor neurologic outcomes are caused by decisions to withdraw or withhold therapy according to a perceived poor neurologic prognosis. This has the potential to artificially elevate positive LRs. Although there is no empirical evidence that this occurred in our study population, this clinical reality does remain a limitation of the existing literature.

It would be potentially useful to assess whether combinations of neurologic findings could improve the accuracy of prognosis in comatose survivors of cardiac arrest. Unfortunately, we were unable to perform this analysis because the available literature does not provide these data. In 3 studies, combinations of findings were assessed. In the analysis of 262 patients by Edgren et al,³⁷ no combination of findings was found to be more predictive than the individual variables. Sasser,⁴² who performed a detailed analysis of combined neurologic findings and demographic, comorbidity, and cardiopulmonary resuscitation variables, did not find any additional predictive value of the algorithm (sensitivity, 59%; specificity, 93%). Only Levy et al³⁹ found practical and useful algorithms that combined various neurologic findings. These are clearly presented in their article.

Finally, the 11 studies included in this meta-analysis represent a diverse and heterogeneous population with various comorbidities. For example, it is unclear what effect individual medications or hypothermic cooling may have on the bedside clinical examination. Consequently, the applicability of our results to individual patients must be made with caution and as part of the larger clinical picture. We do not suggest a direct extension of our results to the decision to proceed with or withdraw from medical care. Rather, we present information that we hope will allow the decision to be made on a more rational basis.

In summary, simple physical examination maneuvers strongly predict death or poor neurologic outcome in coma-

tose survivors of cardiac arrest. Although decisions to proceed with care or withdraw care may take place at later times for a variety of reasons, the most useful signs occur after at least 24 hours and in the case of motor response at 72 hours postcardiac arrest. The existing literature does not allow for an earlier prognosis to be made on the basis of the clinical examination alone.

Author Affiliations at the Time of the Original Publication

Departments of Medicine (Drs Booth, Boone, Tomlinson, and Detsky) and Health Policy Management and Evaluation (Dr Detsky), University of Toronto, and Department of Medicine, Mount Sinai Hospital (Drs Tomlinson and Detsky) and University Health Network (Drs Tomlinson and Detsky), Toronto, Ontario, Canada.

Acknowledgments

This project was partially funded through the financial support of Griffen Rotman.

We thank David Simel, MD, for his valuable guidance during the course of this study. We also thank Tom Stewart, MD, Marika Hohol, MD, Karen Welty-Wolf, MD, James Tulsky, MD, and Howell Sasser, PhD, for their expert advice and comments on this article.

REFERENCES

- Eisenberg MS, Mengert TJ. Cardiac resuscitation. *N Engl J Med*. 2001;344(17):1304-1313.
- Becker LR, Ostrander MP, Barrett J, Kondos GT. Outcome of CPR in a large metropolitan area. *Ann Emerg Med*. 1991;20(4):355-361.
- Bates D, Caronna JJ, Cartlidge NE, et al. A prospective study of nontraumatic coma: methods and results in 310 patients. *Ann Neurol*. 1977;2(3):211-220.
- Shewmon DA, De Giorgio CM. Early prognosis in anoxic coma. *Neurology*. 1989;7(4):823-843.
- Elheff SM, Hanly DF. Postresuscitation prognostication and declaration of brain death. In: Paradis NA, Halperin HR, Novak RA, eds. *Cardiac Arrest: The Science and Practice of Resuscitation Medicine*. Baltimore, MD: Williams & Wilkins; 1996:910-922.
- Zandbergen EG, de Haan RJ, Stoutenbeck CP, Loelman JH, Kijdra A. Systematic review of early prediction of poor outcome in anoxic-ischaemic coma. *Lancet*. 1998;352(9153):1808-1812.
- Robinson LR, Micklesen PJ, Tirschwell DL, Lew HL. Predictive value of somatosensory evoked potentials for awakening from coma. *Crit Care Med*. 2003;31(3):960-967.
- Zingler VC, Krumm B, Bertsch T, Fassbender K, Pohlmann-Eden B. Early prediction of neurological outcome after cardiopulmonary resuscitation: a multimodal approach combining neurobiochemical and electrophysiological investigations may provide high prognostic certainty in patients after cardiac arrest. *Eur Neurol*. 2003;49(2):79-84.
- Safar P, Behringer W, Bottiger BW, Sterz F. Cerebral resuscitation potentials for cardiac arrest. *Crit Care Med*. 2002;30(4 suppl):S140-S144.
- Plum F, Posner JB. *The Diagnosis of Stupor and Coma*. 3rd ed. Philadelphia, PA: FA Davis; 1980;1-86.
- Teasdale G, Jennett B. Assessment of coma and impaired consciousness: a practical scale. *Lancet*. 1974;2(7872):81-84.
- Campbell WW. *DeJong's The Neurologic Examination*. 6th ed. Philadelphia, PA: JB Lippincott Co; 2005:699-732.
- Cummings RO, Chamberlain DA, Abramson NS, et al. Recommended guidelines for uniform reporting of data from out-of-hospital cardiac arrest: the Utstein style. *Circulation*. 1991;84(2):960-975.
- Lederle FA, Simel DL. Does this patient have abdominal aortic aneurysm? *JAMA*. 1999;281(1):77-82.
- Bassetti C, Bomio F, Mathis J, Hess CW. Early prognosis in coma after cardiac arrest: a prospective clinical, electrophysiological, and biochemical study of 60 patients. *J Neurol Neurosurg Psychiatry*. 1996;61(6):610-615.
- Bertini G, Margheri M, Giglioli C, et al. Prognostic significance of early clinical manifestations in postanoxic coma: a retrospective study of 58 patients resuscitated after prehospital cardiac arrest. *Crit Care Med*. 1989;17(7):627-633.
- Longstreth WT Jr, Diehr P, Inui TS. Prediction of awakening after out-of-hospital cardiac arrest. *N Engl J Med*. 1983;308(23):1378-1382.
- Mullie A, Verstringe P, Buylaert W, et al, for the Cerebral Resuscitation Study Group of the Belgian Society for Intensive Care. Predictive value of Glasgow coma score for awakening after out-of-hospital cardiac arrest. *Lancet*. 1988;1(8578):137-140.
- Pifferi S, Codazzi D, Savioli M, Langer M. Early prediction of neurologic prognosis after post-anoxic coma. *Intensive Care Med*. 1998;24(5):535-536.
- Sacco RL, VanGool R, Mohr JP, Hauser WA. Nontraumatic coma: Glasgow coma score and coma etiology as predictors of 2-week outcome. *Arch Neurol*. 1990;47(11):1181-1184.
- Holleman DR, Simel DL. Does the clinical examination predict airflow limitation? *JAMA*. 1995;273(4):313-319.
- Warn DE, Thompson SG, Spiegelhalter DJ. Bayesian random effects meta-analysis of trials with binary outcomes: methods for the absolute risk difference and relative risk scales. *Stat Med*. 2002;21(11):1601-1623.
- Sackett DL, Straus SE, Richardson WS, Rosenberg W, Haynes RB. *Evidence-Based Medicine: How to Practise and Teach EBM*. 2nd ed. Edinburgh, Scotland: Churchill Livingstone; 2000:72-74.
- WinBUGS [computer program]. Version 1.4. Cambridge, England: MRC Biostatistics Unit; 2003.
- Sackett DL. A primer on the precision and accuracy of the clinical examination. *JAMA*. 1992;267(19):2638-2644.
- Braakman R, Avezaat CJ, Maas AI, Roel M, Schouten HJ. Interobserver agreement in the assessment of the motor response of the Glasgow Coma Scale. *Clin Neurol Neurosurg*. 1977;80(2):100-106.
- Teasdale G, Knill-Jones R, Van der Sande J. Observer variability in assessing impaired consciousness and coma. *J Neurol Neurosurg Psychiatry*. 1978;41(7):603-610.
- van den Berge JH, Schouten HJ, Boomstra S, van Drunen Littel S, Braakman R. Interobserver agreement in assessment of ocular signs in coma. *J Neurol Neurosurg Psychiatry*. 1979;42(12):1163-1168.
- Minderhoud JM, Huizenga J, van Woerkom TC, Blomjous CE. The pattern of recovery after severe head injury. *Clin Neurol Neurosurg*. 1982;84(1):15-28.
- Born JD, Hans P, Albert A, Bonnal J. Interobserver agreement in assessment of motor response and brain stem reflexes. *Neurosurgery*. 1987;20(4):513-517.
- Teasdale G, Jennett B. Assessment and prognosis of coma after head injury. *Acta Neurochir (Wien)*. 1976;34(1-4):45-55.
- Van Gijn J, Bonke B. Interpretation of plantar reflexes: biasing effect of other signs and symptoms. *J Neurol Neurosurg Psychiatry*. 1977;40(8):787-789.
- Berek K, Schinneri A, Traweger C, Lechleitner P, Baubin M, Aichner F. The prognostic significance of coma-rating, duration of anoxia and cardiopulmonary resuscitation in out-of-hospital cardiac arrest. *J Neurol*. 1997;244(9):556-561.
- Chen R, Bolton CF, Young GB. Prediction of outcome in patients with anoxic coma: a clinical and electrophysiologic study. *Crit Care Med*. 1996;24(4):672-678.
- Earnest MP, Breckinridge JC, Yarnell PR, Olivia PB. Quality of survival after out-of-hospital cardiac arrest: predictive value of early neurologic evaluation. *Neurology*. 1979;29(1):56-60.
- Edgren E, Hedstrand U, Nordin M, Rydin E, Ronquist G. Prediction of outcome after cardiac arrest. *Crit Care Med*. 1987;15(9):820-825.
- Edgren E, Hedstrand U, Kelsey S, Sutton-Tyrell K, Safar P. Assessment of neurological prognosis in comatose survivors of cardiac arrest. *Lancet*. 1994;343(8905):1055-1059.
- Krumholz A, Stern BJ, Weiss HD. Outcome from coma after cardiopulmonary resuscitation: relation to seizures and myoclonus. *Neurology*. 1988;38(3):401-405.
- Levy DE, Caronna JJ, Singer BH, Lapinski RH, Frydman H, Plum F. Predicting outcome from hypoxic-ischemic coma. *JAMA*. 1985;253(10):1420-1426.
- Madl C, Kramer L, Domanovits H, et al. Improved outcome prediction in unconscious cardiac arrest survivors with sensory evoked potential compared with clinical assessment. *Crit Care Med*. 2000;28(3):721-726.

41. Mall C, Grimm G, Kramer L, et al. Early prediction of individual outcome after cardiopulmonary resuscitation. *Lancet*. 1993;341(8849):855-858.
42. Sasser H. *Association of Clinical Signs With Neurological Outcome After Cardiac Arrest* [dissertation]. Pittsburgh, PA: University of Pittsburgh; 1999.
43. Snyder BD, Loewenson RB, Gumnit RJ, Hauser A, Leppik IE, Ramirez-Lassepas M. Neurologic prognosis after cardiopulmonary arrest, II: level of consciousness. *Neurology*. 1980;30(1):52-58.
44. Snyder BD, Hauser A, Loewenson RB, Leppik IE, Ramirez-Lassepas M, Gumnit RJ. Neurologic prognosis after cardiopulmonary arrest, III: seizure activity. *Neurology*. 1980;30(12):1292-1297.
45. Snyder BD, Gumnit RJ, Leppik IE, Hauser A, Loewenson RB, Ramirez-Lassepas M. Neurologic prognosis after cardiopulmonary arrest, IV: brainstem reflexes. *Neurology*. 1981;31(9):1092-1097.
46. Widjiks EF, Parisi JE, Sharbrough FW. Prognostic value of myoclonus status in comatose survivors of cardiac arrest. *Ann Neurol*. 1994;35(2):239-243.
47. Grubb NR, Elton RA, Fox KA. In-hospital mortality after out-of-hospital cardiac arrest. *Lancet*. 1995;346(8972):417-421.
48. Kentsch M, Stendel M, Berkel H, Mueller-Esch G. Early prediction of prognosis in out-of-hospital cardiac arrest. *Intensive Care Med*. 1990;16(6):378-383.
49. Snyder BD, Ramirez-Lassepas M, Lippert DM. Neurologic status and prognosis after cardiopulmonary arrest, I: a retrospective study. *Neurology*. 1977;27(9):807-811.
50. Steen-Hansen JE, Hansen NN, Vaagenes P, Schreiner B. Pupil size and light reactivity during cardiopulmonary resuscitation: a clinical study. *Crit Care Med*. 1988;16(1):69-70.
51. Thompson RJ, McCullough PA, Kahn JK, O'Neill WW. Prediction of death and neurologic outcome in the emergency department in out-of-hospital cardiac arrest survivors. *Am J Cardiol*. 1998;81(1):17-21.
52. Willoughby JO, Leach BG. Relation of neurological findings after cardiac arrest to outcome. *BMJ*. 1974;3(5829):437-439.
53. Andrews K. Recovery of patients after four months or more in the persistent vegetative state. *BMJ*. 1993;306(6892):1597-1600.
54. Dougherty JH Jr, Rawlinson DG, Levy DE, Plum F. Hypoxic-ischemic brain injury and the vegetative state: clinical and neuropathologic correlation. *Neurology*. 1981;31(8):991-997.
55. Hung TP, Chen ST. Prognosis of deeply comatose patients on ventilators. *J Neurol Neurosurg Psychiatry*. 1995;58(1):75-80.
56. Kano T, Shimoda O, Morioka T, Yagishita Y, Hashiguchi A. Evaluation of the central nervous function in resuscitated comatose patients by multi-level evoked potentials. *Resuscitation*. 1992;23(3):235-248.
57. Levy DE, Bates D, Caronna JJ, et al. Prognosis in nontraumatic coma. *Ann Intern Med*. 1981;94(3):293-301.
58. Brain Resuscitation Clinical Trial I Study Group. Randomized clinical study of thiopental loading in comatose survivors of cardiac arrest. *N Engl J Med*. 1986;314(7):397-403.
59. Brain Resuscitation Clinical Trial II Study Group. A randomized clinical study of a calcium-entry blocker (lidoflazine) in the treatment of comatose survivors of cardiac arrest. *N Engl J Med*. 1991;324(18):1225-1231.
60. Bar-Joseph G, Abramson NS, Jansen-McWilliams L, et al, Brain Resuscitation Clinical Trial III (BRCT III) Study Group. Clinical use of sodium bicarbonate during cardiopulmonary resuscitation—is it used sensibly? *Resuscitation*. 2002;54(1):47-55.

Prepared by Christopher Booth, MD

Reviewed by David L. Simel, MD, MHS

CLINICAL SCENARIO

A 26-year-old woman presented to the emergency department with severe pleuritic chest pain and dyspnea. While waiting for a computed tomographic scan in the radiology department, she had an asystolic cardiac arrest. The resuscitation lasted 20 minutes, after which she had reactive pupils. You have been asked to see her 3 days later for prognosis because the family is considering withdrawing care. On examination, her pupils are now unreactive, and she has no motor response or brainstem reflexes. The nurse tells you she had myoclonus 12 hours ago.

ORIGINAL REVIEW

Booth CM, Boone RH, Tomlinson G, Detsky AS. Is this patient dead, vegetative, or severely neurologically impaired? assessing outcome for comatose survivors of cardiac arrest. *JAMA*. 2004;291(7):870-879.

UPDATED LITERATURE SEARCH

We reviewed 46 citations identified, using the same search strategy used in the original article. From 2003 to January 2006, we found no additional articles on the accuracy of physical examination for predicting outcome of comatose survivors of cardiac arrest.

CLINICAL SCENARIO—RESOLUTION

Three days after resuscitation, she has no pupillary, motor, or brainstem response. Myoclonus has been observed. These are poor prognostic signs, with the lack of motor response conferring a likelihood ratio of 9.2 for a poor response.

See next page for the “Make the Diagnosis” section.

REFERENCE FOR THE UPDATE

- Cummings RO, Chamberlain DA, Abramson NS, et al. Recommended guidelines for uniform reporting of data from out-of-hospital cardiac arrest: the Utstein style. *Circulation*. 1991;84(2):960-975.

OUTCOME FOR COMATOSE SURVIVORS OF CARDIAC ARREST—MAKE THE DIAGNOSIS

PRIOR PROBABILITY

A poor neurologic outcome (severe neurologic disability, vegetative state, or death) occurs in 77% of victims after a non-traumatic cardiac arrest.

ASSESSING THE LIKELIHOOD OF A POOR OUTCOME

See [Table 17-6](#).

Table 17-6 Likelihood Ratios of Signs That Predict Poor Prognosis Change Over Time

	LR+ (95% CI), Finding Absent	LR- (95% CI), Finding Present
Examination at 24 h		
Corneal reflex	13 (2.0-69)	0.6 (0.2-1.9)
Pupillary response	10 (1.8-49)	0.8 (0.4-1.4)
Any motor response to pain	4.9 (1.6-13)	0.6 (0.3-1.3)
Withdrawal to pain	4.7 (2.2-9.8)	0.2 (0.1-0.6)
Examination at 72 h		
Any motor response to pain	9.2 (2.1-49)	0.7 (0.3-1.3)
Pupillary response	3.4 (0.5-24)	0.9 (0.4-2.1)
Seizure or myoclonus at any time after the cardiac arrest	1.4 (0.5-3.9)	0.8 (0.3-2.1)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

- Individual examination findings work better than scales in predicting the likelihood of death or a poor neurologic outcome.
- The examination results at 24 hours and then at 72 hours are more important than the findings immediately after resuscitation.
- The presence of normal findings does not guarantee a good outcome.
- Seizures at 72 hours have minimal effect on predicting the outcome.

REFERENCE STANDARD TESTS

Although death can be defined with traditional biological criteria, there are also cultural and legal definitions of death. A patient who is unaware of his or her surroundings and who has no cognition of or verbal or psychological interaction with the environment characterizes a comatose or vegetative state. No existing tests for recent postcardiac arrest serve as a reference standard for predicting the clinical outcome. When decisions about coma or vegetative states are required, clinicians must often resort to panels of experts to agree on the patient's condition. Other categories of outcomes are described in the Glasgow-Pittsburgh Cerebral Performance Categories.¹

CHAPTER 18

Does This Patient Have Deep Vein Thrombosis?

Sonia Anand, MD, MSc

Philip S. Wells, MD, MSc

Dereck Hunt, MD

Pat Brill-Edwards, MD, MSc

Deborah Cook, MD, MSc

Jeffrey S. Ginsberg, MD

Deep vein thrombosis (DVT) affects approximately 2 million US individuals per year¹ and is the third most common cardiovascular disease, behind acute coronary syndromes and stroke.² Venous thromboembolism represents a single disease entity, with 2 patterns of clinical presentation: DVT and pulmonary embolism (PE). The approach to patients who present with suspected DVT is problematic for several reasons. If left untreated, affected patients can experience fatal PE. The clinical diagnosis of DVT is unreliable when used in isolation without objective testing.^{3,4} About three-quarters of the patients who present with suspected DVT have nonthrombotic causes of leg pain.^{5,6} Finally, although anticoagulant therapy is highly effective in preventing the extension, embolization, and recurrence of DVT, it is associated with an increased risk of major bleeding (approximately 5%) and other potentially serious consequences such as heparin-induced thrombocytopenia (approximately 1%).⁷ Therefore, when possible, anticoagulation should be restricted to those with confirmed DVT. For all of these reasons, it is important to diagnose DVT accurately. This will allow administration of appropriate therapy for patients with documented DVT; for patients without DVT, it will prevent unnecessary exposure of patients to the hazards of anticoagulant therapy and prevent many from being falsely labeled as having venous thromboembolic disease.

The low specificity of clinical symptoms and signs means that most symptomatic patients will not have DVT. Of those symptomatic patients with confirmed DVT at presentation, which represents about one-fourth of patients who are investigated,^{6,8} approximately 80% have proximal DVT (popliteal or more proximal veins) and 20% have DVT that is limited to the calf.⁹ The clinical significance of proximal DVT is different from that of calf vein thrombosis because proximal vein thrombosis is associated with a higher incidence of PE. Pulmonary embolism is detected in approximately 50% of patients with documented proximal DVT.¹⁰ Therefore, proximal DVT should be identified and anticoagulant treatment should be initiated immediately in affected patients. The initiation of appropriate treatment reduces the risk of developing recurrent DVT to about 5% and reduces the incidence of fatal PE to less than 1%.^{1,11} On the other hand, calf vein thrombosis rarely causes PE unless it first extends into the proximal veins. Proximal extension of calf DVT occurs in approximately 30%, with propagation occurring within 1 to 2 weeks of initial presentation.⁶

CLINICAL SCENARIO

A 55-year-old woman is referred to you with suspected DVT. She complains of pain, swelling, warmth, and redness of her right calf. She denies injury to the leg or previous DVT. She has been receiving intravenous combination chemotherapy for ovarian carcinoma that was diagnosed 6 months earlier. Extensive pelvic lymph node involvement, especially on the right side, was present at diagnosis, and you consider the possibility that her leg symptoms are due to extrinsic compression of the right iliac vein. However, no lymph nodes are palpable and a recent pelvic ultrasonographic examination showed a reduction in the previously demonstrated adenopathy. On physical examination, you find pitting edema, erythema, increased warmth of the right calf (diameter 3.5 cm greater than that of the left calf), and tenderness with palpation of the popliteal vein. You apply a clinical prediction rule⁶ and conclude that the probability of proximal DVT is high.

METHODS

Search Strategy

We conducted a MEDLINE search to retrieve all relevant articles pertaining to the clinical assessment of patients with suspected DVT. MEDLINE was searched from 1966 to April 1997 using Medical Subject Headings, EXP (explode) "thrombosis" (tw [textword]) and (EXP "physical examination" or EXP "diagnostic tests" or EXP "sensitivity and specificity") and EXP "phlebography." This was limited to human and English-language studies. One hundred fifteen articles were retrieved (available on request from the senior author); 68 articles that dealt with the diagnosis of DVT were selected for complete review. The bibliographies of the retrieved articles were examined for additional relevant articles. Only 5 studies provided information on the relationship between clinical findings and venographic confirmation of DVT.^{3,4,6,12,13} These studies were graded according to their methodologic quality with a standard scoring system.¹⁴ (See Table 1-7 for a summary of Evidence Grades and Levels.)

Principles of Diagnosis of DVT

The diagnostic assessment of patients with suspected DVT has evolved from reliance on clinical symptoms and signs alone to proof of DVT from objective diagnostic tests.

Table 18-1 Odds Ratios of Risk Factors for Deep Vein Thrombosis^a

Risk Factors	OR (95% CI)
Male sex	1.7 (1.4-2.0)
Age > 60 y	1.6 (1.3-1.9)
Cancer	2.4 (1.9-2.8)
Heart failure	1.8 (1.3-2.3)
Systemic lupus erythematosus	4.4 (3.1-5.5)
Lower limb arteriopathy	1.9 (1.3-2.5)

Abbreviations: CI, confidence interval; OR, odds ratio.

^aData are from Cogo et al.¹⁸

RESULTS

Clinical Assessment

Throughout the past 30 years, the clinical assessment in patients with suspected DVT has been refined and now includes a careful review of risk factors, symptoms, and physical signs.^{5,15-17} Risk factors for DVT include immobility, paralysis, recent surgery or trauma, malignancy, cancer chemotherapy, advancing age (ie, >60 years), family history of venous thromboembolism, pregnancy, and estrogen use.¹⁸ In a recent prospective cohort study, 426 consecutive outpatients referred by general practitioners to a tertiary-care thrombosis unit were assessed for DVT risk factors, and in approximately half of the patients with confirmed DVT, a major risk factor (immobility, trauma, or recent surgery) was present.¹⁸ The odds ratios (ORs) for other risk factors independently associated with the presence of DVT, including male sex, age greater than 60 years, cancer, heart failure, systemic lupus erythematosus, and lower limb arteriopathy, are presented in Table 18-1. Commonly reported symptoms in patients with suspected DVT include leg pain, swelling, and other signs, such as pitting edema, warmth, dilated superficial veins, and erythema.³⁻⁵ Unfortunately, these findings are neither sensitive nor specific for DVT and may be caused by other disease processes,^{5,15} such as leg trauma, cellulitis, obstructive lymphadenopathy, superficial venous thrombosis, postphlebitic syndrome, or Baker cysts.^{5,19} The ORs for these factors range from 1.6 to 4.3.¹⁸ Furthermore, DVT can coexist with each of these processes. For example, the finding of a Baker cyst on an ultrasonographic examination does not rule out the presence of DVT.¹⁹

Traditionally, the routine physical examination in patients with suspected DVT included a careful inspection of the leg, measurement of the leg circumference, and elicitation of Homans sign,²⁰ which refers to the development of pain in the calf or popliteal region on forceful and abrupt dorsiflexion of the ankle while the knee is flexed. Early studies evaluating the properties of individual physical signs such as these to diagnose DVT showed that they were inaccurate.^{3,4} In a study by O'Donnell et al,³ 102 patients with suspected DVT who presented to the outpatient departments of 2 tertiary-care hospitals underwent a clinical assessment and venography. A combination of clinical signs and symptoms that included tenderness, swelling, redness, and the assessment of Homans sign could not adequately differentiate patients with or without DVT. The sensitivity of the clinical examination in this study was 88% (95% confidence interval [CI], 77%-97%) and the specificity was only 30% (95% CI, 18%-40%). Haeger⁴ conducted a prospective study of 72 outpatients in a thrombosis clinic who were examined by 1 or 2 experienced surgeons and who then underwent venography. No differences in the presenting symptoms or physical signs were identified between those with or without venographically confirmed DVT. The sensitivity of the clinical examination in this study was 66% (95% CI, 50%-82%) and the specificity only 53% (95% CI, 38%-69%). In a study by Molloy et al,¹² 100 patients with a

Table 18-2 Frequency of Symptoms and Signs in Patients With Suspected Deep Vein Thrombosis (DVT)^a

Signs and Symptoms	Source					
	O'Donell et al, ³ Grade A, %		Hager, ⁴ Grade B, %		Molloy et al, ¹² Grade A, %	
	DVT+	DVT-	DVT+	DVT-	DVT+	DVT-
Pain	78	75	90	97	48	23
Tenderness	76	89	84	74	43	35
Edema	78	67	42	32	43	26
Homans sign	56	61	33	21	11	11
Swelling ^b	85	56	... ^c	...	41	39
Erythema ^b	24	38

Abbreviations: DVT+, those with DVT; DVT-, those without DVT.

^aThe DVT diagnosis was observed by venography.

^bSee Table 1-7 for a summary of Evidence Grades and Levels.

^cEllipses indicate data not reported.

clinical diagnosis of DVT who were referred to the radiology department of a general hospital were studied; the sensitivity of the clinical examination was 60% (95% CI, 45%-75%) and the specificity was 72% (95% CI, 60%-83%). Overall, these symptoms and signs occur in similar frequency in symptomatic patients with and without DVT (**Table 18-2**).

The results of these studies led to a shift away from the clinical examination to a heavy reliance on noninvasive objective tests for patients with suspected DVT. In a retrospective chart review by Landefeld et al¹³ of 354 inpatients and outpatients with suspected DVT who underwent venography, there were 5 clinical findings independently related to the presence of proximal DVT: swelling below the knee, swelling above the knee, recent immobility, cancer, and fever. These factors were determined by using multiple linear regression, were found to be significantly associated with the presence of proximal DVT in 236 patients, and then were confirmed in the remaining 119 patients. Overall, the sensitivity of a positive clinical examination (associated with the presence of 1 or more independent predictors) was 96% (95% CI, 92%-100%) and the specificity was 20% (95% CI, 15%-25%). The frequency of signs and symptoms seemed to predict the presence of proximal DVT when the absence of any findings was associated with less than a 5% chance of proximal DVT, and the presence of 2 or more clinical findings was associated with a 46% chance of proximal DVT. This was the first study to demonstrate the potential role of a clinical prediction guide in patients with suspected DVT. The likelihood ratio (LR) estimates for the clinical assessment according to the 4 studies described above are shown in **Table 18-3**.

Recall that an LR expresses the odds that a given finding on the medical history or physical examination would occur in a patient with the target disorder as opposed to a patient without it. Given an LR of more than 1.0, the probability of disease (in this case, DVT) increases when the finding is present because the finding is more likely among the patients with the disease than among those without. When the LR is less

Table 18-3 Likelihood Ratio for Clinical Assessment in Patients With Suspected Deep Vein Thrombosis Compared With Venographic Result^a

Source	Positive Clinical Assessment Result for DVT (95% CI)	Negative Clinical Assessment Result for DVT (95% CI)
O'Donnell et al ³	1.2 (1.0-1.5)	0.40 (0.17-0.96)
Haeger ⁴	1.4 (0.95-2.2)	0.64 (0.34-1.1)
Molloy et al ¹²	2.1 (1.3-3.5)	0.55 (0.36-0.80)
Landefeld et al ¹³	1.2 (1.1-1.3)	0.21 (0.08-0.54)

Abbreviations: CI, confidence interval; DVT, deep vein thrombosis.

^aPositive clinical assessment result was defined as 1 or more clinical factors; negative clinical assessment, absence of clinical factors.

than 1.0, the probability of disease decreases because the finding is less likely to occur among patients with the disease than among those without.²¹

Objective Assessment

Venography is the reference standard for the diagnosis of DVT, and it is highly accurate for both proximal and calf DVT.²² However, venography is invasive, expensive, technically inadequate in about 10% of patients (either because of an inability to cannulate a vein or because of lack of adequate visualization of the deep veins), and may induce DVT in approximately 3% of patients.²³ This led to the evaluation and validation of 2 noninvasive tests: impedance plethysmography and compression ultrasonography. These tests have proved to be sensitive to proximal but not to calf vein thrombosis.

Impedance plethysmography reliably detects occlusive thrombi of the proximal veins (popliteal, femoral, or iliac veins) but is less reliable at detecting nonocclusive proximal DVT and is insensitive to calf DVT.²⁴⁻²⁷ Impedance plethysmography does not allow direct visualization of the veins but suggests that DVT is present when significant outflow obstruction is present, particularly in the absence of a comorbid condition that might cause a false-positive result (eg, extrinsic

venous compression or elevated central venous pressure). Although studies before 1990 reported that impedance plethysmography detected more than 90% of proximal DVT, more recent studies reported sensitivities for proximal DVT of about 70%.²⁸⁻³⁰ This apparent decrease in sensitivity is probably caused by changes in referring patterns to specialty centers with a strong interest in DVT.³¹

Compression ultrasonography assesses compressibility of the femoral and popliteal veins and is highly sensitive and specific for detecting proximal DVT (noncompressibility is diagnostic of DVT, whereas compressibility rules out DVT).^{6,32-34} Neither impedance plethysmography nor compression ultrasonography reliably detects isolated calf vein thrombosis.³⁵ Although the specificity of compression ultrasonography and impedance plethysmography for DVT remains high in both symptomatic and asymptomatic patients, the sensitivity declines dramatically when impedance plethysmography and compression ultrasonography are used to evaluate asymptomatic patients (ie, 22% and 58%, respectively) vs symptomatic patients (ie, 96% and 96%, respectively).³⁶ Several diagnostic algorithms using serial compression ultrasonography or impedance plethysmography have been evaluated and validated in large clinical trials.^{24,27,32-34,37-42} Although compression ultrasonography appears to be more accurate than impedance plethysmography, serial testing with either is acceptable in patients with suspected DVT.^{37,43} Therefore, as most clinicians consider clinically important proximal DVT excluded by normal impedance plethysmography or compression ultrasonography on the day of presentation, anticoagulants can be safely withheld in such patients, because the probability of experiencing proximal DVT is less than 2% in the following 3 months.⁴⁴ If the initial test results are normal, repeated testing during the next 5 to 7 days is recommended; if they become abnormal during this period, extending proximal DVT is likely and an anticoagulant therapy should be initiated. However, impedance plethysmography and compression ultrasonography have limitations too, such as availability and the inconvenience and expense of repeated testing.

Recently, the D-dimer assays have been demonstrated to be useful adjuncts to noninvasive testing for suspected DVT because they are highly sensitive and therefore have high negative predictive values.⁴⁹ D-dimer⁴⁵⁻⁴⁷ is formed when cross-linked fibrin contained within a thrombus is proteolyzed by plasmin. Various D-dimer assays are available, including enzyme-linked immunosorbent assays (ELISA), latex agglutination assays, and a whole blood agglutination test.⁴⁶ The whole blood agglutination assay appears to be best for exclusion of DVT because it is suitable for individual testing (unlike ELISA) and has high sensitivity and reasonable specificity. Recent studies show that DVT can be reliably excluded in patients with suspected DVT who have a normal impedance plethysmograph result and a normal D-dimer result (using a high-sensitivity whole blood assay) and that such results occur in about two-thirds of patients.⁴⁵ This supports the role of the assay as a simple and rapid adjunct to noninvasive tests for the exclusion of clinically important DVT.^{45,46} For a summary of diagnostic algorithms for patients with suspected DVT, see Table 18-4.

Clinical Prediction Guide

Recently, the clinical assessment of patients with suspected DVT was reevaluated. This was sparked by 2 observations that many patients with a high pretest probability (using clinical judgment) and a normal impedance plethysmograph had proximal DVT,²⁸ and that the pretest probability of patients had an important influence on diagnosing PE, a closely related disease. For example, in patients with a low pretest probability and a high-probability lung scan, the prevalence of PE was approximately 50% to 60%.⁴⁸ These results generated the hypothesis that when pretest probability and further tests are concordant, DVT can be ruled in or out, whereas when they are discordant, further tests are necessary.

Development of a Clinical Prediction Guide

Recently, a clinical prediction guide that seeks to standardize the estimation of the pretest probability among clinicians was developed⁶ and is described below. This model enables clini-

Table 18-4 Interpretation of Test Results in Patients With Suspected Initial Deep Vein Thrombosis

Tests	Results		
	Venography	Compression Ultrasonography	Impedance Plethysmography
Diagnose DVT	Intraluminal filling defect in at least 2 projections	Noncompressibility of the femoral or popliteal vein	Abnormal impedance plethysmography and a moderate to high clinical probability of DVT
Exclude clinically important DVT	Normal venogram result	Normal compressibility of proximal venous segments combined with a low clinical pretest probability, or normal serial compression ultrasonographic examination result	Normal impedance plethysmography combined with a normal D-dimer or normal serial impedance plethysmography
Nondiagnostic for DVT	Technically inadequate study in which all deep veins are not adequately visualized	Noncompressibility of deep veins of the calf	Abnormal impedance plethysmography combined with a low clinical suspicion

Abbreviation: DVT, deep vein thrombosis.

cians to reliably stratify patients with suspected DVT into high-, moderate-, or low-probability groups by following uniform criteria. After a review of the literature^{3,4,8,18} and input from experienced thrombosis investigators, categories deemed to be important in the estimation of a patient's pretest probability were considered and categorized as follows: signs and symptoms of DVT, risk factors for DVT, and the presence or absence of diagnoses that were deemed at least as likely as DVT to explain the patient's symptoms. These include musculoskeletal injuries, cellulitis, and prominent lymphadenopathy of the inguinal area. The clinical prediction guide uses a scoring system that combines important symptoms and signs, risk factors for DVT, and the presence or absence of an alternative diagnosis. The results stratify patients with suspected DVT into low-, moderate-, or high-probability groups. The original clinical prediction guide was initially developed in a training set of 100 outpatients at a thrombosis referral center at McMaster University, Hamilton, Ontario, Canada, who presented with suspected DVT. All patients underwent venography, and a simple regression model determined the relative importance of individual and various clusters of factors to predict the probability that a patient had DVT.

The clinical prediction guide was then prospectively validated in a test set of 529 patients who presented with suspected DVT to 3 tertiary-care referral centers: 2 in Hamilton and 1 in Padua, Italy.⁶ Clinicians recorded their assessment of pretest probability of DVT, and then all patients underwent venography and compression ultrasonographic examination. This model cannot be applied to certain subgroups of patients who were excluded from the study, such as those with previous venous thromboembolism, those with concomitantly suspected PE, pregnant women, or patients receiving treatment with anticoagulants. With the clinical model, eligible patients were initially stratified into low-, moderate-, or high-pretest-probability groups.

Although individual physical findings on their own are not predictive of DVT, when specific physical signs are incorporated into the clinical prediction guide they contribute to the generation of the pretest probability of DVT. In Table 18-5, the physical signs and the scoring system of the clinical prediction guide are outlined. The physical signs classified as major points include localized tenderness to palpation along the distribution of the deep venous system; thigh and calf swelling, indicating that the entire leg has an increased diameter compared with the asymptomatic side; and calf swelling, in which the calf is measured approximately 10 cm below the tibial plateau (at the tibial tuberosity) and swelling is considered present if the difference between calf diameters is more than 3 cm. Minor points include the presence of a unilateral pitting edema of the leg with standard assessment measures, the presence of dilated superficial veins (nonvaricose) that persist with elevation in the lower limb or if present in any new pattern in the groin region on the symptomatic leg only, and the presence of diffuse or streaking erythema.

The test set confirmed that the clinical model could reliably classify patients into high-, moderate-, and low-probability groups. The prevalence of all DVT (proximal and calf), using the venogram as the criterion standard in patients who were classified by the clinical model into the high-probability strata, was 85% compared with 33% in the moderate-probability and 5% in the low-probability categories. The positive

LRs for the high-, moderate-, and low-risk categories are 16 (95% CI, 9.3-28), 1.3 (95% CI, 1.0-1.7), and 0.2 (95% CI, 0.1-0.3), respectively. The specificity of compression ultrasonography to detect proximal DVT in all strata was between 98% and 100%. When interpreted in conjunction with pretest probability, the ability of compression ultrasonography to reliably diagnose DVT decreased as the pretest probability declined. The sensitivities of compression ultrasonography in the high, moderate, and low strata were 94%, 83%, and 80%, respectively. The corresponding LRs for compression ultrasonography in pretest probability strata are provided in Table 18-6. By combining pretest probability and compression ultrasonography results, the posttest probabilities of DVT for each possible combination of results were generated. In the high-pretest-probability strata, an abnormal compression ultrasonogram result led to a 100% posttest probability; in the moderate strata, a 96% posttest probabili-

Table 18-5 Estimation of Pretest Probability of Deep Vein Thrombosis Using the Clinical Model^a

Major Points
Active cancer (treatment ongoing or within previous 6 mo or palliative)
Paralysis, bedridden > 3 days, or major surgery within 4 wk
Localized tenderness along the distribution of the deep venous system in calf or thigh
Thigh and calf swollen (should be measured)
Calf swelling by > 3 cm when compared with the asymptomatic leg (measured 10 cm below the tibial tuberosity)
Strong family history of DVT (>2 first-degree relatives with history of DVT)
Minor Points
History of recent trauma (≤ 60 d to the symptomatic leg)
Pitting edema in symptomatic leg only
Dilated superficial veins (nonvaricose) in symptomatic leg only
Hospitalization within previous 6 mo
Erythema

Abbreviation: DVT, deep vein thrombosis.

^aItems excluded from the model are age, duration of symptoms, sex, obesity, presence of varicose veins, a palpable cord, and Homans sign. Scoring method: high probability if ≥ 3 major points and no alternative diagnosis, ≥ 2 major points and ≥ 2 minor points and no alternative diagnosis; low probability if 1 major point and ≤ 2 minor points and an alternative diagnosis, 1 major point and ≤ 1 minor point and no alternative diagnosis, 0 major points and ≤ 3 minor points and an alternative diagnosis, 0 major points and ≤ 2 minor points and no alternative diagnosis; and moderate probability for all other combinations.

Table 18-6 Likelihood Ratios for Ultrasonographic Results by Clinical Probability Strata

Pretest Probability	Ultrasonography	LR+ (95% CI)
High	Abnormal	∞ (3- ∞)
Moderate	Abnormal	72 (13-412)
Low	Abnormal	34 (14-76)
High	Normal	0.06 (0.03-0.16)
Moderate	Normal	0.17 (0.07-0.34)
Low	Normal	0.20 (0.06-0.52)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio.

ity; and in the low strata, a 63% posttest probability. In patients whose compression ultrasonogram result was normal, the posttest probabilities of DVT in the high, moderate, and low strata were 24%, 5%, and less than 1%, respectively.

The original clinical prediction guide was recently simplified with stepwise logistic regression and reevaluated.⁴⁹ Recent trauma, family history, erythema, and hospitalization within the previous 6 months did not remain in the simplified model, which, in combination with compression ultrasonography, was prospectively tested in 593 patients with suspected DVT who were referred to tertiary-care thrombosis clinics⁴⁹ (Table 18-7). Similar to the original clinical prediction guide, the simplified guide was able to reliably stratify patients into high-, moderate-, or low-probability groups, with corresponding prevalences of DVT of 75% (95% CI, 63%-81%), 17% (95% CI, 12%-23%), and 3% (95% CI, 1.7%-5.9%), respectively.

These data support the use of a clinical prediction guide to simplify the diagnostic approach for patients with suspected

Table 18-7 Simplified Clinical Model^a

Clinical Characteristic	Score
Active cancer (treatment ongoing or within previous 6 mo or palliative)	1
Paralysis, paresis, or recent plaster immobilization of the lower extremities	1
Recently bedridden for > 3 d of major surgery within 4 wk	1
Localized tenderness along the distribution of the deep venous system	1
Entire leg swelling	1
Calf swelling by > 3 cm compared with the asymptomatic leg (measured 10 cm below the tibial tuberosity) ^b	1
Pitting edema (greater in the symptomatic leg)	1
Collateral superficial veins (nonvaricose)	1
Alternative diagnosis as likely as or greater than that of DVT	-2

Abbreviation: DVT, deep vein thrombosis.

^aScoring method: high probability if score is 3 or higher, moderate if score is 1 or 2, and low if score is 0 or lower.

^bIn patients with symptoms in both legs, the more symptomatic leg was used.

DVT (Figure 18-1). In patients with a high or moderate pretest score who have an abnormal compression ultrasonogram result, DVT can be reliably diagnosed (LR+, ∞ and 72, respectively) and treatment should be initiated. In patients with a low pretest probability of DVT who have a normal compression ultrasonogram result (LR-, 0.2), DVT can be reliably ruled out without further testing. For patients with discordant results (ie, high pretest probability and normal compression ultrasonogram result, or low pretest probability and an abnormal compression ultrasonogram result), further testing is recommended (ie, venography or serial compression ultrasonography). Patients with a moderate pretest probability and a normal ultrasonogram result have a 5% probability of having DVT, and a repeated compression ultrasonographic examination in 7 days is recommended.

CLINICAL SCENARIO—RESOLUTION

The patient described in the “Clinical Scenario” section is a 55-year-old woman who presents with suspected DVT. Using the clinical prediction guide checklist found in Table 18-5, you determine that she has 5 clinical features predictive of DVT: a diagnosis of active cancer, calf swelling, erythema, localized tenderness along the popliteal vein, and pitting edema of the symptomatic leg. Although the possibility of enlarging pelvic lymph nodes in the right inguinal area offers an alternative diagnosis, you observe that a recent pelvic ultrasonographic report indicates that these nodes have shrunk, rendering this a less likely alternative diagnosis. Therefore, with 5 clinical features of DVT and no convincing alternative diagnosis, following the approach of the clinical prediction guide you conclude that she has a high clinical probability of experiencing acute DVT. The next step is to perform a compression ultrasonographic examination, and, if the results are abnormal, the posttest probability of DVT being present approaches 100%. However, if the ultrasonogram result is normal (ie, showing normal compressibility of the proximal veins), the posttest probability is approximately 24%, and further testing with venography would be required.

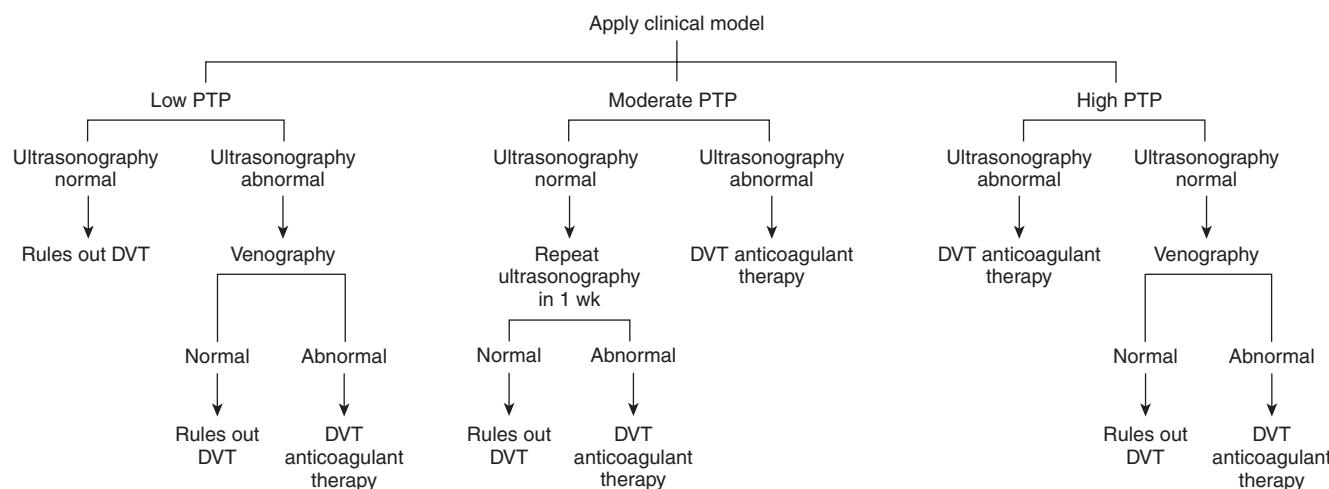


Figure 18-1 Suggested Diagnostic Approach in Patients With Suspected DVT

Abbreviations: DVT, deep vein thrombosis; PTP, pretest probability.

CONCLUSIONS

Although physical findings of patients with suspected DVT are not useful on their own, a clinical prediction guide that includes factors from both the medical history and physical examination is able to assist in the diagnosis of DVT. When used in combination with noninvasive tests, such as compression ultrasonography, it can simplify and reduce the expense of management strategies.

THE BOTTOM LINE

Individual symptoms and signs on their own are not useful to diagnose DVT. However, a systematic review of patients' risk factors, symptoms, and physical signs allows the clinician to reliably determine the pretest probability that a patient has DVT. This strategy, in combination with the results of noninvasive diagnostic test results, guides further diagnostic testing and treatment strategies.

Author Affiliations at the Time of the Original Publication

Department of Medicine (Drs Anand, Hunt, Brill-Edwards, Cook, and Ginsberg) and Department of Clinical Epidemiology (Dr Cook), McMaster University, Hamilton, Ontario, Canada; and Department of Medicine, Ottawa Civic Hospital, Ottawa, Ontario, Canada (Dr Wells).

Acknowledgments

We acknowledge the contributions of Brian Gage, MD, MSc, and Gino Merli, MD, for their critical review of the manuscript.

Dr Anand is a recipient of a Research Fellowship of the Heart and Stroke Foundation of Canada, Ottawa, Ontario; Dr Ginsberg, a Career Investigator Award of the Heart and Stroke Foundation of Ontario, Toronto, Ontario, Canada; Dr Wells, a Research Scholarship of the Heart and Stroke Foundation of Canada, Ottawa, Ontario; and Dr Cook, the Ontario Ministry of Health Career-Investigator Award, Toronto, Ontario, Canada.

REFERENCES

- Hirsh J, Hoak J. Management of deep vein thrombosis and pulmonary embolism. *Circulation*. 1996;93(12):2212-2245.
- Gunintini C, Di Ricco G, Marini C, Melillo E, Palla A. Epidemiology. *Chest*. 1995;107:3S-9S.
- O'Donnell T, Abbott W, Athanasoulis C, Millan V, Callow A. Diagnosis of deep venous thrombosis in the outpatient by venography. *Surg Gynecol Obstet*. 1980;150(1):69-74.
- Haeger K. Problems of acute deep venous thrombosis, I: the interpretation of signs and symptoms. *Angiology*. 1969;20(4):219-223.
- Hull RD, Raskob GE, Leclerc J, Jay RM, Hirsh J. The diagnosis of clinically suspected venous thrombosis. *Clin Chest Med*. 1984;5(3):439-452.
- Wells P, Hirsh J, Anderson D, Lensing A, Foster G, Kearon C. Accuracy of clinical assessment of deep-vein thrombosis. *Lancet*. 1995;345(8961):1326-1330.
- Hirsh J, Raschke R, Warkentin T. Heparin: mechanism of action, pharmacokinetics, dosing considerations, monitoring, efficacy, and safety. *Chest*. 1995;108(4 suppl):258S-275S.
- Vine H, Hillman B, Hessel S. Deep venous thrombosis: predictive value of signs and symptoms. *Am J Radiol*. 1981;136(1):167-171.
- Cogo A, Lensing A, Prandoni P, Hirsh J. Distribution of thrombosis in patients with symptomatic deep vein thrombosis: implications for simplifying the diagnostic process with compression ultrasound. *Arch Intern Med*. 1993;153(24):2777-2780.
- Hull RD, Hirsh J, Carter CJ, Jay RM, Ockelford PA, Buller HR. Pulmonary angiography, ventilation lung scanning, and venography for clinically suspected pulmonary embolism with abnormal perfusion lung scanning. *Ann Intern Med*. 1983;98(6):891-899.
- Hyers TM, Hull RD, Weg JG. Antithrombotic therapy for venous thromboembolic disease. *Chest*. 1995;108(4 suppl):335S-351S.
- Molloy W, English J, O'Dwyer R, O'Connell J. Clinical findings in the diagnosis of proximal deep venous thrombosis. *Ir Med J*. 1982;75(4):119-120.
- Landefeld CS, McGuire E, Cohen A. Clinical findings associated with acute proximal deep vein thrombosis: a basis for quantifying clinical judgment. *Am J Med*. 1990;88(4):382-388.
- Holleman D, Simel D. Does the clinical examination predict airflow limitations? *JAMA*. 1995;273(4):313-319.
- Hirsh J, Hull RD, Raskob GE. Clinical features and diagnosis of venous thrombosis. *J Am Coll Cardiol*. 1986;8(6 suppl B):114B-127B.
- Browse N. Deep vein thrombosis: diagnosis. *BMJ*. 1969;4(5684):676-678.
- McLachlin J, Richard T, Paterson JC. An evaluation of clinical signs in the diagnosis of venous thrombosis. *Arch Surg*. 1962;85:738.
- Cogo A, Bernardi E, Prandoni P, Girolami B, Novanta F, Simioni P. Acquired risk factors for deep-vein thrombosis in symptomatic outpatients. *Arch Intern Med*. 1994;154(2):164-168.
- Simpson PG, Bark M, Robinson PJ, Losowsky MS. Prospective study of thrombophlebitis and pseudothrombophlebitis. *Lancet*. 1980;1(8164):331-333.
- Barnes RW, Wu KK, Hoak JC. Fallibility of the clinical diagnosis of venous thrombosis. *JAMA*. 1975;234(6):605-607.
- Sackett DL. A primer on the precision and accuracy of the clinical examination. *JAMA*. 1992;267(19):2638-2644.
- Rabinov K, Paulin S. Roentgen diagnosis of venous thrombosis in the leg. *Arch Surg*. 1972;104(2):134-144.
- Lensing AWA, Prandoni P, Buller H, Casara D, Cogo A, ten Cate JW. Lower extremity venography with iohexol: results and complications. *Radiology*. 1990;177(2):503-505.
- Hull RD, Hirsh J, Carter CJ. Diagnostic efficacy of impedance plethysmography for clinically-suspected deep-vein thrombosis. *Ann Intern Med*. 1985;102(1):21-28.
- Prandoni P, Lensing AWA, Huisman MV, et al. A new computerized impedance plethysmograph: accuracy in the detection of proximal deep-vein thrombosis in symptomatic outpatients. *Thromb Hemost*. 1991;65(3):229-232.
- Hull RD, van Acken WG, Hirsh J, et al. Impedance plethysmography using the occlusive cuff technique in the diagnosis of venous thrombosis. *Circulation*. 1976;53(4):696-700.
- Huisman MV, Buller HR, ten Cate JW, Heijermans HSF, van der Laan J, van Maanen DJ. Management of clinically suspected acute venous thrombosis in outpatients with serial impedance plethysmography in a community hospital setting. *Arch Intern Med*. 1989;149(3):511-513.
- Anderson DR, Lensing AW, Wells PS, Levine MN, Weitz JI, Hirsh J. Limitations of impedance plethysmography in the diagnosis of clinically suspected deep-vein thrombosis. *Ann Intern Med*. 1993;118(1):25-30.
- Prandoni P, Lensing AWA, Buller HR, et al. Failure of computerized impedance plethysmography in the diagnostic management of patients with clinically suspected DVT. *Thromb Hemost*. 1991;65:233.
- Ginsberg JS, Wells PS, Hirsh J, et al. Reevaluation of the sensitivity of impedance plethysmography for the detection of proximal deep vein thrombosis. *Arch Intern Med*. 1994;154(17):1930-1933.
- Cogo A, Prandoni P, Villalta S, Polistena D, Bernardi E, Simioni P. Changing features of proximal vein thrombosis over time. *Angiology*. 1994;45(5):377-382.
- Heijboer H, Cogo A, Buller HR, Prandoni P, ten Cate JW. Detection of deep vein thrombosis with impedance plethysmography and real time compression ultrasonography in hospitalized patients. *Arch Intern Med*. 1992;152(9):1901-1903.
- White RH, McGahan JP, Daschbach MM, Hartling RP. Diagnosis of deep-vein thrombosis using duplex ultrasound. *Ann Intern Med*. 1989;111(4):297-304.
- Lensing AWA, Prandoni P, Brandjes P, Huisman MV, Vigo M, Tomasella G. Detection of deep vein thrombosis by real time B-mode ultrasonography. *N Engl J Med*. 1989;320(6):342-345.

35. Mattos MA, Londrey GL, Leutz DW, et al. Color-flow duplex scanning for the surveillance and diagnosis of acute deep venous thrombosis. *J Vasc Surg.* 1992;15(2):366-376.
36. Ginsberg JS, Caco CC, Brill-Edwards PA, et al. Venous thrombosis in patients who have undergone major hip or knee surgery: detection with compression US and impedance plethysmography. *Radiology.* 1991;181(3): 651-654.
37. Heijboer H, Buller H, Lensing A, Turpie A, Colly L, ten Cate J. A comparison of real-time compression ultrasonography with impedance plethysmography for the diagnosis of deep-vein thrombosis in symptomatic outpatients. *N Engl J Med.* 1993;329(19):1365-1369.
38. Hull R, Hirsh J, Sackett DL, Powers P. Combined use of leg scanning and impedance plethysmography in suspected venous thrombosis: an alternative to venography. *N Engl J Med.* 1977;296(26):1497-1500.
39. Huisman MV, Buller HR, ten Cate JW, Vreeken J. Serial impedance plethysmography for suspected deep venous thrombosis in outpatients: the Amsterdam General Practitioner Study. *N Engl J Med.* 1986;314(13): 823-828.
40. Cogo A, Lensing AWA, Koopman MMW. Compression ultrasound for the diagnostic management of clinically suspected deep-vein thrombosis. *Thromb Hemost.* 1995;73:1098.
41. Hull R, Hirsh J, Sackett DL. Replacement of venography in suspected venous thrombosis by impedance plethysmography and 125I-fibrinogen leg scanning. *Ann Intern Med.* 1981;94(1):12-15.
42. Cogo A, Lensing AWA, Prandoni P, Buller HR, Girolami A, ten Cate JW. Comparison of real-time B-mode ultrasonography and Doppler ultrasound with contrast venography in the diagnosis of venous thrombosis in symptomatic outpatients. *Thromb Hemost.* 1993;70(3):404-407.
43. Wells PS, Hirsh J, Anderson DR, et al. Comparison of the accuracy of impedance plethysmography and compression ultrasonography in outpatients with clinically suspected deep vein thrombosis. *Thromb Hemost.* 1995;74(6):1423-1427.
44. Ginsberg JS. Management of venous thromboembolism. *N Engl J Med.* 1996;335(24):1816-1828.
45. Wells PS, Brill-Edwards P, Stevens P, et al. A novel and rapid whole-blood assay for D-dimer in patients with clinically suspected deep vein thrombosis. *Circulation.* 1995;91(8):2184-2187.
46. Dale S, Gogstad GO, Brosstad F, et al. Comparison of 3 D-dimer assays for the diagnosis of DVT: ELISA, latex, and an immunofiltration assay (NucoCard D-dimer). *Thromb Hemost.* 1994;71(3):270-274.
47. Bounameaux H, De Moerloose P, Perrier A, Raber G. Plasma measurement of D-dimer as diagnostic aid in suspected venous thromboembolism: an overview. *Thromb Hemost.* 1994;71(1):1-7.
48. PIOPED Investigators. The value of the ventilation/perfusion scan in acute pulmonary embolism. *JAMA.* 1990;26(20):2753-2759.
49. Wells P, Anderson DR, Bormanis J, et al. Value of assessment of pretest probability of deep-vein thrombosis in clinical management. *Lancet.* 1997;350(9094):1795-1798.

CHAPTER 18

Does This Patient Have Deep Vein Thrombosis?

Philip S. Wells, MD, MSc

Carolyn Owen, MD

Steve Doucette, MSc

Dean Fergusson, PhD

Huyen Tran, MD

CLINICAL SCENARIO

A 60-year-old man referred with suspected deep vein thrombosis (DVT) cut the plantar surface of his left foot on glass 10 days ago and has been resting in bed. He presents with left leg pain and mild calf swelling, redness, and heat. There is no history of a DVT or known family history of venous thromboembolism. Physical examination shows the patient is febrile and has pitting edema of the left calf. The calf erythema is hot, tender, and well demarcated. Enlarged left inguinal lymph nodes are present. He has longstanding diabetes mellitus, and the diagnoses that seem most likely are cellulitis and DVT. Can a clinical probability estimate of DVT reliably determine a pretest probability that can be used in decision making?

CLINICAL EVALUATION AND CLINICAL PREDICTION RULES

DVT occurs frequently, with an estimated annual incidence of 0.1% in white populations,^{1,2} creating considerable morbidity. Complications include postphlebitic syndrome and chronic thromboembolic pulmonary hypertension, whereas pulmonary embolism (PE) causes death in 1% to 8% of affected patients despite treatment.³⁻⁵ Although anticoagulant therapy decreases the risk of recurrent thrombosis, the treatment also increases the risk of major hemorrhage and other potentially serious consequences, such as heparin-induced thrombocytopenia. Therefore, diagnostic strategies must correctly diagnose DVT when present and safely rule out DVT when absent. The desire to not miss a patient with DVT, combined with the large number of nonspecific signs and symptoms, makes DVT part of the differential diagnosis in most patients presenting with leg pain or swelling. Unfortunately, the nonspecific signs and symptoms force clinicians to investigate many patients who do not have DVT. In the past, clinical assessment was not quantified in the diagnostic assessment in patients with suspected DVT, and before 1995, the approach was for all patients with suspected DVT to undergo ultrasonography.^{6,7} This approach was inefficient because most patients with suspected DVT did not have the disorder (DVT rates ranging from 10% to 25%).⁷⁻⁹ Because imaging for calf DVT is relatively inaccurate and often inadequate,^{10,11} serial testing in which only the proximal veins were evaluated and testing repeated 1 week later in the case of negative results was the standard. Several studies performed in the last decade successfully incorporated clinical assessment into the diagnostic approach.

In a previous Rational Clinical Examination article, we outlined how categorizing patients as having a low clinical probability for DVT eliminates the need for serial testing, whereas categorizing patients as having a high clinical probability

selects those in whom a negative ultrasonographic result may be a false negative.¹² We also emphasized that false-positive ultrasonographic results were most likely when patients had a low clinical probability for DVT. The clinical prediction rule⁸ described in that article had not been widely evaluated. We conducted a new systematic review to determine the accuracy of the same clinical prediction rule for DVT.

The incorporation of D-dimer testing into diagnostic algorithms has simplified the treatment of a patient presenting with suspected DVT.¹³⁻¹⁶ Clinical trials demonstrate safe, feasible, and validated approaches for the treatment of patients with suspected DVT. However, it is also clear that D-dimer assays differ with respect to sensitivity and specificity. Recent meta-analyses summarize the accuracy of various D-dimer assays compared with gold standard imaging tests for DVT.^{17,18}

Diagnostic algorithms work by combining the pretest probability estimate (or clinical suspicion) with the likelihood ratio (LR) of a diagnostic test result, providing an accurate probability of disease after testing.¹⁹ Given the consequences of failing to detect DVT, a strategy that produces probabilities of 1% or less after testing should provide reassurance that additional tests are unnecessary. The combination of a low or unlikely clinical probability estimate with a negative D-dimer result safely rules out DVT.¹³ The following are not clear: whether the clinical prediction rule (eg, Wells et al¹³) can be used reliably across a broad range of at-risk population; what an estimated pooled risk of DVT is in each pretest category; and how pretest clinical probability estimates should be used with different D-dimer assays. To date, 3 studies have evaluated the literature on clinical prediction rules for the diagnosis of DVT, but all have limitations.²⁰⁻²² Specifically, they included studies and data that either did not use the model or used the model incorrectly by including patients with previous DVT (the most recent changes to the model include a point for patients with previous DVT). Indeed, Goodacre et al²² report that exclusion of persons with a history of thromboembolism is associated with improved diagnostic performance of the model by Wells et al¹³; however, they did not report summary prevalence data, one article reported only events rates in follow-up, and none reported LR data in combination with D-dimer testing. We conducted a systematic review to determine the accuracy of clinical prediction rules for DVT and D-dimer assays in conjunction with the clinical probability estimate.

METHODS

Study Identification

We searched for English- and French-language clinical studies that used a clinical prediction model or clinical assessment in the DVT diagnostic process. To evaluate the role of D-dimer, we also sought studies that used D-dimer in combination with clinical assessment. Published studies were identified by searching MEDLINE from January 1, 1990, to July 1, 2004, using the Medical Subject Headings “venous thrombosis” or “thrombophlebitis,” “fibrin or fibrinogen degradation products,” and “predictive value of tests,” and key words “DVT,” “D-dimer,” “diagnosis,” “sensitivity,” “specificity,” “clinical probability,” “clinical model,”

or “decision rule.” We supplemented the MEDLINE search by scrutinizing the reference lists of all articles selected for inclusion, review articles retrieved, and review of our own reference library of more than 4200 articles.

Study Selection

To be included in the review, all of the following criteria were required: (a) enrollment of consecutive outpatients with symptoms and signs of suspected DVT; (b) prospective trial design involving a minimum 3-month follow-up; (c) objective documentation of all venous thromboembolic events (DVT and PE); (d) exclusion of patients with previous DVT unless the clinical model adjusted for the history of DVT or the reviewers could make that adjustment; (e) assessment of patients with a validated clinical rule to estimate the clinical probability of DVT before D-dimer testing or diagnostic imaging; (f) performance of D-dimer testing before other diagnostic tests (although D-dimer testing was not a requirement for study inclusion); (g) available data on the prevalence of DVT in at least 1 of the 3 risk estimate categories (low, moderate, or high); (h) evaluation of proximal DVT; and (i) study quality graded A or B with the scheme previously appearing in The Rational Clinical Examination series, adapted from Holleman and Simel²³ (see Table 1-7) as shown:

Level 1: Independent, blinded comparison of symptom or sign results with a criterion standard of diagnosis among a large number of consecutive patients (≥ 300) with suspected DVT.

Level 2: Independent, blinded comparison of symptom or sign results with a criterion standard of diagnosis among consecutive patients (< 300) with suspected DVT.

Data Extraction

Two authors independently reviewed and abstracted data for determining prevalence of DVT in low-, moderate-, and high-clinical-probability groups; sensitivity and specificity; and LRs of D-dimer testing in each of the 3 clinical probability groups.

Statistical Analysis

Data were imported into the Comprehensive Meta-Analysis software program version 2.197 (Biostat Inc, Englewood, New Jersey) and analyzed with a random-effects model.

For each study, the overall prevalence of DVT and the prevalence among patients with low, moderate, or high clinical probability estimate were calculated. We confirmed the sensitivity and specificity and 95% confidence intervals (CIs) for each study that included D-dimer testing. The positive and negative likelihood ratios (LR+ and LR-) for each clinical probability estimate according to the D-dimer subset were calculated. An LR+ is a measure of how strongly a positive result increases the odds of disease and an LR- is measure of how well a negative result decreases the odds of disease. The easiest way to interpret LRs is to keep in mind that the likelihood of a disease outcome increases when the LR is greater than 1, the likelihood of disease decreases if the LR is less than 1, and an LR close to 1 does not change the likelihood. We also calcu-

lated the pooled LR because, unlike diagnostic odd ratios (ORs), the LRs can be used for clinical decisions.

Studies were grouped into 2 subsets, depending on the accuracy of the D-dimer that was used (ie, high sensitivity and moderate sensitivity, according to Stein et al¹⁸), and the same calculations were performed. Diagnostic ORs were calculated with correction for 100% sensitivities by adding 0.5 to each cell of the 2 × 2 table.^{24,25} The diagnostic OR is a single indicator of diagnostic test performance, reflecting its accuracy. With the random-effects model, the pooled estimates for the overall diagnostic OR as well as for the 2 subsets of D-dimer assays were calculated. For the 2 subsets of D-dimer assays, we evaluated differences between the sensitivity and specificity of the assays, between the low- and moderate-clinical-probability groups, and between the moderate- and high-pretest-probability groups with a χ^2 test.

RESULTS

After reviewing all titles and abstracts, we identified 67 of 274 articles for further review. Of the 67 articles, 14 met the

inclusion criteria involving 8239 patients (Table 18-8).^{9,13-15,26-35} The only studies eligible used the Wells clinical prediction rule (Table 18-9). One study reported D-dimer data on an earlier study, so it was not included in the calculation of prevalence.²⁷ Twelve of the 14 studies evaluating 5690 patients incorporated D-dimer testing into the diagnostic algorithm.^{9,13-15,27-34}

Does the Clinical Prediction Rule Accurately Categorize the Pretest Probability Estimate?

To be useful, the clinical probability estimate for DVT must be reproducible. Put another way, when the same patient or different patient populations presenting with suspected DVT are assessed, the clinical prediction rule should yield similar estimates for the risk of DVT. All studies included in this systematic review used the same clinical prediction rule. The pooled prevalence of DVT in the studies included in this meta-analysis was 19% (95% CI, 16%-23%). The pooled prevalence of DVT in the low-, moderate-, and high-clinical-probability groups was 5.0% (95% CI, 4.0%-8.0%), 17% (95% CI, 13%-23%), and 53% (95% CI, 44%-61%), respec-

Table 18-8 Summary of Studies of Deep Vein Thrombosis Diagnosis Involving Clinical Prediction Rule With or Without D-dimer Testing in Outpatients

Source, y	Evidence Quality Level	Outpatient Population	Had Ultrasonography, %	Requiring Serial Ultrasonography, %	D-dimer Assay	Score	Previous DVT Excluded	Prevalence of DVT, %
Anderson et al, ²⁶ 1999	1	447	100	27	N/A	Wells	Yes	13
Anderson et al, ²⁷ 2000	2	214	100	N/A	Moderate sensitivity	Wells	Yes	13
Miron et al, ²⁸ 2000	2	270	N/A	N/A	High sensitivity	Wells empirical estimate ^a	Yes	21
Kearon et al, ²⁹ 2001	1	445	60	N/A	Moderate sensitivity	Wells	Yes	14
Aguilar et al, ³⁰ 2002	2	134	100	0	High sensitivity	Wells	Not stated	N/A
Bucek et al, ³¹ 2002	2	99 Patients with low clinical probability	74	0	High sensitivity	Wells	No ^b	N/A
Kraaijenhagen et al, ¹⁵ 2002	1	1756	100	47	Moderate sensitivity	Wells	Yes	24
Shields et al, ³² 2002	2	102	100	0	Moderate sensitivity	Wells	Yes	17
Tick et al, ³³ 2002	1	811	100	10	Moderate sensitivity	Wells	Yes	42
Anderson et al, ³⁴ 2003	1	1075	71	19	Moderate sensitivity	Modified Wells ^c	No	18
Bates et al, ⁹ 2003	1	556	49	7	High sensitivity	Wells	Yes	10
Schutgens et al, ¹⁴ 2003	1	812	78	38	High sensitivity	Wells	Yes	39
Wells et al, ¹³ 2003	1	1082	62	18	Moderate sensitivity	Modified Wells ^c	No	16
Stevens et al, ³⁵ 2004	1	436	100	0	Not done	Wells	Yes	14

Abbreviations: DVT, deep vein thrombosis; N/A, overall prevalence not available.

^aDid not report D-dimer data; clinical prediction tool data from this prospective study was analyzed retrospectively.

^bOnly results for patients without previous DVT used in analysis (n = 87).

^cModified Wells score including 1 point for a history of DVT.

Table 18-9 Simplified Clinical Model for Assessment of Deep Vein Thrombosis^a

Clinical Variable	Score
Active cancer (treatment ongoing or within previous 6 mo or palliative)	1
Paralysis, paresis, or recent plaster immobilization of the lower extremities	1
Recently bedridden for 3 d or more, or major surgery within the previous 12 wk requiring general or regional anesthesia	1
Localized tenderness along the distribution of the deep venous system	1
Entire leg swelling	1
Calf swelling at least 3 cm larger than that on the asymptomatic leg (measured 10 cm below the tibial tuberosity) ^b	1
Pitting edema confined to the symptomatic leg	1
Collateral superficial veins (nonvaricose)	1
Previously documented DVT	1
Alternative diagnosis at least as likely as DVT	-2

Abbreviation: DVT, deep vein thrombosis.

^aScoring method indicates high probability if score is 3 or higher, moderate if score is 1 or 2, and low if score is 0 or lower.

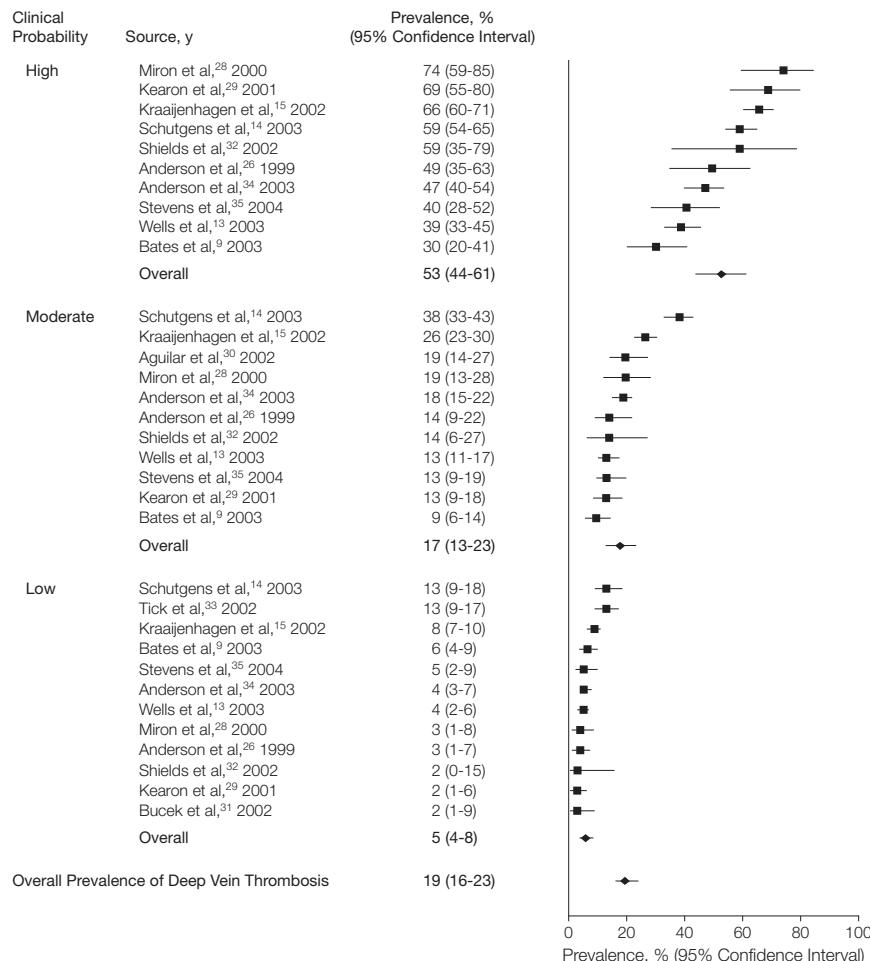
^bIn patients with symptoms in both legs, the more symptomatic leg was used.

tively (Figure 18-2). Interobserver reliability has not been widely evaluated, but the reported studies included many physicians with a wide range of clinical experience, including junior residents.

With the Wells et al¹³ criteria applied, the patient would have a score of 0, summed by pitting edema (1 point), bed rest (1 point), and an alternative diagnosis (cellulitis) at least as likely as DVT (-2 points). Using the clinical prediction rule, the clinician concludes that the patient has a low clinical probability of having an acute DVT. These data suggest that the clinician should be confident that the prevalence of DVT is approximately 5%. Would additional tests decrease the likelihood of DVT below 5%?

D-dimer Testing

D-dimer is a degradation product of a cross-linked fibrin blood clot. Levels of D-dimer are typically elevated in patients with acute venous thromboembolism. D-dimer levels may also be increased by a variety of nonthrombotic disorders, including recent major surgery, hemorrhage, trauma, pregnancy, cancer, or acute arterial thrombosis.³⁶ D-dimer assays are, in general, sensitive but nonspecific markers so that a positive D-dimer result is not useful to “rule in” the diagnosis of DVT. Instead, the value of the

**Figure 18-2** Prevalence of Deep Vein Thrombosis

D-dimer is with a negative test result that works to decrease the likelihood of the diagnosis.

The ability of a negative D-dimer result to “rule out” DVT depends on the type of assay. D-dimer assays are categorized as high sensitivity vs moderate sensitivity. The efficiency of a negative result to rule out DVT increases proportionately with the sensitivity of the assay, but it is inversely related to the prevalence of venous thromboembolism. On the other hand, the specificity of the particular D-dimer assay and the population under study affect its ability to rule out the diagnosis of DVT. For instance, use of a less specific assay or the testing of hospitalized patients who are currently ill limits its value because of the expected number of false-positive results.

The incorporation of D-dimer testing into diagnostic algorithms simplifies the management of a patient presenting with suspected DVT. Since the last review, numerous trials evaluated the accuracy of D-dimer and its incorporation into the diagnostic approach in patients with suspected DVT. Recent meta-analyses summarize the accuracy of various D-dimer assays compared with gold standard imaging tests for DVT.

Returning to the clinical scenario outlined earlier, a D-dimer test is performed. The hospital uses a moderately sensitive D-dimer assay. Does the type of D-dimer assay matter? Does the D-dimer result affect the already low probability of DVT?

How Will D-dimer Testing Simplify DVT Diagnosis?

Although a variety of quantitative and qualitative D-dimer assays are available and with all involving specific monoclonal antibodies, 2 methods have been extensively investigated: enzyme-linked immunosorbent assays (ELISA) and whole blood assays. There is wide variation in the sensitivity, normal reference ranges, and cutoff points among different assays. Current available assays can be divided into highly sensitive or moderately sensitive tests.¹⁸ A recent meta-analysis of different D-dimer assays shows that the ELISAs and certain immunoturbidimetric tests are highly sensitive ($\geq 95\%$) but less specific (approximately 40% at a cutoff value of 500 ng/mL) for excluding DVT.¹⁸ In general, other D-dimer methods such as whole blood and quantitative latex agglutination

assays are moderately sensitive ($\approx 85\%$) but more specific ($> 65\%$). Therefore, the probability after testing varies according to the D-dimer assay used. Before clinicians use a particular D-dimer assay to revise their clinical probability estimate, they should be aware of the differences and interpret the results accordingly. The use of D-dimer testing has improved the diagnostic process in suspected DVT, but the D-dimer result itself does not serve as the reference standard for the presence or absence of DVT.

The pooled sensitivity, specificity, and negative LRs of the D-dimer test in the low-clinical-probability group were 88% (95% CI, 81%-92%), 72% (95% CI, 65%-78%), and 0.18 (95% CI, 0.12-0.28), respectively. Among patients with moderate clinical probability estimate, the pooled values were 90% (95% CI, 80%-95%), 58% (95% CI, 49%-67%), and 0.19 (95% CI, 0.11-0.32), respectively; among patients with high clinical probability estimate, the results were 92% (95% CI, 85%-96%), 45% (95% CI, 37%-52%), and 0.16 (95% CI, 0.09-0.30), respectively. The specificity of D-dimer testing decreased as the clinical suspicion for DVT increased from low to moderate and from moderate to high ($P < .001$) with no change in the sensitivity ($P = .51$ and .28, respectively). The lower specificity of D-dimer testing among patients with a high clinical suspicion for DVT might be due to more comorbid conditions (eg, surgery or cancer) that can cause high D-dimer levels.³⁷ Among patients in this group, the number of false-positive D-dimer results can exceed the number of negative results, thereby limiting its use. The pooled estimates for diagnostic OR for D-dimer tests in the low-, moderate-, and high-clinical-probability groups were 17 (95% CI, 9.9-28), 14 (95% CI, 8.6-21), and 12 (95% CI, 5.7-25), respectively; that is, the diagnostic OR did not differ between clinical probability estimates despite a variation in sensitivity and specificity. These data are summarized in Table 18-10. Because the literature suggests that D-dimer assays can be broadly considered as high-sensitivity or moderate-sensitivity assays, we analyzed the eligible D-dimer studies in these categories.

Moderate-Sensitivity D-dimer Assays

The sensitivity, specificity, negative predictive values, LRs+, LRs-, and their respective 95% CIs for the studies that used

Table 18-10 Accuracy Measures for D-dimer Pooling of All Studies

Measures	Clinical Pretest Probability (95% CI)		
	Low	Moderate	High
Sensitivity, %	88 (81-92)	90 (80-95)	92 (85-96)
Specificity, %	72 (65-78)	58 (49-67)	45 (37-52)
Negative predictive value	99 (98-99)	96 (94-97)	84 (77-89)
Positive predictive value	17 (13-20)	32 (25-41)	66 (56-75)
Positive likelihood ratio	3.3 (2.6-4.1)	2.1 (1.8-2.5)	1.6 (1.5-1.8)
Negative likelihood ratio	0.18 (0.12-0.28)	0.19 (0.11-0.32)	0.16 (0.09-0.30)
Diagnostic OR	17 (9.9-28)	14 (8.6-21)	12 (5.7-25)

Abbreviations: CI, confidence interval; OR, odds ratio.

Table 18-11 Accuracy Measures in the Moderate-Sensitivity D-dimer Studies

Clinical Probability Before Testing	Study	Sensitivity, %	Specificity, %	NPV, %	LR+ (95% CI)	LR- (95% CI)
Low	Wells et al, ¹³ 2003	93	73	100	3.7 (2.9-4.6)	0.10 (0.01-1.4)
	Kraaijenhagen et al, ¹⁵ 2002	87	67	98	2.6 (2.3-3.1)	0.20 (0.11-0.36)
	Kearon et al, ²⁹ 2001	80	88	99	6.4 (3.6-11)	0.23 (0.04-1.3)
	Anderson et al, ²⁷ 2000	90	85	99	6.7 (4.3-10)	0.12 (0.01-1.7)
	Anderson et al, ³⁴ 2003	85	73	99	3.2 (2.5-4.1)	0.20 (0.07-0.58)
	Shields et al, ³² 2002	NE	80	98	5.0 (2.7-9.3)	0.32 (0.03-3.50)
	Weighted average (95% CI)	86 (79-92)	78 (71-83)	99 (98-99)	4.0 (3.0-5.4)	0.20 (0.12-0.31)
Moderate	Wells et al, ¹³ 2003	94	60	98	2.4 (2.0-2.8)	0.10 (0.03-0.38)
	Anderson et al, ³⁴ 2003	80	72	94	2.9 (2.4-3.6)	0.27 (0.17-0.43)
	Kraaijenhagen et al, ¹⁵ 2002	94	57	96	2.2 (2.0-2.5)	0.11 (0.05-0.21)
	Shields et al, ³² 2002	93	53	98	2.1 (1.5-3.0)	0.14 (0.01-2.0)
	Kearon et al, ²⁹ 2001	71	69	94	2.3 (1.6-3.2)	0.42 (0.23-0.80)
	Anderson et al, ²⁷ 2000	67	84	94	4.2 (2.0-9.0)	0.40 (0.16-1.0)
	Weighted average (95% CI)	85 (73-93)	66 (58-73)	95 (93-97)	2.4 (2.1-2.7)	0.23 (0.13-0.39)
High	Wells et al, ¹³ 2003	83	44	79	1.5 (1.2-1.9)	0.39 (0.20-0.77)
	Anderson et al, ³⁴ 2003	84	48	77	1.6 (1.3-2.0)	0.34 (0.20-0.56)
	Shields et al, ³² 2002	80	71	71	2.8 (0.8-9.4)	0.28 (0.07-1.1)
	Kraaijenhagen et al, ¹⁵ 2002	98	44	91	1.7 (1.5-2.1)	0.05 (0.02-0.14)
	Kearon et al, ²⁹ 2001	94	43	75	1.7 (1.0-2.6)	0.13 (0.03-0.59)
	Anderson et al, ²⁷ 2000	87	87	87	6.5 (1.8-24)	0.15 (0.04-0.57)
	Weighted average (95% CI)	90 (80-95)	49 (40-58)	81 (74-86)	1.7 (1.5-1.9)	0.20 (0.10-0.38)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio; NE, not estimable; NPV, negative predictive value.

moderate-sensitivity D-dimer assays are demonstrated in Table 18-11. Data are presented for each clinical probability estimate category. The LRs- are not sufficiently low to rule out DVT without ultrasonography among patients with moderate and high pretest probability estimates. Among these patients, the probability after testing for DVT is greater than 1% (see negative predictive values in Table 18-12). When combined with a negative D-dimer result, diagnostic imaging and anticoagulant therapy can be safely withheld for patients with a low clinical probability estimate because the LR- (0.20; 95% CI, 0.12-0.31) is such that the probability after testing for DVT is less than 1%.

High-Sensitivity D-dimer Assays

The sensitivity, specificity, negative predictive values, LRs+, LRs-, and their respective 95% CIs for the studies that used high-sensitivity D-dimer assays are demonstrated in Table 18-12. When combined with a negative D-dimer result, diagnostic imaging and anticoagulant therapy can be safely withheld in patients with a low (LR, 0.10; 95% CI, 0.03-0.37) or moderate clinical probability estimate (LR, 0.05; 95% CI, 0.01-0.21) because they create a probability estimate after testing for DVT of less than 1%. With a high clinical probability estimate, a normal D-dimer result does not have an LR low enough so that the probability of DVT becomes less than

1%. These results suggest pooling D-dimer data may not be appropriate. Table 18-13 demonstrates the probabilities after testing for the different clinical probability estimates according to the D-dimer results and includes an explanation of the application of Bayes theorem. Assessing the clinical effect of different sensitivity D-dimer assays on venous thromboembolic outcomes requires assumptions about the proportions of patients in each clinical probability category because they have not been compared head to head. This type of assessment is best performed by a formal decision analysis in which D-dimer assay accuracies and DVT prevalence are varied, and this is beyond the scope of this article. Comparative studies are required to provide more definitive conclusions.

Is Serial Ultrasonography Needed?

Should a negative D-dimer result after a normal ultrasonographic result suggest a need for serial ultrasonography?

Five studies reported sufficient data to enable the determination of the LR for a negative D-dimer result when the clinical probability estimate was moderate or high and the initial ultrasonographic result was normal (data not shown).^{9,13-15,34} Two studies used a high-sensitivity D-dimer.^{9,14} Because the probability of DVT after an initially negative ultrasonographic result is low, the LR for a negative D-dimer result ranges from 0.22 to 0.45 and results in a probability of DVT

of less than 1% after testing. Thus, regardless of the clinical probability estimate, a negative D-dimer result using a moderately sensitive D-dimer assay combined with a negative initial ultrasonographic result safely obviates the need for serial ultrasonography. However, caution must be used when performing D-dimer testing in patients with prolonged symptoms of suspected DVT or after a prolonged period of heparin therapy (>24 hours).³⁸

THE BOTTOM LINE

Outpatients presenting with suspected DVT should be initially assessed with a validated clinical prediction rule. The clinical prediction published by Wells et al¹³ has been assessed and validated in multiple clinical studies and can accurately categorize outpatients as having low, moderate, or high clinical probability. With this model, less than 5% of outpatients classified as low clinical probability have DVT. No other prediction tools met our eligibility criteria. A recent study suggests the prediction rule may not work in the primary care

setting, but limitations in the design of that study (in particular, failure to prospectively apply the rule as the diagnostic strategy) necessitate further research in primary care.⁴⁰ Validation studies of the model are required for hospitalized patients.

Incorporating D-dimer testing into a diagnostic algorithm further simplifies the management of a patient's case when he or she presents with suspected DVT. Once the clinical probability has been estimated, the D-dimer result can be combined to determine whether DVT can be safely ruled out without use of diagnostic imaging. Currently, the diagnosis of DVT can be ruled out without the need for ultrasonography by using a combination of low clinical probability estimate and a negative D-dimer result, and this strategy should apply to as many as 40% of patients referred with suspected DVT. Ultrasonography may provide information helpful to establish an alternative diagnosis, but ultrasonographic imaging for DVT is not required for every patient. Although the data are more limited, it seems likely that serial testing after an initially normal ultrasonographic result can be con-

Table 18-12 Accuracy Measures in the High-Sensitivity D-dimer Studies

Clinical Probability Before Testing	Study	Sensitivity, %	Specificity, %	NPV, %	LR+ (95% CI)	LR- (95% CI)
Low	Bates et al, ⁹ 2003	97	69	100	3.3 (2.7-3.9)	0.04 (0-0.65)
	Schutgens et al, ¹⁴ 2003	96	51	99	2.0 (1.7-2.4)	0.07 (0.01-0.51)
	Bucek et al, ³¹ 2002	83	53	99	2.1 (1.7-2.6)	0.32 (0.03-4.0)
	Weighted average (95% CI)	95 (82-99)	58 (45-71)	99 (97-100)	2.4 (1.7-3.3)	0.10 (0.03-0.37)
Moderate	Bates et al, ⁹ 2003	94	52	99	2.0 (1.6-2.4)	0.11 (0.02-0.76)
	Schutgens et al, ¹⁴ 2003	100	40	99	1.7 (1.5-1.9)	0.01 (0-0.16)
	Aguilar et al, ³⁰ 2002	98	32	99	1.5 (1.3-1.7)	0.06 (0-0.85)
	Weighted average (95% CI)	98 (91-100)	41 (31-52)	99 (96-100)	1.7 (1.5-1.9)	0.05 (0.01-0.21)
High	Bates et al, ⁹ 2003	98	40	98	1.7 (1.3-2.1)	0.06 (0-0.85)
	Schutgens et al, ¹⁴ 2003	98	34	90	1.5 (1.3-1.7)	0.07 (0.03-0.20)
	Weighted average (95% CI)	97 (94-99)	36 (29-43)	92 (81-97)	1.5 (1.4-1.7)	0.07 (0.03-0.18)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio; NPV, negative predictive value.

Table 18-13 Probabilities by Clinical Probability Estimate Combined With D-dimer Assays After Testing^a

	Clinical Probability Estimate ^b		
	Low	Moderate	High
Point estimate for DVT likelihood	5	17	53
Probability for positive D-dimer result after testing (high sensitivity)	11	25	63
Probability for negative D-dimer result after testing (high sensitivity)	0.5	1	8.6
Probability for positive D-dimer result after testing (moderate sensitivity)	17	34	67
Probability for negative D-dimer result after testing (moderate sensitivity)	0.9	4.4	19

Abbreviation: DVT, deep vein thrombosis.

^aProbability after testing from application of Bayes theorem.

^bPosttest odds = pretest odds \times likelihood ratio; pretest odds derived from pretest probability as follows: pretest odds = pretest probability/(1 - pretest probability). Similarly, posttest probability derived from posttest odds by posttest odds/(1 + posttest odds). For example, using a negative result with a high-sensitivity D-dimer if patient is low pretest probability, then pretest odds = 0.05/.95 = 0.052. Next, posttest odds = 0.052 \times 0.1 (from Table 18-12) = 0.0052. Convert to posttest probability by 0.0052/1.0052 = 0.052, or 5%.

fined to high-probability patients with positive D-dimer results. Patients with moderate probability and a negative high-sensitivity D-dimer result can have DVT ruled out.

Among patients with high clinical probability estimates, a normal D-dimer result does not have a sufficiently low LR. Therefore, all high-probability patients require diagnostic imaging to safely rule out DVT. Thus, D-dimer assays should not affect initial treatment for patients with a high probability of a DVT, because all of them require diagnostic imaging.

The specificity of D-dimer assays decreases as the clinical probability estimate increases, which leads to more false-positive test results, thereby limiting its utility. This emphasizes that D-dimer should not be used as a screening test, and indeed some advocate that D-dimer assays should not be used for patients at high risk for a false-positive result, ie, elderly patients, patients with cancer, and hospitalized patients.

CLINICAL SCENARIO—RESOLUTION

The clinician has already determined that the patient has a low pretest probability for DVT. The D-dimer result is now determined to be negative and therefore the probability of DVT after testing is sufficiently low (<1%) that the diagnosis can be safely ruled out. If the D-dimer result had been positive, the patient would require ultrasonographic imaging. In patients with low pretest probability, a normal ultrasonographic result reliably rules out clinically important DVT without the need for follow-up ultrasonography (probability after testing < 1%). If the ultrasonographic result is abnormal, it is usually considered predictive of DVT, although the probability after testing may be as low as 90%. Therefore, consideration should be given that it may be a false-positive result. Small, isolated, single-vein, nonocclusive ultrasonographic results have been reported to be falsely positive, mostly because they represent chronic DVT.³⁹

Author Affiliations at the Time of the Original Publication

Department of Medicine, University of Ottawa, and Division of Hematology, Ottawa Hospital, Ottawa, Ontario, Canada (Dr Wells); Department of Epidemiology and Community Medicine, University of Ottawa, Ottawa, Ontario, Canada (Drs Wells and Fergusson); Ottawa Health Research Institute, Ottawa, Ontario, Canada (Drs Wells and Fergusson and Mr Doucette); Department of Medicine, University of British Columbia, Vancouver, British Columbia, Canada (Dr Owen); and Department of Medicine, McMaster University, Hamilton, Ontario, Canada (Dr Tran).

Acknowledgments

We thank Richard White, MD, and Nathan Shapiro, MD, for reviewing an early draft of the manuscript. Dr Wells is supported by the Canada Research Chairs program. Dr Tran is supported by funding from Hamilton Health Sciences, Canada, and from the Haematology Societies of Australasia and New Zealand and Australia.

REFERENCES

- White RH. The epidemiology of venous thromboembolism. *Circulation*. 2003;107(23 suppl 1)I-4-I-8.
- Silverstein MD, Heit JA, Mohr DN, Petterson TM, O'Fallon WM, Melton LJ III. Trends in the incidence of deep vein thrombosis and pulmonary embolism: a 25-year population-based study. *Arch Intern Med*. 1998;158(6):585-593.
- Hirsh J, Bates SM. Prognosis in acute pulmonary embolism. *Lancet*. 1999;353(9162):1375-1376.
- Prandoni P, Lensing AWA, Prins M. Long-term outcomes after deep venous thrombosis of the lower extremities. *Vasc Med*. 1998;3(1):57-60.
- Fedullo PF, Auger WR, Channick RN, Kerr KM, Rubin LJ. Chronic thromboembolic pulmonary hypertension. *Clin Chest Med*. 2001;22(3):561-581.
- Wells PS, Anderson DR. Diagnosis of deep-vein thrombosis in the year 2000. *Curr Opin Pulm Med*. 2000;6(4):309-313.
- Cogo A, Lensing AWA, Koopman MMW, et al. Compression ultrasonography for diagnostic management of patients with clinically suspected deep vein thrombosis: prospective cohort study. *BMJ*. 1998;316(7124):17-20.
- Wells PS, Anderson DR, Bormanis J, et al. Value of assessment of pretest probability of deep-vein thrombosis in clinical management. *Lancet*. 1997;350(9094):1795-1798.
- Bates SM, Kearon C, Crowther MA, et al. A diagnostic strategy involving a quantitative latex D-dimer assay reliably excludes deep venous thrombosis. *Ann Intern Med*. 2003;138(10):787-794.
- Kearon C, Julian JA, Newman TE, Ginsberg JS. Noninvasive diagnosis of deep vein thrombosis. *Ann Intern Med*. 1998;128(8):663-677.
- Gottlieb RH, Voci SL, Syed L, et al. Randomized prospective study comparing routine versus selective use of sonography of the complete calf in patients with suspected deep venous thrombosis. *AJR Am J Roentgenol*. 2003;180(1):241-245.
- Anand SS, Wells PS, Hunt D, Brill-Edwards P, Cook D, Ginsberg JS. Does this patient have deep vein thrombosis? *JAMA*. 1998;279(14):1094-1099.
- Wells PS, Anderson DR, Rodger M, et al. Evaluation of D-dimer in the diagnosis of suspected deep-vein thrombosis. *N Engl J Med*. 2003;349(13):1227-1235.
- Schutgens RE, Ackerman P, Haas FJ, et al. Combination of a normal D-dimer concentration and a non-high pretest clinical probability score is a safe strategy to exclude deep venous thrombosis. *Circulation*. 2003;107(4):593-597.
- Kraaijenhagen RA, Piovella F, Bernardi E, et al. Simplification of the diagnostic management of suspected deep vein thrombosis. *Arch Intern Med*. 2002;162(8):907-911.
- Perrier A, Desmarais S, Miron MJ, et al. Non-invasive diagnosis of venous thromboembolism in outpatients. *Lancet*. 1999;353(9148):190-195.
- Heim SW, Schectman JM, Siadaty MS, Philbrick JT. D-dimer testing for deep venous thrombosis: a metaanalysis. *Clin Chem*. 2004;50(7):1136-1147.
- Stein PD, Hull RD, Patel KC, et al. D-dimer for the exclusion of acute venous thrombosis and pulmonary embolism: a systematic review. *Ann Intern Med*. 2004;140(8):589-602.
- Bayes T. An essay towards solving a problem in the doctrine of chances. *Philos Trans R Soc Lond B Biol Sci*. 1763;53:370-418.
- Tamariz LJ, Eng J, Segal JB, et al. Usefulness of clinical prediction rules for the diagnosis of venous thromboembolism: a systematic review. *Am J Med*. 2004;117(9):676-684.
- Fancher TL, White RH, Kravitz RL. Combined use of rapid D-dimer testing and estimation of clinical probability in the diagnosis of deep vein thrombosis: systematic review. *BMJ*. 2004;329(7470):821-824.
- Goodacre S, Sutton AJ, Sampson FC. Meta-analysis: the value of clinical assessment in the diagnosis of deep venous thrombosis. *Ann Intern Med*. 2005;143(2):129-139.
- Holleman DR Jr, Simel DL. Does the clinical examination predict airflow limitation? *JAMA*. 1995;273(4):313-319.
- Higgins JPT, Green S, eds. *Cochrane Handbook for Systematic Reviews of Interventions Version 5.0.0* [updated February 2008]. Cochrane Collaboration, 2008. <http://www.cochrane-handbook.org>. Accessed June 6, 2008. Page 16.9.2.
- Glas AS, Lijmer JG, Prins MH, Bonsel GJ, Bossuyt PMM. The diagnostic odds ratio: a single indicator of test performance. *J Clin Epidemiol*. 2003;56(11):1129-1135.
- Anderson DR, Wells PS, Stiell I, et al. Thrombosis in the emergency department: use of a clinical model to safely avoid the need for urgent radiological investigation. *Arch Intern Med*. 1999;159(5):477-482.

27. Anderson DR, Wells PS, Stiell I, et al. Management of patients with suspected deep vein thrombosis in the emergency department: combining use of a clinical diagnosis model with D-dimer testing. *J Emerg Med.* 2000;19(3):225-230.
28. Miron MJ, Perrier A, Bounameaux H. Clinical assessment of suspected deep vein thrombosis comparison between a score and empirical assessment. *J Intern Med.* 2000;247(2):249-254.
29. Kearon C, Ginsberg JS, Douketis JD, et al. Management of suspected deep venous thrombosis in outpatients by using clinical assessment and D-dimer testing. *Ann Intern Med.* 2001;135(2):108-111.
30. Aguilar C, Martinez A, Martinez A, Del Rio C, Vazquez M, Rodriguez FJ. Diagnostic value of D-dimer in patients with a moderate pretest probability of deep venous thrombosis. *Br J Haematol.* 2002;118(1):275-277.
31. Bucek RA, Koca N, Reiter M, Haumer M, Zontsich T, Minar E. Algorithms for the diagnosis of deep-vein thrombosis in patients with low clinical pretest probability. *Thromb Res.* 2002;105(1):43-47.
32. Shields GP, Turnipseed S, Panacek EA, Melnikoff N, Gosselin R, White RH. Validation of the Canadian clinical probability model for acute venous thrombosis. *Acad Emerg Med.* 2002;9(6):561-566.
33. Tick LW, Ton E, van Voorthuizen T, et al. Practical diagnostic management of patients with clinically suspected deep vein thrombosis by clinical probability test, compression ultrasonography, and D-dimer test. *Am J Med.* 2002;113(8):630-635.
34. Anderson DR, Kovacs MJ, Kovacs G, et al. Combined use of clinical assessment and D-dimer to improve the management of patients presenting to the emergency department with suspected deep-vein thrombosis (the EDITED Study). *J Thromb Haemost.* 2003;1(4):645-651.
35. Stevens SM, Elliot CG, Chan KJ, Egger MJ, Ahmed KM. Withholding anti-coagulation after a negative result on duplex ultrasonography for suspected symptomatic deep venous thrombosis. *Ann Intern Med.* 2004;140(12):985-991.
36. Kelly J, Rudd A, Lewis RR, Hunt BJ. Plasma D-dimers in the diagnosis of venous thromboembolism. *Arch Intern Med.* 2002;162(7):747-756.
37. Lee AYY, Julian JA, Levine MN, et al. Clinical utility of a rapid whole-blood D-dimer assay in patients with cancer who present with suspected acute deep venous thrombosis. *Ann Intern Med.* 1999;131(6):417-423.
38. Stricker H, Marchetti O, Haeberli A, Mombelli G. Hemostatic activation under anticoagulant treatment: a comparison of unfractionated heparin vs. nadroparin in the treatment of proximal deep vein thrombosis. *Thromb Haemost.* 1999;82(4):1227-1231.
39. Wells PS, Hirsh J, Anderson DR, et al. Comparison of the accuracy of impedance plethysmography and compression ultrasonography in outpatients with clinically suspected deep vein thrombosis—a two centre paired-design prospective trial. *Thromb Haemost.* 1995;74(6):1423-1427.
40. Oudega R, Hoes AW, Moons KGM. The Wells rule does not adequately rule out deep venous thrombosis in primary care patients. *Ann Intern Med.* 2005;143(2):100-107.

This page intentionally left blank

Prepared by David L. Simel, MD, MHS

Reviewed by Sheri Keitz, MD, PhD

CLINICAL SCENARIO

A 60-year-old man referred with suspected deep vein thrombosis (DVT) cut the plantar surface of his left foot on glass 10 days ago and has been resting in bed. He presents with left leg pain and mild calf swelling, redness, and heat. There is no history of a DVT or known family history of venous thromboembolism. Physical examination shows the patient is febrile and has pitting edema of the left calf. The calf erythema is hot, tender, and well demarcated. Enlarged left inguinal lymph nodes are present. He has longstanding diabetes mellitus, and the diagnoses that seem most likely are cellulitis and DVT. Can a clinical probability estimate of DVT reliably determine a pretest probability that can be used in decision making?

UPDATED SUMMARY ON DEEP VEIN THROMBOSIS

Original Review

Anand SS, Wells PS, Hunt D, et al. Does this patient have deep vein thrombosis? *JAMA*. 1998;279(14):1094-1099.

UPDATED REVIEW

Wells PS, Owen C, Doucette S, et al, eds. Does this patient have deep vein thrombosis? *JAMA*. 2006;295(2):199-207.

The Update was prepared within 12 months of publication of The Rational Clinical Examination article so the “Make the Diagnosis” section summarizes the findings published in the original review.

IMPROVEMENTS IN THE DATA PRESENTED IN THE ORIGINAL PUBLICATION

After the original publication, clinical prediction models were studied extensively and validated. The updated review provides evidence supporting the role of clinical prediction models for DVT.

CLINICAL SCENARIO—RESOLUTION

The clinician has already determined that the patient has a low pretest probability for DVT. The D-dimer result is now determined to be negative and therefore the probability of DVT after testing is sufficiently low (<1%) that the diagnosis can be safely ruled out. If the D-dimer result had been positive, the patient would require ultrasonographic imaging. In patients with low pretest probability, a normal ultrasonographic result reliably rules out clinically important DVT without the need for follow-up ultrasonography (probability after testing < 1%). If the ultrasonographic result is abnormal, it is usually considered predictive of DVT, although the probability after testing may be as low as 90%. Therefore, consideration should be given that it may be a false-positive result. Small, isolated, single-vein, nonocclusive ultrasonographic results have been reported to be falsely positive, mostly because they represent chronic DVT.¹

DEEP VEIN THROMBOSIS—MAKE THE DIAGNOSIS

PRIOR PROBABILITY

A validated clinical prediction rule, applied to the appropriate patient population, creates stratified probability estimates of DVT (see Table 18-14).

Table 18-14 Simplified Wells Prediction Rule²

Clinical Variable	Score
Active cancer (treatment ongoing or within previous 6 mo or palliative)	1
Paralysis, paresis, or recent plaster immobilization of the lower extremities	1
Recently bedridden for 3 d or more, or major surgery within the previous 12 wk requiring general or regional anesthesia	1
Localized tenderness along the distribution of the deep venous system	1
Entire leg swelling	1
Calf swelling at least 3 cm larger than the asymptomatic leg (measured 10 cm below the tibial tuberosity) ^a	1
Pitting edema confined to the symptomatic leg	1
Collateral superficial veins (nonvaricose)	1
Previously documented DVT	1
Alternative diagnosis at least as likely as DVT	-2
Simplified Score = Sum of Clinical Variables	
Probability of DVT, % (95% CI)	
Score ≥ 3, high probability	53 (44-61)
Score = 1 to 2, moderate probability	17 (13-23)
Score ≤ 0, low probability	5.0 (4-8)

Abbreviations: CI, confidence interval; DVT, deep vein thrombosis.

^aIn patients with symptoms in both legs, the more symptomatic leg was used.

POPULATION FOR WHOM DEEP VEIN THROMBOSIS SHOULD BE CONSIDERED

Deep vein thrombosis should be considered in patients with an acutely swollen leg that is causing discomfort, even though it can be bilateral and occur without prominent discomfort.

REFERENCE FOR THE UPDATE

- Wells PS, Hirsh J, Anderson DR, et al. Comparison of the accuracy of impedance plethysmography and compression ultrasonography in outpatients with clinically suspected deep vein thrombosis—a two centre paired-design prospective trial. *Thromb Haemost*. 1995;74(6):1423-1427.

DETECTING THE LIKELIHOOD OF DEEP VEIN THROMBOSIS

Because the prediction rule has been validated for the pretest probability and because the likelihood ratio (LR) varies according to the probability estimates and D-dimer assay, it is easier to display the posttest probability estimates without the LRs. Clinicians must know whether their laboratory uses the high-sensitivity D-dimer assay or the moderate-sensitivity assay. The clinical probability estimates *must* be determined before the D-dimer result is revealed to the clinician. Of all the findings, a negative high sensitivity D-dimer result has the biggest effect on the probability of disease and for many patients will provide evidence that obviates the need for imaging (see Table 18-15).

Table 18-15 Probability of Deep Vein Thrombosis After First Determining the Clinical Probability and Then Obtaining the D-dimer Result

Clinical Probability Estimates ^a	Probability of DVT After Applying D-dimer Result to the Clinical Probability Estimate, %		
	High Probability	Moderate Probability	Low Probability
High-sensitivity D-dimer	Positive	63	25
	Negative	8.6	1
Moderate-sensitivity D-dimer	Positive	67	34
	Negative	19	4.4

Abbreviation: DVT, deep vein thrombosis.

^aValues in the table use the exact summary pretest probability estimates, but a clinician might simplify by remembering that a high probability is about 50%; moderate probability, 20%; and low probability, 5%.

REFERENCE STANDARD TESTS

Imaging studies.

- Wells P, Anderson DR, Bormanis J, et al. Value of assessment of pretest probability of deep-vein thrombosis in clinical management. *Lancet*. 1997;350(9094):1795-1798.

CHAPTER 19

Is This Patient Clinically Depressed?

John W. Williams Jr, MD, MHS

Polly Hitchcock Noël, PhD

Jeffrey A. Cordes, MD

Gilbert Ramirez, DrPH

Michael Pignone, MD, MPH

CLINICAL SCENARIO

Mr P is a 52-year-old small-business owner with a 5-year history of controlled hypertension, for which he takes a thiazide diuretic. Otherwise, he is in good health. He presents for routine follow-up and notes a 1-month history of mild to moderate bitemporal headaches and feeling fatigued. The headaches occur about twice a week and are relieved by acetaminophen. He denies chest pain or dyspnea on exertion. He notes wryly that the “new economy” has left him feeling a bit “frazzled.”

You wonder whether the headache and fatigue are stress related, a somatic presentation of depression. What is the most effective and efficient method for diagnosing depression? How does one distinguish between somatic symptoms related to depression vs those related to coexisting physical illness?

WHY IS THIS AN IMPORTANT QUESTION TO ANSWER WITH A CLINICAL EXAMINATION?

Depressive disorders are prevalent, cause marked personal distress, and are associated with increased mortality. In primary care settings, the prevalence of major depression ranges from 4.8% to 8.6%, and dysthymia ranges from 2.1% to 3.7%.¹ The World Health Organization estimates that major depression alone is the fourth leading cause of disability worldwide.² Antidepressants and depression-specific psychological treatments are clearly effective for depression, improving both depressive symptoms and functional status.^{3,4} Many patients can be treated effectively in primary care settings. Quality improvement initiatives⁵ and disease management models⁶⁻¹⁰ are cost-effective compared with usual care and improve patient outcomes in primary care settings. Until effective prevention strategies are developed, high-quality depression care begins with recognition and accurate diagnosis. This evidence-based review will discuss case-finding and clinical interview strategies for depression diagnosis.

DEFINING CLINICAL DEPRESSION

Clinical depression is a syndromal diagnosis based on patient medical history and the exclusion of competing diagnoses. Depressive symptoms are evaluated along several continuums: intensity, duration, and influence on daily functioning. With these elements, symptoms can range from low mood lasting hours or a few days to major depression, characterized by multiple symptoms and substantial effect on daily functioning, according to criteria from the *Diagnostic and Statistical Manual of Mental Disorders*, (Fourth Edition) (*DSM-IV*) (Table 19-1).¹¹ A diagnostic nomenclature that helps guide treatment is “major depression,” “dysthymia,”

Table 19-1 Diagnostic Criteria and Questions to Assess Major Depression^a

Symptom	DSM-IV Diagnostic Criteria	Suggested Questions
Depressed mood	Depressed mood most of the day, nearly every day	How has your mood been lately? OR Do you ever feel down, depressed, or blue? How often does that happen? How long does it last?
Anhedonia	Markedly diminished interest or pleasure in almost all activities most of the day, nearly every day	Have you lost interest in your usual activities? Do you get less pleasure in things you used to enjoy?
Sleep disturbance	Insomnia or hypersomnia nearly every day	How have you been sleeping? How does that compare with your normal sleep?
Appetite or weight change	Substantial change in appetite nearly every day or unintentional weight loss or gain (eg, ≥5% of body weight in 1 mo)	Has there been any change in your appetite or weight?
Decreased energy	Fatigue or loss of energy nearly every day	Have you noticed a decrease in your energy level?
Increased or decreased psychomotor activity	Psychomotor agitation or retardation nearly every day	Have you been feeling fidgety or had problems sitting still? Have you felt slowed down, like you were moving in slow motion or stuck in mud?
Decreased concentration	Diminished ability to think or concentrate, or indecisiveness nearly every day	Have you been having trouble concentrating? Is it harder to make decisions than before?
Guilt or feelings of worthlessness	Feelings of worthlessness or excessive guilt nearly every day	Are you feeling guilty or blaming yourself for things? How would you describe yourself to someone who had never met you before?
Suicidal ideation	Recurrent thoughts of death or suicide	Have you felt that life is not worth living or that you'd be better off dead? Sometimes when a person feels down or depressed, they might think about dying. Have you been having any thoughts like that?

Abbreviation: *DSM-IV*, *Diagnostic and Statistical Manual of Mental Disorders*, (Fourth Edition).

^aThe diagnosis of major depression requires 5 or more symptoms, including depressed mood or anhedonia, which have been present during the same 2-week period and cause clinically significant distress or impairment in social, occupational, or other important areas of functioning.

Adapted from the *DSM-IV*.¹¹

Table 19-2 Diagnostic Categories for Depressive Disorders

Diagnostic Category	DSM-IV Criteria	Symptom Duration
Major depression	≥5 Depressive symptoms, including depressed mood or anhedonia, causing significant impairment in social, occupational, or other important areas of functioning	≥2 wk
Minor depression ^a	2-4 Depressive symptoms, including depressed mood or anhedonia, causing significant impairment in social, occupational, or other important areas of functioning	≥2 wk
Dysthymia	3 Or 4 dysthymic symptoms, ^b causing significant impairment in social, occupational, or other important areas of functioning	≥2 y

Abbreviation: *DSM-IV*, *Diagnostic and Statistical Manual of Mental Disorders*, (Fourth Edition).

^aMinor depression is included in *DSM-IV* as a research criteria diagnosis that requires further evaluation.

^bDysthymic symptoms are depressed mood, poor appetite or overeating, insomnia or hypersomnia, low energy, low self-esteem, poor concentration or indecisiveness, and hopelessness.

and “depression not otherwise specified.” Major depression is defined by depressed mood or loss of interest in nearly all activities for at least 2 weeks, accompanied by a minimum of 3 to 4 (for a total of 5) psychological (eg, decreased concentration) or somatic symptoms (eg, insomnia).¹¹ Dysthymia is characterized by fewer symptoms than major depression (<5) and a chronic course lasting at least 2 years (Table 19-2). Depression not otherwise specified includes syndromes without a sufficient number of symptoms (<5) or duration (<2 weeks) to meet major depression criteria. Within this category, minor depression, an unofficial diagnosis that has been nominated for further study, is an example with an insufficient number of symptoms.^{11,12}

HOW TO EVALUATE PATIENTS FOR CLINICAL DEPRESSION

There are 2 recommended approaches to recognizing and diagnosing depression. One approach, endorsed by the US Preventive Services Task Force, is to cue physicians to possible clinical depression by asking patients to complete a depression questionnaire during a routine appointment, an approach known as case-finding.¹³ Patients who score above a specified threshold are evaluated more carefully for depression. A second approach is to

evaluate patients for depression only when the clinical presentation triggers the suspicion of depression.¹ Chronic medical illness, chronic pain syndromes, recent life changes or stressors, fair or poor self-rated health, and unexplained physical symptoms are associated with depression.¹⁴ The likelihood of a depressive disorder increases by approximately 1.5 to 3.5 times if any of these factors is present.¹⁵

For either approach, a clinical interview is used to make a definitive diagnosis, in which the interviewer begins with open-ended questions and then proceeds as necessary to narrowly focused questions. In patients such as Mr P who present with somatic symptoms, a transition is recommended from inquiry about these symptoms to questions about emotional health. Many experts create useful transitions with questions such as, How are things at home? or, How are things at work? More narrowly focused questions should follow (Table 19-1), with priority given to questions about mood and anhedonia (a loss of interest or decreased pleasure in activities) because at least 1 of these 2 cardinal symptoms is required to diagnose clinically significant depression. Because successive generations use different synonyms for depressed mood, several alternatives should be offered in the question. For example, it may be helpful to ask, Have you been feeling sad, down, depressed, or blue? If answers to questions about mood and anhedonia are no, clinically significant depression is unlikely and alternative diagnoses should be considered more strongly.

Patients admitting to either depressed mood or anhedonia should be asked additional questions to determine whether there are sufficient symptoms to warrant a diagnosis of *clinical depression*. Assessing the effect of depressive symptoms on functioning and suicide risk are critical elements in the initial treatment decision. A helpful question to assess functioning is, Have these symptoms of [fill in patient's symptoms] affected your home or work life? Suicide assessment is more complex. Because patients rarely volunteer thoughts of suicide or their intentions to their physicians, it is important to ask directly. There is no evidence to suggest that asking about suicide precipitates suicidal thinking or acts.¹⁶ One useful screening question is, Have you been feeling that life is not worth living or that you would be better off dead?¹⁷ Another approach is to say, "Sometimes when a person feels down or depressed, they might think about dying. Have you been having any thoughts like that?" For patients with suicidal ideation, the next step is to ask, "Do you have a plan?" If a patient answers yes, inquire about the plan and determine whether he or she has assembled the materials required, has set a time, and whether there are any factors that may precipitate or keep the patient from carrying out the plan. Major risk factors for suicide include hopelessness, substance abuse, and previous suicide attempts. Patients at high risk of suicide should be referred for psychiatric evaluation; those at imminent risk should be evaluated immediately.¹⁸

Expert guidelines recommend a careful review of systems to detect general medical conditions that may masquerade as depression or complicate its treatment.¹ Physical conditions, such as hypothyroidism or Cushing disease, may cause depression, and some experts recommend a thyrotropin measurement in women older than 50 years because of the increased prevalence of hypothyroidism.^{1,19} Because these physical conditions are

etiologic, treatment is directed at the underlying condition rather than the depressive symptoms. Similarly, medication such as glucocorticoids, anabolic steroids, and high-dose reserpine or withdrawl from cocaine or amphetamine can cause depression.²⁰⁻²² Other medical conditions such as malignancies, diabetes mellitus, autoimmune disorders, and coronary heart disease are highly associated but not causative for depression, and treatment is directed simultaneously at the clinical depression and the associated physical illness.^{1,4,23-26} Diagnostic testing for these disorders is indicated only when clinical symptoms suggest the condition. For example, patients with weight loss out of proportion to the depression should be evaluated for malignancy or other systemic disorders associated with weight loss. Psychiatric illnesses such as alcohol abuse are common in primary care settings and often co-occur with depression.²⁷ The combination is difficult to treat, often requiring mental health specialty care. The CAGE questions (Have you ever felt the need to cut down on your drinking? Have you ever felt annoyed by criticism of your drinking? Have you ever felt guilty about your drinking? Have you ever taken a drink [eye opener] first thing in the morning?) are a pragmatic and effective screen for alcohol abuse.²⁸

Once depression is diagnosed, additional history should be elicited about factors that may affect treatment. First, explore the patient's understanding and acceptance of the diagnosis. Stigmatizing beliefs about depression or outright rejection of the diagnosis may interfere with treatment adherence. Second, elicit the patient's treatment preferences and information on response to therapy for previous episodes of depression. This is particularly important for pharmacotherapy because antidepressant agents that have been used successfully for past depressive episodes are likely to be effective and well tolerated for the current episode.⁴ Finally, assess the number of previous episodes. The risk of relapse, and hence the need for longer-term treatment, increases with the number of previous episodes.^{1,4}

CRITERION STANDARD DIAGNOSIS

Clinical depression is a syndromal diagnosis. There is no physiologic or laboratory test, radiologic examination, or tissue diagnosis to definitively establish the diagnosis. Instead, a trained interviewer conducts a clinical interview to determine whether the patient meets established criteria. The most commonly used criteria, which are updated periodically, are the *DSM-IV* or the *International Classification of Diseases, Tenth Revision, Clinical Modification*.^{11,29}

METHODS

Search Strategy and Inclusion/Exclusion Criteria

We conducted separate searches of MEDLINE and a specialized registry of depression trials³⁰ for English-language medical literature published from 1970 through July 2000 for studies evaluating the performance of case-finding instruments in primary care settings and the reliability of the clinical interview. All searches included the terms "depressive disorder" or "depression" and additional terms as appropriate for the specific search. Unpublished data were not sought.

For case-finding, we modified inclusion criteria used in our previous literature synthesis³¹ to select instruments that are most readily used in clinical situations. Studies were included if they were conducted in a primary care setting, administered a case-finding instrument, and used a standard interview such as the Structured Clinical Interview for *Diagnostic and Statistical Manual of Mental Disorders, Third Edition Revised* (*SCID*, *DSM-III-R*)³² to make a criterion-based diagnosis (eg, *DSM-III*, *DSM-III-R*, *DSM-IV*) of depression. Furthermore, we specified that the case-finding instrument have easy to average literacy requirements,³³ be scored without a calculator, have a depression-specific component, and be evaluated in at least 1 study with at least 100 subjects. Of 1766 articles identified by the search strategy, 379 potentially eligible studies were reviewed. Twenty-eight studies, involving 11 case-finding instruments, met all inclusion criteria.³⁴⁻⁶¹

For reliability studies, we required criterion-based diagnoses made by 2 or more clinicians who interviewed the same patient or reviewed an audiotape or videotape interview. Clinicians evaluated patients with known or suspected psychiatric illness who were recruited from inpatient or outpatient settings in both mental health and general medical settings. Studies using non-clinician interviewers were excluded. Among studies using semi-structured interviews, we only included those using the *SCID*, a commonly used research instrument for diagnosing psychiatric illness. The search yielded 6103 potentially eligible articles, of which 14 met all inclusion/exclusion criteria.⁶²⁻⁷⁵

Data Abstraction and Statistical Methods

Two independent reviewers abstracted articles. For case-finding studies, quality assessment addressed sample size

greater than 100, whether patients were selected consecutively or randomly, whether the criterion standard was administered and interpreted independently of and blind to the results of the case-finding instrument, and whether the proportion of persons receiving the criterion standard assessment was less than or more than 50% of those approached for criterion standard assessment. For reliability studies, quality assessments addressed whether key patient characteristics were described (eg, depression severity), whether the interviewers collected clinical history independently, and whether diagnoses were made blinded to other clinicians' evaluations.

Established cut points for case-finding instruments were used except for short versions of original instruments that had proportionally lower thresholds^{35,43,46,60} and one study that used a higher threshold than originally recommended.⁴⁴ Two-by-two tables were used to categorize the number of persons whose results were positive and negative and who did and did not meet criterion standard diagnosis for major depression. When it was appropriate, we adjusted for verification bias.⁷⁶ Of 11 authors contacted for additional information, 10 responded with the needed data. The average positive likelihood ratio (*LR+*) and negative likelihood ratio (*LR-*), weighted by study precision and corrected for 2-stage assessment techniques when indicated, were computed for each case-finding instrument.⁷⁷⁻⁷⁹ A scattergram plotting true-positive against false-positive rates was constructed to visually evaluate variability among studies. To provide a visual reference for the consistency of study results, we modeled a summary receiver operating characteristic curve based on the logit transformation of the true-positive and false-positive rates. The effectiveness score was used to evaluate overall performance and study heterogeneity.⁸⁰ Studies of reliability were not combined quantitatively because of marked heterogeneity in study design.

Box 19-1 Web Sites for Case-Finding Instruments^a

Beck Depression Inventory (BDI):

http://en.wikipedia.org/wiki/Beck_Depression_Inventory

Center for Epidemiologic Studies Depression (CES-D):

<http://www.chcr.brown.edu/pcoc/cesdscale.pdf>

Duke Anxiety-Depression Scale (DADS):

<http://healthmeasures.mc.duke.edu/>

Geriatric Depression Scale (GDS)—Long or Short Versions:

<http://www.stanford.edu/~yesavage/GDS.html>

Primary Care Evaluation of Mental Disorders (PRIME-MD)⁴⁸:

<http://jama.ama-assn.org/cgi/content/full/282/18/1737>

PRIME-MD Patient Health Questionnaire (PHQ):

<http://www.phqscreeners.com/>

The Zung Self-Rating Depression Scale (SDS):

<http://healthnet.umassmed.edu/mhealth/ZungSelfRatedDepressionScale.pdf>

^aAll Web sites accessed May 28, 2008.

RESULTS

Accuracy of Case-Finding Questionnaires for Depression

Eleven questionnaires, ranging from 1 to 30 items, met all inclusion criteria (Table 19-3). Six are depression-specific (Beck Depression Inventory [BDI], Center for Epidemiologic Studies Depression Screen [CES-D], Depression Scale [DEPS], Geriatric Depression Scale [GDS], Zung Self-Rating Depression Scale [SDS], and Single Question [SQ]), 1 addresses depression and anxiety (Duke Anxiety and Depression Scale [DADS]), and 4 are multicomponent (Hopkins Symptom Check List [HSCL], Primary Care Evaluation of Mental Disorders [PRIME-MD], PRIME-MD Patient Health Questionnaire [PHQ], and Symptom Driven Diagnostic System for Primary Care [SDDS-PC]). All of the questionnaires can be self-administered in less than 5 minutes, all include specific questions aimed at assessing depressed mood, and, except for the SQ instrument, all assess anhedonia. Resources to obtain the full instruments are listed in Box 19-1.

Table 19-3 Characteristics of Depression Case-Finding Instruments Validated in Primary Care Settings

Instrument	No. of Items ^a	Scope	Response Format	Period of Questions	Score Range	Usual Cut Point ^b	Literacy Level ^c	Administration Time, min	Monitor Severity or Response
BDI	21, 13, 7	Depression-specific (multiple versions)	4 Statements of symptom severity per item	Today	0-63	10-19 Mild, 20-29 moderate, ≥30 severe	Easy	2-5	Yes
CES-D	20, 10	Depression-specific (2 versions)	4 Frequency ratings: "less than 1 d" to "most or all (5-7 d)"	Past week	0-60	≥16	Easy	2-5	Yes
DEPS	10	Depression-specific	4 Frequency ratings: "not at all" to "extremely"	Last month	0-30	≥9	Average	≤2	Unknown
DADS	7	For anxiety and depression	3 Frequency ratings: "yes, somewhat, no" for 3 items; "none, some, a lot" for 4 items	Past week	0-100	>30	Average	≤2	Unknown
GDS	30, 15	Depression-specific (2 versions)	Yes or no	Past week	0-30	≥11	Easy	2-5	Yes
HSCL	25, 13	Multiple versions and multiple components with depression category	4 Frequency ratings: "not at all" to "much more than usual"	Past week	25-100	≥43	Average	2-5	Yes
PRIME-MD	2	Multiple components with depression	Yes or no	Past month	0-2	≥1	Average	<1	No
PRIME-MD (PHQ)	9	Multiple components with depression	4 Frequency ratings: "not at all" to "nearly every day"	Past 2 wk	0-9 For diagnosis; 0-27 for response	For diagnosis: 5 symptoms. For severity: 0-4 none; 5-9 mild; 10-14 moderate; 15-19 major; 20-27 severe	Average	<2	Yes
SDDS-PC	5	Multiple components with depression	Yes or no	Past month	0-5	≥2	Easy	<2	Unknown
SDS	20	Depression-specific	4 Frequency ratings: "little of the time" to "most of the time"	Recently	25-110 (Sum/80 × 100)	50-59 Mild, 60-69 moderate, ≥70 severe	Easy	2-5	Yes
SQ	1	Depression-specific	Yes or no	Past year	0-1	1	Easy	<1	No

Abbreviations: BDI, Beck Depression Inventory; CES-D, Center for Epidemiologic Studies Depression Screen; DADS, Duke Anxiety and Depression Scale; DEPS, Depression Scale; GDS, Geriatric Depression Scale; HSCL, Hopkins Symptom Check List; PRIME-MD, Primary Care Evaluation of Mental Disorders; PRIME-MD (PHQ), PRIME-MD Patient Health Questionnaire; SDDS-PC, Symptom-Driven Diagnostic System for Primary Care; SDS, Zung Self-Rating Depression Scale; SQ, single question.

^aNumbers refer to different versions of the same instrument and are listed from most to least number of items. Item numbers for the DADS, PRIME-MD, PRIME-MD (PHQ), and SDDS-PC refer to depression questions only; item numbers for the HSCL refer to depression plus anxiety questions.

^bCut point is given for the instrument version with the highest number of items.

^cEasy indicates third- to fifth-grade reading level; average, sixth- to ninth-grade reading level.³³

The BDI, the CES-D, and the SDS were developed specifically to identify depression. They include similar numbers of questions and use response formats that rely either on ranking symptom severity or on classifying frequency of symptoms. These 3 instruments are among the most thoroughly evaluated in primary care and can be used to rate the severity of depression and monitor response to therapy. Shortened versions of

the BDI and the CES-D have been tested recently in primary care.^{35,43} The GDS exists in both 30- and 15-item versions and uses a yes-or-no response format that simplifies telephone administration. It has been evaluated only in populations aged 60 years and older. DADS, DEPS, and SQ (Have you felt depressed or sad much of the time in the past year?) are newer, brief instruments that have been evaluated in single studies.

The PRIME-MD and SDDS-PC instruments are multidimensional questionnaires. Each has screening questions arranged in several categories (eg, depression, anxiety, alcohol abuse) that are used to trigger more extensive diagnostic interviewing sections for specific *Diagnostic and Statistical Manual (DSM)* diagnoses. Recently, the PHQ, a completely self-administered version of the PRIME-MD, has been evaluated. It scores each *DSM-IV* depression symptom as present or absent to diagnose depres-

sion, and can also be scored continuously to monitor treatment response. The HSCL screens for general psychiatric illness and has a specific category for depression.

These instruments, encompassing 37 evaluations in 28 published studies,³⁴⁻⁶¹ involved 25550 screened patients, of whom 9218 were administered an acceptable criterion standard for diagnosing depression (Table 19-4). Nine of 28 studies had potential major selection biases because more

Table 19-4 Case-Finding Instrument Performance in Primary Care Settings

Source, y	Instrument	Quality Score ^a	Sample Size ^a	Population ^b	Positive Likelihood Ratio (95% CI) ^c	Negative Likelihood Ratio (95% CI) ^d
Major Depressive Disorder						
Whooley et al, ³⁴ 1997	BDI	1	536	Veterans Affairs	2.5	0.18
Steer et al, ³⁵ 1999	BDI, 7 items	3	120	Academic	97	0.03
Perez-Stable et al, ³⁶ 1990	BDI	4	265	Mixed	1.5	0.25
Zich et al, ³⁷ 1990	BDI	4	31	Mixed	3.4	0.17
Summary BDI					4.2 (1.2-14)	0.17 (0.1-0.3)
Kirmayer et al, ³⁸ 1993	CES-D		685	Academic	3.0	0.28
Whooley et al, ³⁴ 1997	CES-D	1	536	Veterans Affairs	3.0	0.10
Williams et al, ³⁹ 1999	CES-D	1	296	Mixed	3.2	0.13
Fechner-Bates et al, ⁴⁰ 1994	CES-D	2	425	Community	2.7	0.29
Hendrie et al, ⁴¹ 1995	CES-D	2	125	Academic (age ≥ 60 y)	2.9	0.60
Hough et al, ⁴² 1983	CES-D	2	525	Health maintenance organization	3.9	0.23
Irwin et al, ⁴³ 1999	CES-D, 10 items	3	68	Academic (age ≥ 60 y)	11	0.14
Lyness et al, ⁴⁴ 1997	CES-D	4	130	Community (age ≥ 60 y)	12	0.15
Perez-Stable et al, ³⁶ 1990	CES-D	4	214	Mixed	1.9	0.40
Zich et al, ³⁷ 1990	CES-D	4	34	Mixed	1.8	0.31
Summary CES-D					3.3 (2.5-4.4)	0.24 (0.2-0.3)
Neal and Baldwin, ⁴⁵ 1994	GDS	2	45	Academic (age > 65 y)	4.0	0.25
D'Ath et al, ⁴⁶ 1994	GDS, 15 items	4	120	Community (age ≥ 65 y)	3.3	0.12
Summary GDS					3.3 (2.4-4.7)	0.16 (0.1-0.3)
Schmitz et al, ⁴⁷ 1999	HSCL	1	421	Community	2.0	0.37
Hough et al, ⁴² 1983	HSCL	2	525	Health maintenance organization	5.4	0.17
Summary HSCL					3.2 (1.7-6.2)	0.24 (0.1-0.5)
Spitzer et al, ⁴⁸ 1999	PHQ	1	585	Mixed	12 (8.4-18)	0.28 (0.2-0.5)
Spitzer et al, ⁴⁹ 1994	PRIME-MD	1	431	Mixed	3.4	0.19
Whooley et al, ³⁴ 1997	PRIME-MD	1	536	Veterans Affairs	2.2	0.07
Summary PHQ or PRIME-MD					2.7 (2.0-3.7)	0.14 (0.1-0.3)
Leon et al, ⁵⁰ 1996	SDDS-PC	1	501	Community	5.4	0.34
Whooley et al, ³⁴ 1997	SDDS-PC	1	536	Veterans Affairs	2.0	0.08
Broadhead et al, ⁵¹ 1995 ⁱ	SDDS-PC	3	388	Community	4.0	0.12
Broadhead et al, ⁵¹ 1995	SDDS-PC	3	257	Mixed	3.9	0.40
Summary SDDS-PC					3.5 (2.4-5.1)	0.22 (0.1-0.4)
Spitzer et al, ⁴⁹ 1994	SDS	1	337	Mixed	3.3	0.19
Okimoto et al, ⁵² 1982	SDS	3	55	Veterans Affairs (age ≥ 60 y)	2.2	0.05
Magruder-Habib et al, ⁵³ 1990	SDS	4	206	Veterans Affairs	15	0.27
Raft et al, ⁵⁴ 1977	SDS	4	69	Academic	1.0	0.97
Summary SDS					3.3 (1.3-8.1)	0.35 (0.2-0.8)
Williams et al, ³⁹ 1999	SQ	1	291	Mixed	2.3 (1.8-2.9)	0.16 (0.0-0.6)
Median performance for all instruments					3.3	0.19

(continued)

Table 19-4 Case-Finding Instrument Performance in Primary Care Settings (Continued)

Source, y	Instrument	Quality Score ^a	Sample Size ^a	Population ^b	Positive Likelihood Ratio (95% CI) ^c	Negative Likelihood Ratio (95% CI) ^d
Major Depressive Disorder or Dysthymia						
Klinkman et al, ⁵⁵ 1997	CES-D	2	425	Mixed	2.9	0.27
Schulberg et al, ⁵⁶ 1985	CES-D	4	294	Community	5.2	0.26
Salokangas et al, ⁵⁷ 1995	DEPS	2	436	Community	4.9	0.31
Parkerson and Broadhead, ⁵⁸ 1997	DADS	3	481	Academic	2.3	0.28
van Marwijk et al, ⁵⁹ 1995	GDS, 30 items	1	586	Community (age \geq 65 y)	3.9	0.53
Arthur et al, ⁶⁰ 1999	GDS, 15 items	3	201	Community (age \geq 75 y)	3.4	0.05
Nettelbladt et al, ⁶¹ 1993	HSCL	2	186	Community	2.8	0.36
Median performance for all instruments					3.9	0.30

Abbreviations: BDI, Beck Depression Inventory; CES-D, Center for Epidemiologic Studies Depression Screen; CI, confidence interval; DEPS, Depression Scale; DADS, Duke Anxiety and Depression Scale; GDS, Geriatric Depression Scale; HSCL, Hopkins Symptom Check List; PHQ, PRIME-MD Patient Health Questionnaire; PRIME-MD, Primary Care Evaluation of Mental Disorders; SDDS-PC, Symptom Driven Diagnostic System for Primary Care; SDS, Zung Self-Rating Depression Scale; SQ, Single Question.

^aThe sample size refers to the actual number who received the criterion standard.

^bMixed-community and university-affiliated clinics, academic university-affiliated clinics, and community-private practice clinics.

^cThe positive likelihood ratio describes how much more likely it is that a positive depression screening result would be observed in an individual with depression than in someone without depression. It is calculated as sensitivity/(1 – specificity). Summary is a weighted average across all studies.

^dThe negative likelihood ratio describes how much more likely it is that a negative depression screening result would be observed in an individual with depression than in someone without depression. It is calculated as (1 – sensitivity)/(specificity). Summary is a weighted average across all studies.

^eLower scores indicate higher quality.

^fThe study by Broadhead et al⁵¹ is listed twice for the same instrument because it included both an initial test set of patients and a validation set of patients.

than 50% of persons selected did not receive a criterion standard assessment, either because they refused the assessment or failed to keep an appointment.^{36,37,41,44,51-53,56,58} Considering independent and blind administration of the criterion standard, major selection bias, and sample size, 15 (54%) of the 28 studies were of reasonably high quality for diagnostic test evaluations.

Figure 19-1 plots the study results for true-positive and false-positive rates for case-finding instruments used to detect major depression. Standard cut points were used for these calculations (Table 19-3) except for one study using higher than recommended thresholds for the CES-D.⁴⁴ The cut point for mild depression was used for the 2 scales with 3 listed cut points (BDI and SDS); the study by Raft et al⁵⁴ only had information corresponding to moderate depression for the SDS. Two studies were identified as outliers.^{36,54} The study by Raft et al⁵⁴ used the higher cut point for the SDS scale and had an unusually low sensitivity (31%; 95% confidence interval [CI], 16%-51%). The study by Perez-Stable et al³⁶ had an unusually low specificity of 40% for the BDI. They studied patients with high levels of medical comorbidity and high ethnic minority representation, factors that may have decreased specificity.

The median LR+ for all studies was 3.3 (range, 2.3-12), meaning that a positive depression screening result is 3.3 times more likely to be observed in someone with depression than in someone without the illness. The median LR- was 0.19 (range, 0.14-0.35), meaning that a negative depression screening result was about 0.2 times as likely to be observed in someone with depression than in someone without the illness. Performance did not differ significantly between instruments. With the

effectiveness score as a measure of overall performance, there was statistically significant heterogeneity for the BDI ($P < .01$), CES-D ($P < .04$), HSCL ($P = .04$), and SDS ($P < .01$), indicating that the instruments performed variably across the individual studies. The variability may be due to differences in the patient populations or study design. Finally, a subset of studies reported instrument performance for major depression and separately for the combined category of major depression or dysthymia.⁵⁵⁻⁶¹ Performance characteristics for detecting this combined category were not statistically significantly different from those for detecting major depression alone.

Given the similar performance, case-finding instruments should be selected according to brevity, response format (particularly if telephone administration is planned), the desire to screen for other psychiatric illnesses, and the need to monitor response. The PHQ best meets these criteria with only 9 items for depression, modules for other psychiatric illness, and a simple response format that is sensitive to change. For clinicians who wish to screen only for depression, the SQ is an attractive alternative that could be asked during preventive medicine evaluations or in response to triggers that increase the likelihood of depression. Positive responses would need to be explored by a more careful clinical interview. In a clinic with an 8% prevalence of major depression or dysthymia, a clinician treating 100 patients per week can expect that 30 will have a positive screening result for depression, of whom 7 would meet criteria for clinical depression after a more careful clinical interview. Among the 70 patients who have a negative screening result for depression, 1 would have clinical depression. If case-finding were used only in selected high-risk patients (eg,

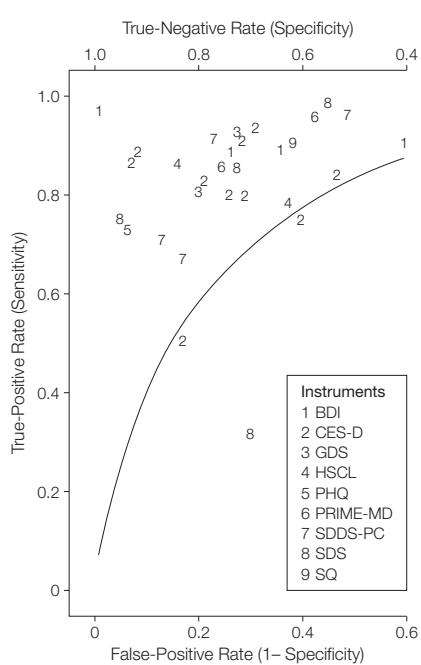


Figure 19-1 True-Positive Rate Against False-Positive Rate for Case-Finding Instruments to Detect Major Depression

The numbered points represent individual studies. The curve represents the summary receiver-operating curve through all data points. Standard cut points (see Table 19-3) were used for these calculations except for one study⁴⁴ that used higher than recommended thresholds for CED-S. Abbreviations: BDI, Beck Depression Inventory; CES-D, Center for Epidemiologic Studies Depression Screen; GDS, Geriatric Depression Scale; HSCL, Hopkins Symptom Check List; PHQ, PRIME-MD Patient Health Questionnaire; PRIME-MD, Primary Care Evaluation of Mental Disorders; SDDS-PC, Symptom Driven Diagnostic System for Primary Care; SDS, Zung Self-Rating Depression Scale; SQ, Single Question.

those with chronic pain), a positive screening result would more likely be a true positive, but more patients with clinical depression would be missed.

Accuracy and Reliability of the Clinical Interview for Depression

Because the criterion standard diagnosis is based on a clinical interview, there is no simple method for establishing its accuracy. However, we identified relevant studies comparing the diagnostic agreement between 2 mental health professionals, between primary care physicians and a mental health professional, and the effects of coexisting medical illness on reliability.

We identified 7 studies using the SCID, which evaluated interrater reliability for major depression (Table 19-5).⁶²⁻⁶⁸ The SCID is a widely used research instrument that uses a semistructured interview to elicit symptoms that are applied to the current *DSM* criteria to establish a diagnosis.³² It is designed in part to decrease variability related to the range of symptoms explored and the manner in which a clinical interviewer presents questions. Study design varied considerably, ranging from multiple clinicians viewing a videotaped interview to paired interviewers conducting sequential interviews. Examiners' training and experience ranged from psychology trainees to practicing psychia-

trists with a special interest in mood disorders. All were conducted in mental health specialty settings. Diagnoses were made blind to the other raters' diagnosis in 6 studies, patient medical histories were elicited independently in only 1 study, and no study described depression severity. Despite the variability in study design and examiner training, interrater agreement corrected for chance was substantial to almost perfect ($\kappa = 0.64$ -0.93). These studies show that major depression can be diagnosed reliably by a mental health professional who uses a careful, semistructured interview.

Studies that use nonstandardized interviews to make *DSM*-based diagnoses may better simulate clinical practice. Seven studies, involving psychiatry trainees to practicing psychiatrists, evaluated interrater agreement with this approach.⁶⁹⁻⁷⁵ Diagnoses were based typically on paired interviewers conducting joint or sequential interviews; one study used a videotaped interview.⁷⁰ Diagnoses were made blind to the other raters' diagnosis in 5 studies, patient medical histories were elicited independently for most patients in 3 studies, and no study described depression severity. Interrater agreement corrected for chance was moderate to substantial ($\kappa = 0.55$ -0.74). Compared with semistructured interviews, agreement was somewhat lower for nonstandardized interviews. However, chance-corrected agreement remained good compared with many other clinical findings.⁸¹⁻⁸³ Less is known about the reliability of depression diagnoses made by primary care physicians. We were able to identify only 1 study that compared a primary care clinician's diagnoses based on *DSM* criteria to that of a mental health professional using the same criteria. Spitzer et al⁴⁹ compared primary care clinicians' diagnoses using a semistructured instrument to mental health professionals' diagnoses with an SCID-based *DSM* measure of depression. This study found good agreement (simple agreement, 88%; $\kappa = 0.71$). It is unknown how well primary care physicians using a nonstandardized interview would agree with diagnoses made by mental health professionals.

These studies have a number of design limitations. The severity of major depression and spectrum of competing medical and psychological illnesses that may make diagnosis more difficult were not typically described. In addition, studies using joint interviews and videotape review may overestimate interrater reliability because both interviewers hear identical information. Two of the studies compared diagnoses made by emergency department psychiatrists to those made by the patient's inpatient treating physician and were thus not blinded evaluations, again potentially biasing these studies toward higher agreement.^{72,75} Finally, only 1 study reported 95% CIs for the estimate of interrater agreement.⁶⁸

Effect of Physical Illness on Diagnosis

Because the psychological and physical symptoms of depression may overlap with other physical illness, diagnosing depression in patients with severe or multiple chronic medical illnesses can be especially challenging.⁸⁴ If symptoms caused by the physical illness (eg, fatigue related to congestive heart failure) are attributed to depression, then patients may receive unnecessary treatment.

Table 19-5 Interrater Reliability for Depressive Disorder With Semistructured and Nonstructured Interviews

Source, y	Examiners (No.)	No. of Patients Evaluated	No. of Patients With MDD Diagnosis	Setting	Design	Independent Assessment ^a	Blinding ^b	Simple Agreement (κ)
Semistructured Interview								
Fuhrer et al, ⁶² 1986	Psychiatrists (136)	11	2	Inpatient	Videotape review	No	Yes	NA (0.78)
Riskind et al, ⁶³ 1987	Psychologists (16)	75	25	Outpatient	Videotape review	No	Yes	82% (0.72)
Stukenberg et al, ⁶⁴ 1990	Psychology trainees (4)	75	NA	Community	NS	NS	NS	NA (0.92)
Skre et al, ⁶⁵ 1991	Psychiatrist (1) Psychologists (4)	54	25	Mixed	Live vs audiotape review	No	Yes	NA (0.93)
Williams et al, ⁶⁶ 1992	Psychiatrists (14) Psychologists (6) Master's degree (4)	390	121	Mixed	Live, sequential interview	Yes	Yes	NA (0.64)
Segal et al, ⁶⁷ 1994	Psychology trainees (NS)	33	15	Outpatient	Live vs audiotape or videotape review	No	Yes	85% (0.70)
Keller et al, ⁶⁸ 1995	Master's degree (NS)	80	68	Mixed	Live vs videotape review	No	Yes	NA (0.72)
Nonstandardized Interview								
Spitzer et al, ⁶⁹ 1979	Mental health clinicians (274)	281	83	Mixed	Joint or sequential interview	Mixed	Yes	NA (0.70)
Webb et al, ⁷⁰ 1981	Mental health clinicians (78)	1	1	NA	Videotape review	No	Yes	83% (NA)
Hyler et al, ⁷¹ 1982	Psychiatrists (31) Psychologists (3) Social workers (7)	46	14 ^c	Mixed	Joint or sequential interview	Mixed	Yes	NA (0.55)
Lieberman and Baker, ⁷² 1985	Psychiatrists (NS)	50	6	Emergency department	Sequential interview	NS	No	NA (0.62)
Mellsop et al, ⁷³ 1991	Psychiatrist (5)	60	32 ^c	Inpatient	Joint interview	No	Yes	NA (0.70)
Buchwald and Rudick-Davis, ⁷⁴ 1993	Psychiatry residents (25) Psychology trainee (1)	43	38	Emergency department	Joint or sequential interview	Mixed	Yes	88% (0.74)
Warner and Peabody, ⁷⁵ 1995	Psychiatry residents (30) Psychiatrists (6)	190	74	Emergency department	Sequential interview	NS	No	NA (0.64)

Abbreviations: MDD, major depressive disorder; NA, not available (not reported by authors); NS, not stated.

^aYes indicates history obtained independently by 2 or more observers; mixed, history obtained independently for some but not all subjects.

^bYes indicates diagnosis made without knowledge of other examiners' diagnosis.

^cPatients had affective disorder rather than the more specific MDD.

Conversely, if depressive symptoms are misattributed to a concurrent physical illness, then effective depression treatment may be withheld. A number of strategies have been proposed in an attempt to improve the accuracy and reliability of diagnosis in physically ill patients. The “inclusive” approach counts depressive symptoms toward the diagnosis of depression, regardless of whether the clinician judges that the symptom is due to medical or psychological causes. The *DSM-IV* criteria use an “etiological” approach that counts symptoms toward a major depression diagnosis unless the symptom is “clearly and fully accounted for by a general medical condition.”¹¹ Because clinicians must make a judgment about the cause of individual symptoms, this approach may be less reliable than the inclusive approach. A third strategy, called the “substitutive” approach, replaces depression criterion symptoms that are most likely to be confused with medical illness (eg, loss of energy, weight loss,

impaired concentration) with psychological symptoms.⁸⁵ This approach was developed in an attempt to better discriminate between patients with depression and physical illness and those with only physical illness. Koenig et al⁸⁶ evaluated these strategies in a consecutive series of elderly, hospitalized patients. The prevalence of major depression was 21% using the inclusive approach, 16% using the etiologic approach, and 15% using the substitutive approach. Measures of depression severity and the need for treatment did not differ significantly across the 3 diagnostic approaches. For minor depression, both the prevalence and measures of severity varied more significantly. In a related study, interobserver agreement among mental health professionals was slightly higher for the inclusive approach ($\kappa = 1.0$) than for the etiologic approach ($\kappa = 0.88$).⁸⁷ Two other studies have shown high levels of agreement between the etiologic and substitutive approaches.^{88,89} Although the data are limited, these studies

show high concordance between the different approaches and high interobserver agreement in physically ill patients. Because the substitutive approach requires learning new criteria and does not offer a clear advantage, we recommend the inclusive or etiologic approaches.

How Can I Improve My Skills for Diagnosing Depression?

Observational and trial data suggest that specific communication and interviewing skills are related to diagnostic performance. Three studies using “standardized patients,” or actors presenting with a scripted set of complaints, suggest that physicians are more likely to recognize or diagnose depression when they ask questions about feelings or psychosocial issues.⁹⁰⁻⁹² In one of these studies, recognition rates approached 100% if physicians asked about mood and anhedonia.⁹³

We did not identify any trials designed to improve the accuracy or reliability of diagnostic interviews for depression. Existing trials focus primarily on improving recognition rates, or sensitivity, which is only one aspect of diagnostic accuracy. Four randomized controlled trials of continuing medical education programs for physicians ($n = 329$) show generally positive results.⁹³⁻⁹⁶ Three of the trials focused specifically on or included recognizing depression⁹³⁻⁹⁵ and the fourth trial focused more generally on communication skills training designed to address patients’ emotional distress.⁹⁶ Trained vs untrained physicians were significantly more likely to recognize depression or psychosocial problems in the 2 trials that provided 8-hour training sessions and emphasized communication or interviewing skills^{94,95} or in the trial that provided access to an on-site consulting psychiatrist after a shorter training session.⁹³ These data suggest that motivated physicians can improve their communication skills and sensitivity to emotional distress or depressive disorder. Medical schools and residency programs should consider incorporating similar training in their curricula.

CLINICAL SCENARIO—RESOLUTION

You follow up on Mr P’s “frazzled” comment and learn that he has been under intense work-related stress. Knowing that recent stress increases the likelihood of clinical depression, you ask, “Have you been feeling sad or depressed much of the time?” Mr P has been feeling down nearly every day for several weeks and on further questioning meets criteria for major depression. A focused review of systems and physical examination does not suggest a secondary cause of depression. He does not drink alcohol and has no history of depression. You discuss both antidepressant medication and psychological treatment options for depression.

Author Affiliations at the Time of the Original Publication

The South Texas Veterans Health Care System, Audie Murphy Division, San Antonio (Drs Williams, Noël, and Cordes); San Antonio Evidence-based Practice Center (Dr Ramirez) and Department of Psychiatry (Dr Cordes), University of Texas Health Science Center at San Antonio, San Antonio; and Department of Medicine, University of North Carolina

and RTI-UNC Evidence-based Practice Center, Chapel Hill (Dr Pignone). Dr Williams is now with the Center for Health Services Research in Primary Care, HSR&D, Department of Veterans Affairs Medical Center, and Duke University Medical Center, Durham, North Carolina.

Acknowledgments

This study was supported by the San Antonio Veterans Affairs Health Services Research and Development Center of Excellence and the Agency for Healthcare Quality and Research.

The views expressed in this article are those of the authors and do not necessarily represent the views of the Department of Veterans Affairs.

We thank Ramon Velez, MD, David Steffens, MD, MHSc, and David Simel, MD, MHSc, for their careful manuscript review, and Elizabeth O. Cain, BA, who assisted with manuscript preparation.

REFERENCES

- Depression Guideline Panel. *Depression in Primary Care, I: Detection and Diagnosis: Clinical Practice Guideline No. 5*. Rockville, MD: US Dept of Health and Human Services, Public Health Service, Agency for Health Care Policy and Research; 1993. AHCPR Publication 93-0550.
- World Health Organization. Depression. http://www.who.int/mental_health/management/depression/definition/en/. Accessed May 28, 2008.
- Mulrow CD, Williams JW Jr, Trivedi M, et al. *Treatment of Depression: New Pharmacotherapies*. Rockville, MD: Agency for Health Care Policy and Research; 1999. AHCPR Publication 99-E014.
- Depression Guideline Panel. *Depression in Primary Care, II: Treatment of Major Depression*. Rockville, MD: US Dept of Health and Human Services, Public Health Service; 1993. AHCPR Publication 93-0551.
- Wells KB, Sherbourne C, Schoenbaum M, et al. Impact of disseminating quality improvement programs for depression in managed primary care: a randomized controlled trial. *JAMA*. 2000;283(2):212-220.
- Katon W, Von Korff M, Lin E, et al. Collaborative management to achieve treatment guidelines: impact on depression in primary care. *JAMA*. 1995;273(13):1026-1031.
- Katon W, von Korff M, Lin E, et al. Collaborative management to achieve depression treatment guidelines. *J Clin Psychiatry*. 1997;58(suppl 1):20-23.
- Hunkeler EM, Meresman JF, Hargreaves WA, et al. Efficacy of nurse tele-health care and peer support in augmenting treatment of depression in primary care. *Arch Fam Med*. 2000;9(8):700-708.
- Rost K, Nutting PA, Smith J, Werner J, Duan N. Designing and implementing a primary care intervention trial to improve the quality and outcome of care for major depression. *Gen Hosp Psychiatry*. 2000;22(2):66-77.
- Katzelnick DJ, Simon GE, Pearson SD, et al. Randomized trial of a depression management program in high utilizers of medical care. *Arch Fam Med*. 2000;9(4):345-351.
- American Psychiatric Publishing. *DSM-IV-TR Diagnostic and Statistical Manual of Mental Disorders* (Fourth Edition, Text Revision). <http://www.psychiatryonline.com/content.aspx?aid=2016>. Accessed May 28, 2008.
- Pincus HA, Davis WW, McQueen LE. “Subthreshold” mental disorders: a review and synthesis of studies on minor depression and other “brand names.” *Br J Psychiatry*. 1999;174:288-296.
- Pignone MP, Gaynes BN, Rushton JL, et al. Screening for depression in adults: a summary of the evidence for the US Preventive Services Task Force. *Ann Intern Med*. 2002;136(10):765-766.
- Kroenke K, Spitzer RL, Williams JB, et al. Physical symptoms in primary care: predictors of psychiatric disorders and functional impairment. *Arch Fam Med*. 1994;3(9):774-779.
- Kroenke K, Jackson JL, Chamberlin J. Depressive and anxiety disorders in patients presenting with physical complaints: clinical predictors and outcome. *Am J Med*. 1997;103(5):339-347.

16. Zimmerman M, Lish JD, Lush DT, Farber NJ, Plescia G, Kuzma MA. Suicidal ideation among urban medical outpatients. *J Gen Intern Med.* 1995;10(10):573-576.
17. Zimmerman M. *Diagnosing DSM-IV Psychiatric Disorders in Primary Care Settings: An Interview Guide for the Nonpsychiatrist Physician*. East Greenwich, RI: Psych Products Press; 1994.
18. Hirschfeld RMA, Russell JM. Assessment and treatment of suicidal patients. *N Engl J Med.* 1997;337(13):910-915.
19. Whooley MA, Simon GE. Primary care: managing depression in medical outpatients. *N Engl J Med.* 2000;343(26):1942-1950.
20. Goodwin FK, Bunney WE Jr. Depressions following reserpine: a reevaluation. *Semin Psychiatry.* 1971;3(4):435-448.
21. Lewis DA, Smith RE. Steroid-induced psychiatric syndromes: a report of 14 cases and a review of the literature. *J Affect Disord.* 1983;5(4):319-332.
22. Pope HG Jr, Katz DL. Affective and psychotic symptoms associated with anabolic steroid use. *Am J Psychiatry.* 1988;145(4):487-490.
23. Bottomley A. Depression in cancer patients: a literature review. *Eur J Cancer Care (Engl).* 1998;7(3):181-191.
24. Gavard JA, Lustman PJ, Clouse RE. Prevalence of depression in adults with diabetes: an epidemiological evaluation. *Diabetes Care.* 1993;16(8):1167-1178.
25. Fifield J, Tennen H, Reisine S, McQuillan J. Depression and the long-term risk of pain, fatigue, and disability in patients with rheumatoid arthritis. *Arthritis Rheum.* 1998;41(10):1851-1857.
26. Schleifer SJ, Macari-Hinson MM, Coyle DA, et al. The nature and course of depression following myocardial infarction. *Arch Intern Med.* 1989;149(8):1785-1789.
27. Ballenger JC, Davidson JR, Lecrubier Y, et al. Consensus statement on the primary care management of depression from the International Consensus Group on Depression and Anxiety. *J Clin Psychiatry.* 1999;60(suppl 7):54-61.
28. Buchsbaum DG, Buchanan RG, Centor RM, Schnoll SH, Lawton MJ. Screening for alcohol abuse using CAGE scores and likelihood ratios. *Ann Intern Med.* 1991;115(10):774-777.
29. World Health Organization. *International Classification of Disease, Version 2007*. <http://www.who.int/classifications/apps/icd/icd10online/>. Accessed May 28, 2008.
30. The Cochrane Library, Issue 3, 2000. Chichester: Wiley. The Cochrane Central Register of Controlled Trials. http://www.mrw.interscience.wiley.com/cochrane/cochrane_ccentral_articles_fs.html. Accessed May 28, 2008.
31. Mulrow CD, Williams JW Jr, Gerety MB, Kerber CA, Ramirez G. Case-finding instruments for depression in primary care settings. *Ann Intern Med.* 1995;122(12):913-921.
32. First MB, Spitzer RL, Gibbon M, Williams JBW. *Structured Clinical Interview for DSM-IV-TR Axis I Disorders, Research Version, Patient Edition (SCID-I/P)*. New York, NY: Biometrics Research, New York State Psychiatric Institute; 2002.
33. Gunning R. *The Technique of Clear Writing*. New York, NY: McGraw-Hill International; 1952.
34. Whooley MA, Avins AL, Miranda J, Browner WS. Case-finding instruments for depression: two questions are as good as many. *J Gen Intern Med.* 1997;12(7):439-445.
35. Steer RA, Cavalieri TA, Leonard DM, Beck AT. Use of the Beck Depression Inventory for Primary Care to screen for major depression disorders. *Gen Hosp Psychiatry.* 1999;21(2):106-111.
36. Perez-Stable EJ, Miranda J, Munoz RF, Ying YW. Depression in medical outpatients: underrecognition and misdiagnosis. *Arch Intern Med.* 1990;150(5):1083-1088.
37. Zich JM, Attkisson CC, Greenfield TK. Screening for depression in primary care clinics: the CES-D and the BDI. *Int J Psychiatry Med.* 1990;20(3):259-277.
38. Kirmayer LJ, Robbins JM, Dworkind M, Yaffe MJ. Somatization and the recognition of depression and anxiety in primary care. *Am J Psychiatry.* 1993;150(5):734-741.
39. Williams JW Jr, Mulrow CD, Kroenke K, et al. Case-finding for depression in primary care: a randomized trial. *Am J Med.* 1999;106(1):36-43.
40. Fechner-Bates S, Coyne JC, Schwenk TL. The relationship of self-reported distress to depressive disorders and other psychopathology. *J Consult Clin Psychol.* 1994;62(3):550-559.
41. Hendrie HC, Callahan CM, Levitt EE, et al. Prevalence rates of major depressive disorders: the effects of varying the diagnostic criteria in an older primary care population. *Am J Geriatr Psychiatry.* 1995;2:119-131.
42. Hough R, Landsverk J, Stone J, et al. *Comparison of Psychiatric Screening Questionnaires for Primary Care Patients*. Bethesda, MD: National Institute of Mental Health; 1983. Final report for NIMH contract 278-81-0036 (DB) edition.
43. Irwin M, Artin KH, Oxman MN. Screening for depression in the older adult: criterion validity of the 10-item Center for Epidemiological Studies Depression Scale (CES-D). *Arch Intern Med.* 1999;159(15):1701-1704.
44. Lyness JM, Noel TK, Cox C, King DA, Conwell Y, Caine ED. Screening for depression in elderly primary care patients: a comparison of the Center for Epidemiologic Studies-Depression Scale and the Geriatric Depression Scale. *Arch Intern Med.* 1997;157(4):449-454.
45. Neal RM, Baldwin RC. Screening for anxiety and depression in elderly medical outpatients. *Age Ageing.* 1994;23(6):461-464.
46. D'Ath P, Katona P, Mullan E, Evans S, Katona C. Screening, detection and management of depression in elderly primary care attenders, I: the acceptability and performance of the 15 item Geriatric Depression Scale (GDS15) and the development of short versions. *Fam Pract.* 1994;11(3):260-266.
47. Schmitz N, Kruse J, Heckrath C, Alberti L, Tress W. Diagnosing mental disorders in primary care: the General Health Questionnaire (GHQ) and the Symptom Check List (SCL-90-R) as screening instruments. *Soc Psychiatry Psychiatr Epidemiol.* 1999;34(7):360-366.
48. Spitzer RL, Kroenke K, Williams JBW, and the Patient Health Questionnaire Primary Care Study Group. Validation and utility of a self-report version of PRIME-MD: the PHQ primary care study. *JAMA.* 1999;282(18):1737-1744.
49. Spitzer RL, Williams JB, Kroenke K, et al. Utility of a new procedure for diagnosing mental disorders in primary care: the PRIME-MD 1000 study. *JAMA.* 1994;272(22):1749-1756.
50. Leon AC, Olfson M, Weissman MM, et al. Brief screens for mental disorders in primary care. *J Gen Intern Med.* 1996;11(7):426-430.
51. Broadhead WE, Leon AC, Weissman MM, et al. Development and validation of the SDDS-PC screen for multiple mental disorders in primary care. *Arch Fam Med.* 1995;4(3):211-219.
52. Okimoto JT, Barnes RF, Veith RC, Raskind MA, Inui TS, Carter WB. Screening for depression in geriatric medical patients. *Am J Psychiatry.* 1982;139(6):799-802.
53. Magruder-Habib K, Zung WW, Feussner JR. Improving physicians' recognition and treatment of depression in general medical care: results from a randomized clinical trial. *Med Care.* 1990;28(3):239-250.
54. Raft D, Spencer RF, Toomey T, Brogan D. Depression in medical outpatients: use of the Zung scale. *Dis Nerv Syst.* 1977;38(12):999-1004.
55. Klinkman MS, Coyne JC, Gallo S, Schwenk TL. Can case-finding instruments be used to improve physician detection of depression in primary care? *Arch Fam Med.* 1997;6(6):567-573.
56. Schulberg HC, Saul M, McClelland M, Ganguli M, Christy W, Frank R. Assessing depression in primary medical and psychiatric practices. *Arch Gen Psychiatry.* 1985;42(12):1164-1170.
57. Salokangas RK, Poutanen O, Stengard E. Screening for depression in primary care: development and validation of the Depression Scale, a screening instrument for depression. *Acta Psychiatr Scand.* 1995;92(1):10-16.
58. Parkerson GR Jr, Broadhead WE. Screening for anxiety and depression in primary care with the Duke Anxiety-Depression Scale. *Fam Med.* 1997;29(3):177-181.
59. van Marwijk HW, Wallace P, de Bock GH, Hermans J, Kaptein AA, Mulder JD. Evaluation of the feasibility, reliability and diagnostic value of shortened versions of the geriatric depression scale. *Br J Gen Pract.* 1995;45(393):195-199.
60. Arthur A, Jagger C, Lindesay J, Graham C, Clarke M. Using an annual over-75 health check to screen for depression: validation of the short Geriatric Depression Scale (GDS15) within general practice. *Int J Geriatr Psychiatry.* 1999;14(6):431-439.
61. Nettelbladt P, Hansson L, Stefansson CG, Borgquist L, Nordstrom G. Test characteristics of the Hopkins Symptom Check List-25 (HSCL-25) in Sweden, using the Present State Examination (PSE-9) as a caseness criterion. *Soc Psychiatry Psychiatr Epidemiol.* 1993;28(3):130-133.
62. Fuhrer R, Rouillon F, Lelouch J. Diagnostic reliability among French psychiatrists using DSM-III criteria. *Acta Psychiatr Scand.* 1986;73(1):12-16.
63. Riskind JH, Beck AT, Berchick RJ, Brown G, Steer RA. Reliability of DSM-III diagnoses for major depression and generalized anxiety disorder using the structured clinical interview for DSM-III. *Arch Gen Psychiatry.* 1987;44(9):817-820.

64. Stukenberg KW, Dura JR, Kiecolt-Glaser JK. Depression screening scale validation in an elderly, community-dwelling population. *Psychol Assess*. 1990;2:134-138.
65. Skre I, Onstad S, Torgersen S, Kringsen E. High interrater reliability for the Structured Clinical Interview for DSM-III-R Axis I (SCID-I). *Acta Psychiatr Scand*. 1991;84(2):167-173.
66. Williams JB, Gibbon M, First MB, et al. The Structured Clinical Interview for DSM-III-R (SCID), II: multisite test-retest reliability. *Arch Gen Psychiatry*. 1992;49(8):630-636.
67. Segal DL, Hersen M, Van Hasselt VB. Reliability of the Structured Clinical Interview for DSM-III-R: an evaluative review. *Compr Psychiatry*. 1994;35(4):316-327.
68. Keller MB, Klein DN, Hirschfeld RM, et al. Results of the DSM-IV mood disorders field trial. *Am J Psychiatry*. 1995;152(6):843-849.
69. Spitzer RL, Forman JB, Nee J. DSM-III field trials, I: initial interrater diagnostic reliability. *Am J Psychiatry*. 1979;136(6):815-817.
70. Webb LJ, Gold RS, Johnstone EE, Diclemente CC. Accuracy of DSM-III diagnoses following a training program. *Am J Psychiatry*. 1981;138(3):376-378.
71. Hyler SE, Williams JB, Spitzer RL. Reliability in the DSM-III field trials: interview v case summary. *Arch Gen Psychiatry*. 1982;39(11):1275-1278.
72. Lieberman PB, Baker FM. The reliability of psychiatric diagnosis in the emergency room. *Hosp Community Psychiatry*. 1985;36(3):291-293.
73. Mellsoop GW, Thomas CS, Ellis PM, Purdie G, Crawshaw J, Mendis N. Reliability of the draft diagnostic criteria for research of ICD-10 in comparison with ICD-10 and DSM-III-R. *Acta Psychiatr Scand*. 1991;84(4):332-335.
74. Buchwald AM, Rudick-Davis D. The symptoms of major depression. *J Abnorm Psychol*. 1993;102(2):197-205.
75. Warner MD, Peabody CA. Reliability of diagnoses made by psychiatric residents in a general emergency department. *Psychiatr Serv*. 1995;46(12):1284-1286.
76. Greenes RA, Begg CB. Assessment of diagnostic technologies: methodology for unbiased estimation from samples of selectively verified patients. *Invest Radiol*. 1985;20(7):751-756.
77. Sackett DL. A primer on the precision and accuracy of the clinical examination. *JAMA*. 1992;267(19):2638-2644.
78. Simel DL, Samsa GP, Matchar DB. Likelihood ratios with confidence: sample size estimation for diagnostic test studies. *J Clin Epidemiol*. 1991;44(8):763-770.
79. Irwig L, Macaskill P, Glasziou P, Fahey M. Meta-analytic methods for diagnostic test accuracy. *J Clin Epidemiol*. 1995;48(1):119-130.
80. Hasselblad V, Hedges LV. Meta-analysis of screening and diagnostic tests. *Psychol Bull*. 1995;117(1):167-178.
81. Norden C, Philipps E, Levy P, Kass E. Variation in interpretation of intravenous pyelograms. *Am J Epidemiol*. 1970;91(2):155-160.
82. Edelman D, Hough DM, Glazebrook KN, Oddone EZ. Prognostic value of the clinical examination of the diabetic foot ulcer. *J Gen Intern Med*. 1997;12(9):537-543.
83. Deyo RA, McHugh LM, Cone RO 3rd. Observer variability in the interpretation of lumbar spine radiographs. *Arthritis Rheum*. 1985;28(9):1066-1070.
84. Koenig HG, Cohen HJ, Blazer DG, Krishnan KR, Sibert TE. Profile of depressive symptoms in younger and older medical inpatients with major depression. *J Am Geriatr Soc*. 1993;41(11):1169-1176.
85. Endicott J. Measurement of depression in patients with cancer. *Cancer*. 1984;53(10 suppl):2243-2249.
86. Koenig HG, George LK, Peterson BL, Pieper CF. Depression in medically ill hospitalized older adults: prevalence, characteristics, and course of symptoms according to six diagnostic schemes. *Am J Psychiatry*. 1997;154(10):1376-1383.
87. Koenig HG, Pappas P, Holsinger T, Bachar JR. Assessing diagnostic approaches to depression in medically ill older adults: how reliably can mental health professionals make judgments about the cause of symptoms? *J Am Geriatr Soc*. 1995;43(5):472-478.
88. Rapp SR, Vrana S. Substituting nonsomatic for somatic symptoms in the diagnosis of depression in elderly male medical patients. *Am J Psychiatry*. 1989;146(9):1197-2000.
89. Kathol RG, Mutgi A, Williams JW Jr, Clamon G, Noyes R. Diagnosis of major depression in cancer patients according to four sets of criteria. *Am J Psychiatry*. 1990;147(8):1021-1024.
90. Stange KC, Fedirko T, Zyzanski SJ, Jaen CR. How do family physicians prioritize delivery of multiple preventive services? *J Fam Pract*. 1994;38(3):231-237.
91. Carney PA, Dietrich AJ, Eliassen MS, Owen M, Badger LW. Recognizing and managing depression in primary care: a standardized patient study. *J Fam Pract*. 1999;48(12):965-972.
92. Carney PA, Eliassen MS, Wolford GL, Owen M, Badger LW, Dietrich AJ. How physician communication influences recognition of depression in primary care. *J Fam Pract*. 1999;48(12):958-964.
93. Worrall G, Angel J, Chaulk P, Clarke C, Robbins M. Effectiveness of an educational strategy to improve family physicians' detection and management of depression: a randomized controlled trial. *CMAJ*. 1999;161(1):37-40.
94. Gerrity MS, Cole SA, Dietrich AJ, Barrett JE. Improving the recognition and management of depression: is there a role for physician education? *J Fam Pract*. 1999;48(12):949-957.
95. Thompson C, Kinmonth AL, Stevens L, et al. Effects of a clinical-practice guideline and practice-based education on detection and outcome of depression in primary care: Hampshire Depression Project randomised controlled trial. *Lancet*. 2000;355(9199):185-191.
96. Roter DL, Hall JA, Kern DE, Barker LR, Cole KA, Roca RP. Improving physicians' interviewing skills and reducing patients' emotional distress: a randomized clinical trial. *Arch Intern Med*. 1995;155(17):1877-1884.

Prepared by John W. Williams Jr, MD, MHS

Reviewed by David Steffens, MD, MHS

CLINICAL SCENARIO

You decide to implement the US Preventive Services Task Force (USPSTF) recommendation to screen for depression. Which questionnaire will you use? Will you ask the questions yourself or ask your staff to administer the questionnaire as part of the check-in process? Will you screen all adults or screen more selectively? What other care components should be in place to effectively follow through on the screening results?

UPDATED SUMMARY ON SCREENING FOR DEPRESSION

Original Review

Williams JW Jr, Noel PH, Cordes JA, Ramirez G, Pignone M. Is this patient clinically depressed? *JAMA*. 2002;287(9):1160-1170.

UPDATED LITERATURE SEARCH

The high prevalence of depressive disorders, suboptimal recognition rates, and availability of efficacious treatments has long provided the impetus to evaluate screening approaches. The USPSTF updated their recommendations (2002) according to new evidence concerning the validity of screening instruments, the effectiveness of screening, and treatment approaches that ensure adequate follow-up.¹ We conducted an updated MEDLINE search for English-language medical literature published between 2000 and August 2004 evaluating the performance of depression case-finding instruments in primary care settings. Search terms were “depressive disorder” or “depression,” textword terms for each instrument, and a search filter for articles on diagnosis. The search yielded 307 articles; an additional 5 articles were identified from citations. We retained studies from primary care settings that administered a case-finding instrument and used a standard interview such as the Structured Clinical Interview for DSM-IV-TR (SCID) to make a criterion-based diagnosis (eg, *Diagnostic and Statistical Manual of Mental Disorders* [Third Edition Revised] [DSM-III-R], and *Diagnostic and Statistical Manual of Mental Disorders* [Fourth Edition] [DSM-IV]).^{2,3} We limited these studies further with requirements that case-finding instruments have (1) easy to average literacy, (2) scores estimable without a calculator, (3) a depression-specific component, and (4) evaluation in at least 1 study of 100 or more subjects. After

review of titles and abstracts, 25 articles underwent full-text review; 5 met all eligibility criteria. We also retrieved a recent systematic review published by the USPSTF.

NEW FINDINGS

- In adults, 2- to 9-item screening instruments perform comparably to longer depression questionnaires.
- Ultrashort questionnaires, such as the 2-item Primary Care Evaluation of Mental Disorders (PRIME-MD), can be administered easily in writing or verbally (see Appendix).
- The brief 9-item Patient Health Questionnaire (PHQ-9) gives the best discriminatory power and more diagnostic information for depression diagnoses. The PHQ-9 can also quantify treatment responses (see Appendix).
- Instruments tailored to subgroups (eg, older adults) lack proof of superiority to instruments developed for general primary care populations.
- Brief, well-validated instruments have not been developed for children treated in primary care settings.

Details of the Update

IMPROVEMENTS IN THE DATA PRESENTED IN THE ORIGINAL PUBLICATION

The data presented in the original publication have not changed; instead, recent literature provides more information and revised estimates of the sensitivity, specificity, and likelihood ratios (LRs).

CHANGES IN THE REFERENCE STANDARD

The reference standard for depressive disorders remains the criterion in the *DSM-IV* text revision.³

RESULTS OF THE LITERATURE REVIEW

Three brief screening instruments were identified; a total of 14 instruments have now been evaluated in primary care. The Hospital Anxiety and Depression Scale (HADS),⁴ the World Health Organization-Five Well-Being Scale (WHO-5),⁵ and the Yale

1-question screen⁶ range from 1 to 7 questions and take fewer than 2 minutes to administer. The HADS, designed for medically ill inpatients and outpatients, has 7-item depression and anxiety subscales in which a score of 8 or higher is considered a positive result (range, 0-21). The WHO-5 is a 5-item quality of life measure in which scores of 12 or lower are considered a positive result (range, 0-25). Although the WHO-5 is described as a quality-of-life measure, it asks about specific depression symptoms. The Yale 1-question screen asks, “Do you often feel sad or depressed?” In addition to these 3 new instruments, new data were published on the PHQ-9 and the 2-item version of the PHQ and the PRIME-MD,^{7,8} the Geriatric Depression Scale (GDS),⁹ and the Center for Epidemiological Studies Depression Scale (CES-D).⁹ Five studies evaluated these instruments in 5652 adult patients, of whom 1653 underwent criterion interviews. Studies were conducted in the United States ($n = 2$), Germany ($n = 2$), and New Zealand.

These studies add to previous evidence that brief, 2- to 9-item screening instruments perform comparably to or better than longer questionnaires (see Table 19-6). In a high-quality study, Henkel et al¹⁰ compared the PHQ-9 and WHO-5 with the 12-item General Health Questionnaire (GHQ-12), a general measure of psychological well-being. The PHQ-9 had a significantly higher positive likelihood ratio (LR+) (5.2; 95% confidence interval [CI], 3.9-6.8) than the WHO-5 (LR+, 2.6; 95% CI, 2.2-3.0) or GHQ-12 (LR+, 2.2; 95% CI, 1.9-2.6). The PHQ-9 negative likelihood ratio (LR-) (0.26; 95% CI, 0.17-0.4) was comparable to the WHO-5 (LR-, 0.11; 95% CI, 0.05-0.25) and GHQ-12 (LR-, 0.24; 95% CI, 0.14-0.42). Three other studies support these findings. Löwe et al¹¹ compared the PHQ-9 to the HADS and WHO-5 in 1619 patients drawn from academic-affiliated family medicine practices. The PHQ-9 had a signifi-

cantly higher LR+ and a comparable LR- to other instruments. Kroenke et al^{7,8} combined data on the PHQ-9 from primary care medical and obstetrics and gynecology populations.^{12,13} This high-quality study found an LR+ of 7.3 (95% CI, 5.6-9.4) and LR- of 0.14 (95% CI, 0.06-0.32). Collectively, these studies show that the PHQ-9 performs better than other brief instruments in head-to-head comparisons and has LRs that are comparable or superior to other longer depression questionnaires. The PHQ-9 has the advantage of asking specifically about *DSM-IV* criterion symptoms for major depression and has been shown responsive to change in clinical status. Therefore, it provides essential symptom data for a diagnostic interview and can also monitor treatment responses.

Depression instruments are typically given to patients in paper form for self-administration. In 15 New Zealand general practices, physicians verbally administered the 2-item PRIME-MD to 421 consecutive patients.¹⁴ Verbal administration performed well, with an LR+ of 2.9 (95% CI, 2.5-3.5) and LR- of 0.05 (LR-, 0.02-0.11). These performance characteristics using verbal administration are almost identical to previous studies of the PRIME-MD that used self-administration. For clinicians wishing to adopt a streamlined approach to depression screening, this study supports verbal administration of the simple 2-item screen.

Another salient issue is how well depression screening instruments perform in important subgroups such as older adults, the medically ill, and children. One study evaluated the Yale 1-question, the PRIME-MD, the GDS, and CES-D in 360 adults aged 60 years or older; 125 of 360 patients were recruited from primary care settings.⁸ The PRIME-MD performed less well than studies conducted in mixed age populations, whereas the GDS and CES-D had statistically similar diagnostic odds ratios. However, these results were biased because major depression diagnoses were made blind to PRIME-MD screening results but with knowledge of the GDS and CES-D screening results. The Yale 1-question performed poorly. For older adults, these data raise the possibility that brief, general depression instruments perform worse than longer instruments or those designed specifically for older adults. This hypothesis remains unproven and needs testing in a larger, higher-quality study with subgroup analyses of older and younger patients. As discussed above, the PHQ-9 performed better than the HADS in primary care patients, although comparisons in patients with severe or multiple chronic medical illness are not available. Finally, we could not identify well-validated instruments for children and adolescents in primary care settings.

Table 19-6 Likelihood Ratios of Brief Screening Instruments to Identify Depression or Dysthymia

Screening Test (No. of Studies)	Major Depressive Disorder	
	Summary LR+ (95% CI) or Range	Summary LR- (95% CI) or Range
Yale 1 item (1)	1.8 (1.1-2.8)	0.56 (0.27-1.14)
PRIME-MD (4)	2.6 (2.1-3.2)	0.15 (0.08-0.28)
WHO-5 (1)	2.2 (2.0-2.4)	0.01 (0-0.2)
HADS (2)	2.9 (2.5-3.5)	0.25 (0.15-0.41)
PHQ-9 (2) ^a	4.9-7.3	0.02-0.14
GDS (3) ^b	2.4-4.0	0.12-0.32
Major Depressive Disorder or Dysthymia		
WHO-5 (2)	2.4 (2.2-2.7)	0.11 (0.06-0.19)
HADS (1)	3.2 (2.7-3.9)	0.25 (0.18-0.37)
PHQ-9 (3)	5.9 (4.2-8.3)	0.29 (0.23-0.38)

Abbreviations: CI, confidence interval; GDS, Geriatric Depression Scale; HADS, Hospital Anxiety and Depression Scale; LR+, positive likelihood ratio; LR-, negative likelihood ratio; PHQ-9, 9-item Patient Health Questionnaire; PRIME-MD, Primary Care Evaluation of Mental Disorders; WHO-5, World Health Organization—Five Well-Being Scale.

^aData are the range of values across studies; summary statistics were not calculated because 2 thresholds (10 and 11) were used.

^bData are the range of values across the studies; summary statistics were not calculated because of varying thresholds and instrument versions.

EVIDENCE FROM GUIDELINES

The USPSTF recommends “screening adults for depression in clinical practices that have systems in place to assure accurate diagnosis, effective treatment and follow-up.”¹⁵ This recommendation came from an evidence synthesis showing that although trials of screening alone had small benefits, even larger benefits accrued in trials that used feedback of depression screening results coordinated with effective treatment and follow-up. No specific screening instrument was recommended, but it was observed that the 2-item PRIME-MD might be as effective as

longer instruments. The task force recommended that clinicians choose the method that best fits their personal preference, the patient population served, and the practice setting. The optimal screening interval is unknown, and all positive screening test results should trigger full diagnostic interviews.

The USPSTF concluded that evidence is insufficient to recommend for or against routine screening of children or adolescents for depression. They found limited evidence on the accuracy and reliability of screening tests in children and adolescents.

The Canadian Task Force on Preventive Health Care updated its guidelines in 2005,¹⁶ relying on the same 2002 evidence synthesis used by the USPSTF. Their recommendations were concordant with the USPSTF. Highlighted issues were as follows:

1. “Recurrent screening may be most productive in patients with a history of depression, unexplained somatic symptoms, comorbid psychological conditions, substance abuse, or chronic pain.”
2. Elements of systems that ensured good depression care were education of the patients, health care providers, or both, a mechanism to ensure that screening results are reported to the patient’s clinician, who can confirm the diagnosis and provide appropriate treatment, access to case management or mental health care, and telephone follow-up.

Web Resources for Depression Screening

The US Preventive Services Task Force Recommendation on depression screening is available online (<http://www.ahrq.gov/clinic/uspstf/uspsdepr.htm>). Accessed May 28, 2008).¹

CLINICAL SCENARIO—RESOLUTION

During the past year, your practice has focused on improving preventive care, and your staff is receptive to implementing depression screening. Keeping in mind the tensions between practice efficiency, costs, and the benefits of improved patient outcomes, you decide on the following strategy. You elect to screen all adults annually for depression. By screening all adults, you establish a consistent approach for your staff that simplifies logistics. With input from your nurses, you decide that nurses will verbally ask the 2 PRIME-MD questions (a brief survey that makes you feel confident that you will not miss many depressed patients) and that patients with positive results will be given the PHQ-9 to complete (a slightly longer survey that helps you assess whether the patients who have positive results on the 2-item survey really are depressed).

You believe your practice has an average prevalence of major depression ($\approx 7\%$); hence, patients who have negative results on the 2-item PRIME-MD (LR-, 0.15) will have a posttest probability of 1% for major depression. Patients who have positive results on the PRIME-MD and then score 10 or higher on the PHQ-9 (LR+, 4.9-7.3) will have a posttest probability of 49% to 59% for major depression. Patients who have positive results on the 2-item PRIME-MD but have a more normal score or lower than 10 on the PHQ-9 (LR-, 0.02-0.14) will have a posttest probability of 1% to 4% for major depression. All patients who score 10 or higher on the PHQ-9 will require further evaluation to establish a specific diagnosis.

See next page for the “Make the Diagnosis” section.

SCREENING FOR DEPRESSION—MAKE THE DIAGNOSIS

PRIOR PROBABILITY

Although most depression screening focuses on major depressive disorders, a number of treatable conditions may be detected by depression screening. In primary care, the combined prevalence of depression and dysthymia is approximately 7% to 12% ([Table 19-7](#)).

Table 19-7 Prevalence of Depression and Dysthymia in Primary Care

Illness	Prevalence, %
Major depressive disorder	4.8-8.6
Dysthymic disorder	2.1-3.7
Depression NOS	4.4-5.4

Abbreviation: NOS, not otherwise specified.

POPULATIONS FOR WHOM DEPRESSION SHOULD BE ASSESSED

All adults.

DETECTING THE LIKELIHOOD OF MAJOR DEPRESSIVE DISORDER

See [Table 19-8](#).

Table 19-8 Likelihood Ratios for Detecting Major Depressive Disorder

Instrument and Threshold	Summary LR+ (95% CI) or Range	Summary LR− (95% CI) or Range
PHQ-9 (score ≥ 10 is positive) ^a	4.9-7.3	0.02-0.14
PRIME-MD (score ≥ 1 is positive)	2.6 (2.1-3.2)	0.15 (0.08-0.28)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR−, negative likelihood ratio; PHQ-9, 9-item Patient Health Questionnaire; PRIME-MD, Primary Care Evaluation of Mental Disorders.

^aData are the range of values across studies; summary statistics were not calculated because 2 thresholds (10 and 11) were used.

DETECTING THE LIKELIHOOD OF MAJOR DEPRESSIVE DISORDER OR DYSTHYMIA

See [Table 19-9](#).

Table 19-9 Likelihood Ratios for Detecting Major Depressive Disorder or Dysthymia

Instrument	Summary LR+ (95% CI)	Summary LR− (95% CI)
PHQ-9 (score ≥ 10 is positive)	5.9 (4.2-8.3)	0.29 (0.23-0.38)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR−, negative likelihood ratio; PHQ-9, 9-item Patient Health Questionnaire.

REFERENCE STANDARD TESTS

A structured (eg, Diagnostic Interview Schedule) or semi-structured diagnostic interview (eg, Structured Clinical Interview for DSM-IV) to establish diagnoses.

APPENDIX—DEPRESSION SCREENING INSTRUMENTS

Patient Health Questionnaire (PHQ-9)

See Table 19-10.

Table 19-10 Patient Health Questionnaire (PHQ-9)^a

During the last 2 weeks, how often have you been bothered by any of the following problems?	Not at all	Several days	More than half the days	Nearly every day
1. Little interest or pleasure in doing things	0	1	2	3
2. Feeling down, depressed, or hopeless	0	1	2	3
3. Trouble falling or staying asleep, or sleeping too much	0	1	2	3
4. Feeling tired or having little energy	0	1	2	3
5. Poor appetite or overeating	0	1	2	3
6. Feeling bad about yourself—or that you are a failure or have let yourself or your family down	0	1	2	3
7. Trouble concentrating on things, such as reading the newspaper or watching television	0	1	2	3
8. Moving or speaking so slowly that other people could have noticed, or the opposite—being so fidgety or restless that you have been moving around a lot more than usual	0	1	2	3
9. Thoughts that you would be better off dead or of hurting yourself in some way	0	1	2	3

Scoring: Add up the results for each item. A score ≥ 10 is positive for depression or dysthymia.

If you checked off *any* problems, how *difficult* have these problems made it for you to do your work, take care of things at home, or get along with other people?

Not difficult at all	Somewhat difficult	Very difficult	Extremely difficult
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

^aThe PHQ was developed from the PRIME-MD. PRIME-MD is a trademark of Pfizer Inc. Copyright © 1999, Pfizer Inc. All rights reserved. Reproduced with permission.

PRIME-MD (2 Items)

See Table 19-11.

Table 19-11 PRIME-MD^a (2 Items)

During the past month have you been bothered a lot by:	Yes	No
1. Little interest or pleasure in doing things	0	1
2. Feeling down, depressed, or hopeless	0	1

Scoring: A "yes" answer on either question is considered a positive result for depression.

^aPRIME-MD is a trademark of Pfizer Inc. Copyright © 1999, Pfizer Inc. All rights reserved. Reproduced with permission.

REFERENCES FOR THE UPDATE

- US Preventive Services Task Force. *Screening for Depression: Recommendations and Rationale*. Rockville, MD: Agency for Healthcare Research and Quality; 2002. <http://www.ahrq.gov/clinic/uspstf/uspsdepr.htm>. Accessed May 28, 2008.
- First MB, Spitzer RL, Gibbon M, Williams JBW. *Structured Clinical Interview for DSM-IV-TR Axis I Disorders, Research Version, Patient Edition (SCID-I/P)*. New York, NY: Biometrics Research, New York State Psychiatric Institute; 2002.
- American Psychiatric Publishing. *DSM-IV-TR Diagnostic and Statistical Manual of Mental Disorders (Fourth Edition) Text Revision*. <http://www.psychiatryonline.com/content.aspx?aID=2016>. Accessed May 28, 2008.
- Zigmond AS, Snaith RP. The hospital anxiety and depression scale. *Acta Psychiatr Scand*. 1983;67(6):361-370.
- Psychiatric Research Unit, Frederiksberg General Hospital, Hillerød, Denmark. WHO-5 Questionnaires. <http://www.who-5.org>. Accessed May 28, 2008.
- Lachs MS, Feinstein AR, Cooney LM, et al. A simple procedure for general screening for functional disability in elderly patients. *Ann Intern Med*. 1990;112(9):669-706.
- Kroenke K, Spitzer RL, Williams JBW. The PHQ-9: validity of a brief depression severity measure. *J Gen Intern Med*. 2001;16(9):606-613.^a
- Kroenke K, Spitzer RL, Williams JBW. The Patient Health Questionnaire-2: validity of a two-item depression screener. *Med Care*. 2003;41(11):1284-1292.^a
- Blank K, Gruman C, Robison JT. Case-finding for depression in elderly people: balancing ease of administration with validity in varied treatment settings. *J Gerontol*. 2004;59A(4):378-384.^a
- Henkel V, Mergl R, Kohnen R, Maier W, Möller H, Hegerl U. Identifying depression in primary care: a comparison of different methods in a prospective cohort study. *BMJ*. 2003;326(7382):200-201.^a
- Löwe B, Spitzer RL, Gräfe K, et al. Comparative validity of three screening questionnaires for DSM-IV depressive disorders and physicians' diagnoses. *J Affect Disord*. 2004;78:131-140.^a
- Spitzer RL, Williams JBW, Kroenke K, et al, for the Patient Health Questionnaire Obstetrics-Gynecology Study Group. Validity and utility of the PRIME-MD patient health questionnaire in assessment of 3000 obstetric-gynecologic patients: the PRIME-MD Patient Health Questionnaire Obstetrics-Gynecology Study. *Am J Obstet Gynecol*. 2000;183(3):759-769.
- Spitzer RL, Kroenke K, Williams JBW. Validation and utility of a self-report version of PRIME-MD: the PHQ primary care study. *JAMA*. 1999;282(18):1737-1744.
- Arroll B, Khin N, Kerse N. Screening for depression in primary care with two verbally asked questions: cross sectional study. *BMJ*. 2003;327(7424):1144-1146.^a
- US Preventive Services Task Force, Agency for Healthcare Research and Quality. Screening for depression: recommendations and rationale. *Ann Intern Med*. 2002;136(10):760-764.
- Macmillan HL, Patterson CJ, Wathen CN. Screening for depression in primary care: recommendation statement from the Canadian Task Force on Preventive Health Care. *CMAJ*. 2005;172(1):33-35.

^aFor Evidence to Support the Update for this topic, see <http://www.JAMAEvidence.com>.

This page intentionally left blank

EVIDENCE TO SUPPORT THE UPDATE:

19

Depression

TITLE Screening for Depression in Primary Care With 2 Verbally Asked Questions: Cross-Sectional Study.

AUTHORS Arroll B, Khin N, Kerse N.

CITATION *BMJ*. 2003;327(7424):1144-1146.

QUESTION How well does the 2-question Primary Care Evaluation of Mental Disorders (PRIME-MD)¹ perform for detecting depression?

DESIGN General practitioners asked the 2 screening questions, and major depression diagnoses were made using a structured interview, blind to the screening results.

SETTING Fifteen general practices in Auckland, New Zealand.

PATIENTS Four hundred twenty-one consecutive patients (median age, 46 years) who agreed to participate; 194 who declined, 47 taking psychotropic drugs, and 8 who were not asked the screening questions were excluded.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

The PRIME-MD prompts:

1. “During the past month have you often been bothered by feeling down, depressed, or hopeless?” and
2. “During the past month have you often been bothered by little interest or pleasure in doing things?”

A yes response to either question was considered a positive result. An interviewer used the computer-assisted structured Composite International Diagnostic Interview to establish major depression diagnoses.

MAIN OUTCOME MEASURES

Sensitivity, specificity, and likelihood ratios.

MAIN RESULTS

One hundred fifty-seven (37%) of 421 patients had a positive result; 29 patients (6.8%) were diagnosed with major depression ([Table 19-12](#)).

Table 19-12 Likelihood Ratio for the 2-Question PRIME-MD

Test	Sensitivity	Specificity	LR+ (95% CI)	LR- (95% CI)	DOR (95% CI)
PRIME-MD	0.97	0.67	2.9	0.05	62 (24-156)
2 items			(2.5-3.5)	(0.02-0.11)	

Abbreviations: CI, confidence interval; DOR, diagnostic odds ratio; LR+, positive likelihood ratio; LR-, negative likelihood ratio; PRIME-MD, Primary Care Evaluation of Mental Disorders.

CONCLUSIONS

LEVEL OF EVIDENCE Level 1.

STRENGTHS Consecutive patients were studied. Because there were few exclusion criteria, there was a high participation rate and the patients had a broad age range.

LIMITATIONS Dysthymia, a chronic depressive disorder that is responsive to antidepressants and psychotherapy, is not considered.

This study continues a trend of evaluating brief depression screening instruments. The study methodology was strong; excluding patients taking psychotropic drugs likely skewed the spectrum toward milder major depressive illnesses. The unique contribution of this study is that practitioners simply asked the 2 screening questions, which is logically simple. The high sensitivity and relatively low specificity are consistent with studies using paper versions of the questionnaire.

Reviewed by John W. Williams Jr, MD

REFERENCE FOR THE EVIDENCE

1. Spitzer RL, Williams JB, Kroenke K, et al. Utility of a new procedure for diagnosing mental disorders in primary care: the PRIME-MD 1000 study. *JAMA*. 1994;272(22):1749-1756.

TITLE Case-Finding for Depression in Elderly People: Balancing Ease of Administration With Validity in Varied Treatment Settings.

AUTHORS Blank K, Gruman C, Robison JT.

CITATION *J Gerontol.* 2004;59A(4):378-384.

QUESTION Compared with longer depression screening instruments, how well do the 2-question Primary Care Evaluation of Mental Disorders (PRIME-MD) and the Yale 1-question screens perform for detecting depression?

DESIGN Patients completed written screening instruments. Major depression diagnoses were made with a structured interview, blind to the PRIME-MD and Yale screening results but not to results from the longer screening instruments.

SETTING Two urban primary care practices, a general medical hospital, and 8 nursing homes in the United States; results reported here are limited to the primary care practices.

PATIENTS Three hundred sixty consecutive patients (125 from primary care) agreed to participate; 35 were ineligible because of psychosis, a positive screening result for alcoholism, or a positive screening result for dementia. An additional 12 patients declined to participate. Most were white; mean age was 77 ± 8.9 years.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

The PRIME-MD prompts, “During the past month have you often been bothered by feeling down, depressed, or hopeless?” and “During the past month have you often been bothered by little interest or pleasure in doing things?” A yes response to either question was considered a positive result. The Yale question is “Do you often feel sad or depressed?” Comparison instruments were the 20-item Center for Epidemiologic Studies Depression Scale¹ (CES-D; available at <http://www.chcr.brown.edu/pcoc/cesdscale.pdf>; accessed May 28, 2008) and the 30-item Geriatric Depression Scale² (GDS Long Version; available at <http://www.stanford.edu/~yesavage/GDS.html>; accessed May 28, 2008). An interviewer used the mood sections of the diagnostic interview schedule to establish major depression diagnoses.

MAIN OUTCOME MEASURES

Sensitivity, specificity, likelihood ratios, and diagnostic odds ratios (calculated from data provided in the article).

MAIN RESULTS

Fourteen (11%) of 125 patients were diagnosed with major depression (**Table 19-13**).

Table 19-13 Likelihood Ratios for the Longer Screening Instruments Compared With the Shorter Instruments^a

Test	Sensitivity	Specificity	LR+ (95% CI)	LR- (95% CI)	DOR (95% CI)
Longer Screening Instruments					
CES-D	0.79	0.75	3.1 (2.0-4.8)	0.29 (0.10-0.79)	11 (2.8-42)
GDS	0.79	0.67	2.4 (1.6-3.4)	0.32 (0.12-0.88)	7.3 (1.9-28)
Shorter Screening Instruments					
PRIME-MD	0.79	0.58	1.9 (1.3-2.6)	0.37 (0.13-1.0)	5.0 (1.3-19)
2 items					
Yale	0.64	0.64	1.8 (1.1-2.8)	0.56 (0.27-1.1)	3.2 (1.0-10)

Abbreviations: CES-D, Center for Epidemiologic Studies Depression Scale; CI, confidence interval; DOR, diagnostic odds ratio; GDS, Geriatric Depression Scale; LR+, positive likelihood ratio; LR-, negative likelihood ratio; PRIME-MD, Primary Care Evaluation of Mental Disorders.

^aThresholds for a positive screening result were CESD ≥ 16 and GDS ≥ 10 .

CONCLUSIONS

LEVEL OF EVIDENCE Level 2 for the PRIME-MD and Yale questionnaires; level 4 for the CES-D and GDS questionnaires.

STRENGTHS Few exclusion criteria; consecutive patients; high participation rate.

LIMITATIONS The sample size was small, as evidenced by the broad confidence intervals around the likelihood ratio. The criterion interviewers were unblinded to CES-D and GDS results, potentially biasing results toward improved performance compared with the brief screening instruments.

Results of this study should be interpreted with caution because of the small sample size and lack of blinding for the CES-D and GDS questionnaires. These 2 questionnaires, studied with a lower level of quality because of the lack of blinding, showed the best accuracy, as evidenced by the highest diagnostic odds ratio. However, there is no statistical difference in the diagnostic odds ratios between these questionnaires ($P = .57$).

The study gives useful information on test performance in older adults. Compared with past studies, the CES-D and GDS performed similarly. The PRIME-MD performed less well than in studies conducted in mixed-age populations. Brief 1- and 2-item questionnaires may perform less well in older (perhaps more medically ill) patients. This hypothesis needs testing in a larger study that allows for subgroup analyses of older and younger patients.

Reviewed by John W. Williams Jr, MD

REFERENCES FOR THE EVIDENCE

- Radloff LS. The Center for Epidemiological Studies Depression Scale (CES-D): a self-report depression scale for research in a general population. *Appl Psychol Meas.* 1977;1:385-401.
- Yesavage JA, Brink TL, Rose TL, et al. Development and validation of a geriatric depression screening scale: a preliminary report. *J Psychiatr Res.* 1982;17(1):37-49.

TITLE Identifying Depression in Primary Care: A Comparison of Different Methods in a Prospective Cohort Study.

AUTHORS Henkel V, Mergl R, Kohnen R, Maier W, Möller H, Hegerl U.

CITATION *BMJ*. 2003;326(7382):200-201.

QUESTION How well do 3 brief screening instruments perform for detecting depression in primary care?

DESIGN On a single day, all patients were asked to complete the 3 screening instruments. A telephone interviewer, blind to screening results, used the structured Composite International Diagnostic Interview to establish major depression diagnoses.

SETTING Eighteen primary care facilities in Germany.

PATIENTS A total of 487 patients gave informed consent. Of these, 431 completed all study assessments; 56 had incomplete data and were excluded. Demographic descriptors were not given.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

The Patient Health Questionnaire (PHQ) is a 9-item, depression-specific, self-administered instrument that asks about the criterion symptoms of major depression.¹ It is scored 0 to 27; a score of 10 or higher is considered a positive result. The General Health Questionnaire (GHQ) is a general measure of psychological well-being that exists in multiple versions; a 12-item version (GHQ-12) was used in this study and a score of 3 or higher was considered a positive result.² The World Health Organization–Five Well-Being Index (WHO-5) is a 5-item depression measure scored from 0 to 25; scores of 12 or lower are considered a positive result.³ The Composite International Diagnostic Interview was used to establish the diagnosis by the *Diagnostic and Statistical Manual of Mental Disorders*, (Fourth Edition) or *International Classification of Diseases and Related Health Problems*, Tenth Revision criteria.^{4,5}

MAIN OUTCOME MEASURES

Sensitivity and specificity. Likelihood ratios (LRs) were calculated by the reviewer.

MAIN RESULTS

Seventy-one (17%) of 431 patients were diagnosed with a depressive disorder, including 43 with major depression, 22 with dysthymia, 3 with bipolar depression, and 3 with a mood disorder caused by a general medical condition (Table 19-14).

Table 19-14 Test Performance

Test	Sensitivity	Specificity	LR+ (95% CI)	LR- (95% CI)	DOR (95% CI)
PHQ-9	0.78	0.85	5.2 (3.9-6.8)	0.26 (0.17-0.40)	20 (11-37)
WHO-5	0.93	0.64	2.6 (2.2-3.0)	0.11 (0.05-0.25)	24 (9.5-61)
GHQ-12	0.85	0.62	2.2 (1.9-2.6)	0.24 (0.14-0.42)	9.2 (4.7-18)

Abbreviations: CI, confidence interval; DOR, diagnostic odds ratio; GHQ-12, General Health Questionnaire-12; LR+, positive likelihood ratio; LR-, negative likelihood ratio; PHQ-9, Patient Health Questionnaire-9; WHO-5, World Health Organization-Five Well-Being Index.

CONCLUSIONS

LEVEL OF EVIDENCE Level 1.

STRENGTHS Consecutive patients were enrolled. There was a large sample size, and 3 instruments, all feasible for primary care settings, were compared.

LIMITATIONS The study population is not well described. Process of adapting instruments to German was not described. Instruments can perform differently in different languages.

The PHQ-9 had the best positive likelihood ratio (LR+) and a negative likelihood ratio (LR-), comparable to the GHQ-12, a nonspecific measure of psychological distress. The WHO-5 had the best LR- but only a modest LR+ compared with the PHQ-9. Overall, the PHQ-9 and WHO-5 had similar accuracy, as evidenced by their diagnostic odds ratio, but we can be more certain about the performance of the PHQ-9 because of its narrower confidence interval. Future studies might confirm that 2 are similar, on average. However, the goals of the clinician should drive whether to select one vs the other. Clinicians who want to maximize the chance that a patient who has a positive screening result actually has depression should choose the PHQ-9 (highest LR+). Clinicians who want to maximize the chance of correctly detecting patients without depression should choose the WHO-5 (lowest LR-).

The performance characteristics reported in this study may have been affected by translation into German. Another caveat is that almost one-third of depressed patients have dysthymia, a more chronic and milder depression, that tends to decrease the reported sensitivity compared to results of studies restricted to patients with major depression. Inclusion of dysthymic patients makes clinical sense because anti-depressant medications are effective for dysthymia.

Reviewed by John W. Williams Jr, MD

REFERENCES FOR THE EVIDENCE

- Spitzer RL, Kroenke K, Williams JB. Validation and utility of a self-report version of PRIME-MD: the PHQ primary care study: Primary Care Evaluation of Mental Disorders: Patient Health Questionnaire. *JAMA*. 1999;282(18):1737-1744.

2. Goldberg DG. *Manual of the General Health Questionnaire*. Windsor, Berksire, UK: NFER Publishing; 1978.
3. Psychiatric Research Unit, Frederiksberg General Hospital, Hillerød, Denmark. WHO-5 questionnaires. <http://www.who-5.org>. Accessed May 28, 2008.
4. American Psychiatric Publishing. *DSM-IV-TR. Diagnostic and Statistical Manual of Mental Disorders* (Fourth Edition, Text Revision). <http://www.psychiatryonline.com/content.aspx?aID=2016>. Accessed May 26, 2008.
5. World Health Organization. *International Classification of Disease*, Version 2007. <http://www.who.int/classifications/apps/icd/icd10online>. Accessed May 28, 2008.

TITLE The PHQ-9: Validity of a Brief Depression Severity Measure.

AUTHORS Kroenke K, Spitzer RL, Williams JMW.

CITATION *J Gen Intern Med*. 2001;16(9):606-613.

TITLE The Patient Health Questionnaire-2: Validity of a 2-item Depression Screener.

AUTHORS Kroenke K, Spitzer RL, Williams JMW.

CITATION *Med Care*. 2003;41(11):1284-1292.

QUESTION How well do the Patient Health Questionnaire (PHQ) 9- and 2-item versions perform for detecting depression in primary care?^{1,2}

DESIGN Adults (consecutive or every *n*th) making an outpatient visit were asked to complete the PHQ. A mental health clinician, blind to screening results, completed a structured telephone interview in a subset of patients (selected without regard to screening results) to establish depressive diagnoses.

SETTING Seven obstetrics-gynecology clinics, 5 general internal medicine clinics, and 3 family practices in the United States.

PATIENTS Three thousand primary care patients (mean age, 46 ± 17 years) and 3000 obstetrics-gynecology patients (mean age, 31 ± 11 years) participated. Patients who declined ($n = 435$) or who did not complete the questionnaire ($n = 1091$) were excluded. Of the 6000 participants, a diagnostic interview by a mental health clinician was completed in 580.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

The PHQ-9 is a 9-item, depression-specific, self-administered instrument that asks about the criterion symptoms of major depression. It is scored 0 to 27; a score of 10 or higher is considered a positive result. The 2-item screen, PHQ-2, uses the first 2 items of the PHQ-9 and asks about depressed mood and anhedonia. It is scored 0 to 6; a threshold of 3 or higher (set post hoc) was used in this study. The Structured Clinical Interview for *Diagnostic and Statistical*

Manual of Mental Disorders (Fourth Edition) was used as the criterion standard.

MAIN OUTCOME MEASURES

Sensitivity, specificity, and likelihood ratios.

MAIN RESULTS

One hundred three (18%) of 580 criterion standard patients scored 10 or higher on the PHQ-9; 88 (15%) scored 3 or higher on the PHQ-2. Forty-one (7.1%) were diagnosed with major depression; 65 (11%) were diagnosed with nonmajor depressive disorders (Table 19-15).

When diagnoses were broadened to include nonmajor depressive disorders, sensitivities decreased to 66% and 62% for the PHQ-9 and PHQ-2, respectively. Specificities were 93% and 95%, respectively.

CONCLUSIONS

LEVEL OF EVIDENCE Level 1 for the PHQ-9; level III for the PHQ-2.

STRENGTHS Large sample size.

LIMITATIONS Threshold set post hoc for the PHQ-2, which may overestimate performance characteristics. It is unclear how many of those chosen for a criterion standard evaluation completed the examination.

This report combines data from 2 studies conducted in primary care and obstetrics and gynecology clinics. The PHQ-2 (a revision of the Primary Care Evaluation of Mental Disorders [PRIME-MD] that uses 4 response categories instead of a yes/no format) performed well, but the threshold for a positive screen was set post hoc, which typically leads to overestimates of performance. If these results are replicated in other populations, the PHQ-2 could be endorsed as a feasible and accurate option for screening. The PHQ-9 performed well in these midlife adults with low rates of comorbid medical conditions. In contrast with the PHQ-2, the 9-item version can be used to monitor treatment response and gives more information for specific depressive diagnoses.³ Some experts recommend using the first 2 PHQ

Table 19-15 Likelihood Ratios for Patient Heath Questionnaire (PHQ)-9 and PHQ-2 for Depression

Test	Sensitivity	Specificity	LR+ (95% CI)	LR- (95% CI)	DOR (95% CI)
PHQ-2	0.83	0.90	8.3 (6.2-11)	0.19 (0.10-0.37)	44 (18-103)
PHQ-9	0.88	0.88	7.3 (5.6-9.4)	0.14 (0.06-0.32)	52 (20-139)

Abbreviations: CI, confidence interval; DOR, diagnostic odds ratio; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

items (using a yes/no response format) and then administering the complete PHQ-9 to those who endorse a yes response to either of the first 2 items.

Reviewed by John W. Williams Jr, MD

REFERENCES FOR THE EVIDENCE

- Spitzer RL, Williams JBW, Kroenke K, Hornyak R, McMurray J, for the Patient Health Questionnaire Obstetrics-Gynecology Study Group. Validity and utility of the PRIME-MD Patient Health Questionnaire in assessment of 3000 obstetric-gynecologic patients: the PRIME-MD Patient Health Questionnaire Obstetrics-Gynecology Study. *Am J Obstet Gynecol.* 2000;183(3):759-769.
- Spitzer RL, Kroenke K, Williams JBW. Validation and utility of a self-report version of PRIME-MD: the PHQ primary care study. *JAMA.* 2000;282(18):1737-1744.
- Lowe B, Kroenke K, Herzog W, Grafe K. Measuring depression outcome with a brief self-report instrument: sensitivity to change of the Patient Health Questionnaire (PHQ-9). *J Affect Disord.* 2004;81(1):61-66.

TITLE Comparative Validity of 3 Screening Questionnaires for *DSM-IV* Depressive Disorders and Physicians' Diagnoses.

AUTHORS Löwe B, Spitzer RL, Gräfe K, et al.

CITATION *J Affect Disord.* 2004;78(2):131-140.

QUESTION How well do 3 brief screening instruments perform for detecting depression in primary care?

DESIGN On selected days, all adult patients were asked to complete the 3 screening instruments. An interviewer, blind to screening results, completed diagnostic interviews in a subset of patients (selected without regard to screening results) to establish depressive diagnoses.

SETTING Twelve family practices and the outpatient clinics of an academic medical center in Germany.

PATIENTS A total of 1619 consecutive patients (mean age, 42 ± 14 years) participated; 431 declined or were unable to complete the questionnaires. The diagnostic interview was completed in 528 (88%) of the 600 selected for diagnostic confirmation; 27 were excluded because of missing data, leaving a final sample of 501.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

The Patient Health Questionnaire-9 (PHQ-9) is a 9-item, depression-specific, self-administered instrument that asks about the criterion symptoms of major depression. It is scored 0 to 27; a score of 10 or higher is the usual threshold but a score of 11 or higher was used in this study. The Hospital Anxiety and Depression Scale (HADS) has 7-item depression and anxiety subscales.¹ It is scored 0 to 21; a score of 8 or higher is considered a positive result. The World Health Organization-5

Well-Being Scale (WHO-5) is a 5-item depression measure scored from 0 to 25; scores of 12 or lower are considered a positive result.² The Structured Clinical Interview for *DSM-IV* was used as the criterion standard.

MAIN OUTCOME MEASURES

Sensitivity and specificity. The reviewer calculated likelihood ratios.

MAIN RESULTS

Sixty-six (13%) of 501 patients were diagnosed with major depression; 126 (25%) were diagnosed as having any depressive disorder that included adjustment disorder and dysthymia (**Table 19-16**).

When diagnoses were broadened to include any depressive disorder, sensitivities decreased to 81%, 81%, and 94% for the HADS, PHQ (threshold ≥ 10), and WHO-5, respectively. Specificities were 75%, 82%, and 60%, respectively.

CONCLUSIONS

LEVEL OF EVIDENCE Level 1.

STRENGTHS Large sample size; community and academic settings; criterion standard completed for a high proportion of individuals chosen for evaluation.

LIMITATIONS Results for the PHQ-9 and WHO-5 were reported for atypical thresholds and may have been determined post hoc. The author supplied supplemental data for the usual WHO-5 threshold, but the higher PHQ-9 threshold creates an expectation of lower sensitivity. Despite an expected lower sensitivity, the sensitivity was high.

A unique contribution is the comparison of a brief depression-specific instrument designed for primary care populations (the PHQ-9 and WHO-5) to a brief depression-specific instrument designed for medically ill patients (the HADS). Both the PHQ-9 and WHO-5 were excellent in ruling out

Table 19-16 Likelihood Ratios of Screening Instruments for Major Depression

Test	Sensitivity	Specificity	LR+ (95% CI)	LR- (95% CI)	DOR (95% CI)
PHQ-9	0.98	0.80	4.9 (4.1-6.0)	0.02 (0-0.14)	193 (27-1408)
HADS	0.88	0.69	2.8 (2.4-3.3)	0.18 (0.09-0.34)	16 (7.5-35)
WHO-5 ^a	1.00	0.54	2.2 (2.0-2.4)	0 (0-0.22) (2.0-2.4)	156 (9.6-2540)

Abbreviations: CI, confidence interval; DOR, diagnostic odds ratio; HADS, Hospital Anxiety and Depression Scale; LR+, positive likelihood ratio; LR-, negative likelihood ratio; PHQ-9, Patient Health Questionnaire-9; WHO-5, World Health Organization-5 Well-Being Scale.

^aData for the WHO-5, with a threshold of 12 or lower, were supplied by the author.

depression, but the WHO-5 had a modest positive likelihood ratio.

One caution when interpreting these results for English-speaking populations is that the instruments were given in German. Even when careful cultural adaptations are made, questionnaires may perform differently across language groups and cultures.

Reviewed by John W. Williams Jr, MD

REFERENCES FOR THE EVIDENCE

1. Zigmond AS, Snaith RP. The Hospital Anxiety and Depression Scale. *Acta Psychiatr Scand.* 1983;67(6):361-370.
2. Psychiatric Research Unit, Frederiksberg General Hospital, Hillerød, Denmark. WHO-5 questionnaires. <http://www.who-5.org>. Accessed May 28, 2008.

CHAPTER 20

Does This Patient Have a Family History of Cancer?

Harvey J. Murff, MD, MPH

David R. Spigel, MD

Sapna Syngal, MD, MPH

CLINICAL SCENARIO

A 35-year-old woman presents for an initial visit and during the medical interview mentions that her mother and grandmother had breast cancer. She reports that her mother was diagnosed with cancer at age 42 years, and she believes that her grandmother, on her mother's side, was diagnosed in her 30s. Because of her family history, she is concerned about her risk of developing breast cancer. Despite having no symptoms of breast cancer, she wonders at what age she should start having mammograms and whether she should have genetic testing.

WHY IS IT IMPORTANT TO RECORD AN ACCURATE FAMILY HISTORY OF CANCER?

Individuals with a family history for certain kinds of cancers can have an increased risk of developing cancer themselves.^{1,2} Two meta-analyses found relative risks of 2.1 (95% confidence interval [CI], 2.0-2.2) for breast cancer³ and 3.1 (95% CI, 2.6-3.7) for ovarian cancer⁴ in women with affected first-degree relatives. Similar higher risks have been observed for endometrial cancer,⁵⁻⁷ colon cancer,⁷⁻¹⁰ and prostate cancer.¹¹⁻¹³

Accurate reporting of family history helps risk-stratify patients, which in turn determines screening and prevention interventions. Individuals with family histories that are suggestive of a hereditary cancer syndrome (Box 20-1)¹⁴⁻¹⁷ are typically considered at high or very high risk of developing cancer. Individuals with family histories for cancers not recognized as hereditary are generally at a moderately increased risk of developing cancer compared with the general population. Several organizations have recommended initiating screening earlier, more frequently, or both in patients at moderately increased risk of developing cancer according to their family history.¹⁸⁻²⁷ Guidelines have also been published regarding the management of individuals who are at high cancer risk.^{26,28-30} A family history of malignancy can not only influence cancer screening initiation and frequency but also affect treatment strategies. Family histories affect decisions about cancer chemoprevention,^{31,32} and those individuals identified as being at very high risk may also be considered for risk-reducing surgeries.³³ Likewise, algorithms that predict individuals who might be candidates for genetic testing rely almost exclusively on family history information.^{14,16}

Box 20-1 Examples of Clinical Diagnostic Criteria for 2 Familial Cancer Syndromes and Recommendations Regarding Genetic Testing for Cancer Susceptibility

HEREDITARY NONPOLYPOsis COLON CANCER (HNPCC)^{14,15}

All of the following criteria should be present:

At least 3 relatives must have a cancer associated with HNPCC (colon, endometrial, ovarian, stomach, small bowel, hepatobiliary, ureter, renal-pelvis, brain).

One should be a first-degree relative of the other 2.

At least 2 successive generations should be affected.

At least 1 of the relatives with cancer associated with HNPCC should have received the diagnosis before aged 50 years.

HEREDITARY BREAST/OVARIAN CANCER^{16a}

Any of the following criteria should be present:

Two breast cancers in a first- or second-degree relative and mean age at diagnosis of 40 years.

One breast cancer and 1 ovarian cancer in a first- or second-degree relative and mean age at diagnosis of 41 to 50 years^a.

Two or more breast cancers and 1 ovarian cancer in a first- or second-degree relative.

Ovarian cancer in 2 relatives.

AMERICAN SOCIETY OF CLINICAL ONCOLOGISTS POLICY STATEMENT ON GENETIC TESTING¹⁷

Indications for genetic testing:

The individual has personal or family history features suggestive of a genetic cancer susceptibility condition.

The test can be adequately interpreted.

The results will aid in diagnosis or influence the medical or surgical treatment of the patient or family members at hereditary risk of cancer.

^aIdentified relatives must be on the same side of the family (either maternal or paternal relatives).

Because many screening and prevention strategies for cancer rely on self-reported family history information, inaccurate information could result in inappropriate care. A false-negative family cancer history results in an underestimation of cancer risk and missed opportunities for cancer screening. Failure to collect adequate family history information and appropriately manage a patient's cancer risk may result in substandard care and in some cases has resulted in malpractice litigation.³⁴

Conversely, a patient's false belief in a positive family cancer history can cause stress³⁵ and, when compounded by the physician's overestimation of risk, may lead to unnecessary procedures or surgeries.³⁶ The overestimation of cancer risk based on pedigree data creates unneeded referrals for genetic testing or cancer risk counseling.^{37,38} The increased availability and demand for genetic services require an even more

important role for primary care physicians in recording an accurate family cancer history^{39,40}; however, many physicians lack adequate training in genetics to accurately identify and refer appropriate candidates for genetic services.^{41,42} In addition, with direct-to-consumer advertising for genetic testing now a reality,⁴³ accurate family history collection and cancer risk assessment by primary care physicians might help decrease the likelihood of inappropriate referrals for genetic counseling and testing.

Few data exist describing how often inaccurate risk assessments are made according to faulty pedigree data. In one retrospective study³⁵ that examined patients referred to 2 cancer genetic clinics, patient treatment was changed in 23 (11%) of 213 patients after their previously reported family history information was found to be inaccurate. In 15 of these patients, screening was thought to be unnecessary, although in 8 patients cancer risk was determined to be greater than initially believed. Further studies have supported these findings, with one study⁴⁴ determining that 6 (5%) of 120 patients referred to a cancer clinic had changes in treatment after confirmation of the family cancer history revealed discrepancies. In most of these patients, the cancer risk had been overestimated.

Prevalence of a Positive Family History of Specific Familial Cancers

The prevalence of a family history of cancer varies, depending on the cancer type. The prevalence of a family history of breast cancer has been estimated to range from 5% to 22%⁴⁵⁻⁴⁸; colon cancer, 2.0% to 9.4%^{8,45,46}; ovarian cancer, 1.1% to 3.5%^{45,46,48}; endometrial cancer, 0.5% to 1.4%^{45,46}; and prostate cancer, 4.6% to 9.5%.^{11,13,46,49} Most of this variation is based on methodology and study population. Some studies included distant relatives in the definition of a positive family cancer history, whereas other studies have focused only on first-degree relatives. Variability in rates also occurs when the results are derived from the general population as opposed to patients referred to cancer or genetic centers, which have higher prevalence rates.

How to Elicit a Family Cancer History

Family medical history information is important for risk assessment in numerous chronic medical conditions in addition to cancer, such as diabetes mellitus and cardiovascular disease; therefore, eliciting a family cancer history can serve as a model for collecting family history information for other disorders. Typically, family history information is collected directly from the patient or from screening questionnaires filled out by the patient. Alternatively, the patient's parent or another family member may provide the information.

Screening questionnaires are often either a list of relatives, with space to provide information on overall health, age, and cause of death, or a list of adult-onset diseases with space to list the affected relatives. Disease history should be collected on first-degree relatives (mother, father, sisters, brothers, and children) and second-degree relatives (maternal and paternal grandparents, aunts, uncles, nieces, and nephews). It is

important to inquire about various types of cancers because certain hereditary cancer syndromes can be identified by specific cancers that cluster within families (Box 20-1), such as endometrial with colon⁵⁰ and breast with ovarian.^{51,52} If the initial screening interview or questionnaire reveals a potential familial predisposition to a particular disease, the family history should be expanded.

Although establishing the numbers of both affected and unaffected relatives is important for determining penetrance and predicting the likelihood of gene mutations, this information for primary care physicians would seldom influence cancer screening decisions. For affected relatives, documenting the age at cancer diagnosis is important because patients developing cancer at ages significantly earlier than typically expected increases the possibility of a hereditary cancer syndrome. Inaccurate reporting of ages at diagnosis for breast cancer can have a considerable influence on risk prediction in families with fewer than 4 affected relatives.⁵³ A 3-generation pedigree, displayed graphically in Figure 20-1, offers a convenient symbolic method of summarizing information. Because of previous inconsistencies in pedigree symbol usage, the Pedigree Standardization Task Force, organized through the National Society of Genetic Counselors, has proposed recommendations for a standardized pedigree nomenclature.⁵⁴

When recording a pedigree, particularly for breast and gynecologic cancers, it is important to inquire about disease in both maternal and paternal lineages because mutations can be transmitted through either parent. When collecting information on second-degree relatives, it is important to note the lineage to which the relative belongs (such as paternal vs maternal grandparents) because the degree of risk might vary if affected relatives do not belong to the same lineage. A brief reference for physicians on the family medical history has been prepared by the American Medical Association (<http://www.ama-assn.org/ama/pub/category/2380.html>; accessed May 29, 2008). Many other electronic sources and texts⁵⁵ are available.

Family medical history information can be collected during a patient care visit or outside of the clinical encounter. Methods of collecting family history information outside of the clinical encounter can include paper questionnaires,⁵⁶ computer questionnaires in kiosks within a clinic waiting area,³⁷ Web-based electronic collection, and interviews by health care professionals. The optimal means of collection has not been determined.

METHODS

Two of the authors (H.J.M. and D.R.S.) performed independent searches of the MEDLINE database for English-language articles dated 1966 to June 2004 from the PubMed search engine. The following Medical Subject Headings were used: “family,” “genetic predisposition to disease,” “medical history taking,” “neoplasm,” and “reproducibility of results.” We also searched using the following textwords: “accuracy,” “sensitivity,” “specificity,” and “family history,” combined with the conditions “breast cancer,” “colon cancer,” “ovarian cancer,” “prostate cancer,” “endometrial cancer,” or “uterine

cancer.” We specifically included cancers that were likely to be commonly encountered by primary care physicians and whose management might be altered according to family history information. The reviewers evaluated article abstracts and chose studies for full-text review according to the abstract. We searched the bibliographies of all retrieved articles to identify additional sources.

Articles were included if they were original articles describing the accuracy of the site-specific family history for the pre-specified cancers and contained a criterion standard. Studies presenting aggregate data (all cancer types combined into a single measure) for self-reported family cancer history information were excluded. For purposes of this study, the criterion standard for a positive family history of cancer required verification from the identified relative’s medical record, physician, or death certificate or verification within a population cancer registry. For studies to be included in our analysis, verification of a negative family history for cancer had to have been performed. Thus, if a study participant reported that a relative had no history of breast cancer, the relative’s medical records, death certificate if applicable, or local cancer registry was examined for verification of this report.

The completeness of case findings within tumor registries varied, with 83% to 99% of cancers identified through medical

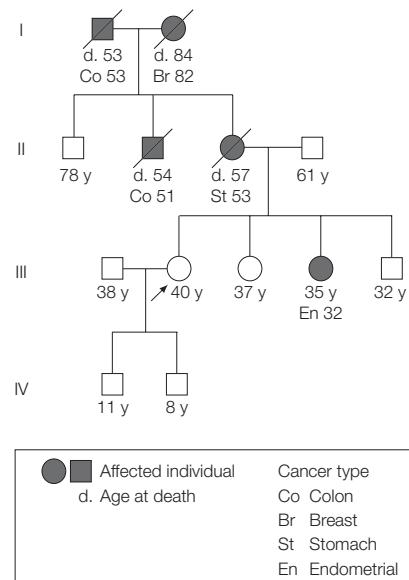


Figure 20-1 Hypothetical Pedigree for a Consultand With a Family History Suggestive of a Hereditary Nonpolyposis Colon Cancer

A consultand is an individual under evaluation for predicting his or her own future risk or the risk of his or her offspring. The arrow identifies the consultand. The 2-letter abbreviation with number (eg, Co 53) represents the diagnosis an individual received, followed by the age at the diagnosis. This pedigree meets Amsterdam II criteria,¹⁴ which include 3 relatives with a hereditary nonpolyposis colorectal cancer-associated tumor, such as colon cancer, endometrial cancer, ureteral cancer, cancer of the renal pelvis, ovarian cancer, stomach cancer, or small bowel cancer; 1 relative who is a first-degree relative of the other 2; cancers affecting at least 2 generations; and 1 or more cases diagnosed before aged 50 years.

record reviews and patient interviews also being present within the registry.⁵⁷⁻⁶¹ Specific cancer sites are correctly recorded within the registry in 93% to 97% of cases. Forty-nine percent of discrepancies within tumor registries result from changes in an initial diagnosis with a failure to update registry information.⁶¹ For breast cancer, the sensitivity and specificity of tumor registries are high.⁶² Other tumors listed within registries have similar high sensitivities.^{63,64} The National Program of Cancer Registries of the Centers for Disease Control and Prevention has created a system that provides the rationale for accepting these data in studies that attempt validation of the patient's family cancer history.⁶⁵ Although death certificates probably lack the accuracy of tumor registries, the poorer performance of death certificates is more likely attributed to poor sensitivity (the death certificates do not record the information when in fact the decedent had cancer).⁶⁶ According to autopsy studies, the death certificate is estimated to have a sensitivity of 87% for identifying cancer.⁶⁷

We identified 22 studies from our search, using the listed criteria.^{44,68-88} Of these, only 7 provided information on both the test characteristics of a positive and negative report of a family cancer history.^{71,73,74,76,78-80} One study specifically assessed pedigrees suggestive of hereditary nonpolyposis colorectal cancer (HNPCC)⁷⁶ and was included within the analysis. Sensitivity and specificity were determined for the family history interview for HNPCC, but this information was not combined with other colon cancer studies. We used techniques described from previous Rational Clinical Examination articles to determine study quality, and all 7 evaluated studies were assigned a quality score of C,⁸⁹ which reflects a study with an independent blind comparison of sign or symptom and a criterion standard of diagnosis among nonconsecutive patients suspected of having the target condition.

Because the population studied could influence reporting accuracy, test characteristics were calculated separately for individuals with a personal history of cancer, as well as individuals without a personal history of cancer. Sensitivity and specificity of patient self-report of a family history of cancer and likelihood ratios (LRs) of a positive or negative report were calculated according to raw data supplied by the original articles that met our search criteria. CIs for LRs were computed with previously described methods.⁹⁰ We used random-effects summary measures for combining the data because this provided broader CIs that display the uncertainty around the point estimates. The summary measures described this uncertainty better than the simple range of possible data from the original studies. For colon cancer, one study⁷⁶ was not included within the summary LRs because it specifically evaluated the family history for HNPCC rather than colon cancer in general.

RESULTS

Precision

Precision reflects the reproducibility of a measurement. Assessing the precision of the family history interview is difficult because it can be influenced by both patient and physician

factors. Although we were unable to identify any studies assessing the reliability of the physician's family cancer history assessment, one study⁹¹ in breast cancer examined the reliability of patient self-report. In this nested case-control study,⁹¹ comparisons were made between self-reported family history information in women before the development of the disease and after the development of the disease. Follow-up surveys were completed 2 years after the initial survey. Women who had developed breast cancer, as well as those who had not developed breast cancer, were surveyed. The agreement for maternal history of breast cancer was $\kappa = 0.92$ and $\kappa = 1.0$ for cases and controls, respectively; and for a history of breast cancer in a sister, $\kappa = 0.65$ and $\kappa = 0.88$, respectively. Although the study did not assess whether a real change in family history might have occurred during the study period, these results suggest that self-reported family breast cancer history is probably only slightly influenced by recall bias. Patient precision regarding the family history interview for other cancers has not been reported.

Accuracy

Accuracy represents how well a particular test measures the value it is intending to measure. Seven studies concerning family cancer history were ultimately included in this analysis (Table 20-1).^{71,73,74,76,78-80} Three studies collected family history from personal patient interviews,^{73,74,79} whereas the other 4 relied on a self-completed survey.^{71,76,78,80} Four studies solely relied on cancer registry data as their criterion standard^{73,74,79,80}; 2 studies used a combination of medical records and death certificates,^{71,78} whereas the remaining study used all 3 sources as its criterion standard.⁷⁶ Only information for first-degree relatives was extracted.

For individuals with cancer (Table 20-2), the positive likelihood ratio (LR+) and negative likelihood ratio (LR-) of a self-reported family cancer history in a first-degree relative were 23 (95% CI, 8.1-64) and 0.29 (95% CI, 0.13-0.67) for colon, 41 (95% CI, 23-75) and 0.07 (95% CI, 0.03-0.13) for breast, 20 (95% CI, 4.3-89) and 0.55 (95% CI, 0.35-0.86) for endometrial, 44 (95% CI, 15-132) and 0.21 (95% CI, 0.12-0.37) for ovarian, and 24 (95% CI, 2.3-262) and 0.25 (95% CI, 0.16-0.39) for prostate cancers, respectively. For patients without a personal history of cancer (Table 20-3), the LR+ and LR- of a family history for the following cancers in a first-degree relative were 23 (95% CI, 6.4-81) and 0.25 (95% CI, 0.10-0.63) for colon, 8.9 (95% CI, 5.4-15) and 0.20 (95% CI, 0.08-0.49) for breast, 14 (95% CI, 2.2-83) and 0.68 (95% CI, 0.31-1.5) for endometrial, 34 (95% CI, 5.7-202) and 0.51 (95% CI, 0.13-2.1) for ovarian, and 12 (95% CI, 6.5-24) and 0.32 (95% CI, 0.18-0.55) for prostate cancers, respectively. The estimates for sensitivity, specificity, and LRs for unaffected individuals for prostate, breast, endometrial, and ovarian cancers are based on data from a single study by Kerber and Slattery.⁷⁴

Of the remaining 15 studies, we excluded 8 studies⁸¹⁻⁸⁸ because the tumor data were presented in aggregate (ie, family history of any cancer) or were unclear; therefore, we were unable to extrapolate site-specific numbers. Seven studies evaluated only the positive predictive value of self-reported family history information for breast, colon, ovarian,

Table 20-1 Characteristics of Included Studies of Patient Report of a Family History of Cancer in a First-Degree Relative^a

Source, y	Cancer Site		Method of Family History Information Collection	Criterion Standard
	Affected Individuals	Unaffected Individuals		
Love et al, ⁶⁸ 1985	Colon, breast		Personal interview	Medical records, death certificate
Breuer et al, ⁶⁹ 1993	Breast		Self-completed survey	Medical records
Theis et al, ⁷⁰ 1994	Colon, prostate, breast, ovarian		Personal interview and self-completed survey	Medical records, death certificate, cancer registry
Aitken et al, ⁷¹ 1995	Colon		Self-completed survey	Medical records, death certificate
Parent et al, ⁷² 1995	Breast	Breast	Personal interview	Medical records
Anton-Culver et al, ⁷³ 1996	Breast		Personal interview	Cancer registry
Kerber and Slattery, ⁷⁴ 1997	Colon, prostate, breast, endometrial, ovarian	Colon, prostate, breast, endometrial, ovarian	Personal interview	Cancer registry
Sijmons et al, ⁴⁴ 2000	Colon, breast, ovarian		Personal interview and self-completed survey	Medical records, death certificate
Erola et al, ⁷⁵ 2000	Breast		Personal interview and self-completed survey	Medical records, cancer registry
Katballe et al, ⁷⁶ 2001	Colon		Self-completed survey	Medical records, death certificate, cancer registry
King et al, ⁷⁷ 2002	Colon, prostate, breast, endometrial, ovarian		Personal interview	Medical records, death certificate
Ziegas and Anton-Culver, ⁷⁸ 2003	Colon, prostate, breast, endometrial, ovarian		Self-completed survey	Medical records, death certificate
Mitchell et al, ⁷⁹ 2004	Colon	Colon	Personal interview	Cancer registry
Verkooijen et al, ⁸⁰ 2004	Breast, ovarian		Self-completed survey	Cancer registry

^aAffected individuals are patients who have a personal diagnosis of cancer and unaffected individuals are patients with no personal diagnosis of cancer.

Table 20-2 Studies Evaluating Both Sensitivity and Specificity of Patient Report of a Family History of Cancer in a First-Degree Relative in Individuals With Cancer

Source, y	Cancer Type	No. of Patients/Total (%)		LR (95% CI)	
		Sensitivity	Specificity	Positive	Negative
Kerber and Slattery, ⁷⁴ 1997	Colon	11/17 (65)	98/108 (91)	6.9 (3.5-13)	0.39 (0.20-0.74)
Katballe et al, ⁷⁶ 2001 ^a	Colon	11/18 (61)	66/69 (96)	14 (4.4-45)	0.41 (0.23-0.73)
Ziegas and Anton-Culver, ⁷⁸ 2003	Colon	174/194 (90)	1454/1498 (97)	31 (23-41)	0.11 (0.07-0.16)
Mitchell et al, ⁷⁹ 2004	Colon	30/53 (57)	1256/1269 (99)	55 (31-100)	0.44 (0.32-0.60)
Summary ^b				23 (8.1-64)	0.29 (0.13-0.67)
Kerber and Slattery, ⁷⁴ 1997	Prostate	11/16 (69)	101/109 (93)	9.4 (4.5-20)	0.34 (0.16-0.70)
Ziegas and Anton-Culver, ⁷⁸ 2003	Prostate	46/58 (79)	557/564 (99)	64 (30-135)	0.21 (0.13-0.35)
Summary				24 (2.3-262)	0.25 (0.16-0.39)
Anton-Culver et al, ⁷³ 1996	Breast	54/59 (92)	364/370 (98)	56 (25-125)	0.09 (0.04-0.20)
Kerber and Slattery, ⁷⁴ 1997	Breast	11/13 (85)	107/112 (96)	19 (7.8-46)	0.16 (0.05-0.58)
Verkooijen et al, ⁸⁰ 2004	Breast	60/61 (98)	247/249 (99)	122 (31-487)	0.02 (0-0.12)
Ziegas and Anton-Culver, ⁷⁸ 2003	Breast	188/197 (95)	850/873 (97)	36 (24-54)	0.05 (0.03-0.09)
Summary				41 (23-75)	0.07 (0.03-0.13)
Kerber and Slattery, ⁷⁴ 1997	Endometrial	2/7 (29)	114/118 (97)	8.4 (1.9-38)	0.74 (0.46-1.2)
Ziegas and Anton-Culver, ⁷⁸ 2003	Endometrial	10/18 (56)	1035/1052 (98)	34 (18-64)	0.45 (0.27-0.76)
Summary				20 (4.3-89)	0.55 (0.35-0.86)
Kerber and Slattery, ⁷⁴ 1997	Ovarian	2/3 (67)	117/122 (96)	16 (5.0-53)	0.35 (0.07-1.7)
Verkooijen et al, ⁸⁰ 2004	Ovarian	4/6 (67)	168/170 (99)	57 (13-251)	0.34 (0.11-1.0)
Ziegas and Anton-Culver, ⁷⁸ 2003	Ovarian	35/42 (83)	1017/1028 (99)	78 (43-142)	0.17 (0.09-0.33)
Summary				44 (15-132)	0.21 (0.12-0.37)

Abbreviations: CI, confidence interval; LR, likelihood ratio.

^aSensitivity and specificity of a patient having a high-risk colon cancer pedigree according to Amsterdam II criteria.¹⁴

^bComposite does not include data from Katballe et al⁷⁶ (see "Methods").

Table 20-3 Studies Evaluating Both Sensitivity and Specificity of Patient Report of a Family History of Cancer in a First-Degree Relative in Healthy Individuals

Source, y	Cancer Type	No. of Patients/Total (%)		LR (95% CI)	
		Sensitivity	Specificity	Positive	Negative
Aitken et al, ⁷¹ 1995	Colon	70/81 (86)	219/239 (92)	10 (6.7-16)	0.15 (0.09-0.26)
Kerber and Slattery, ⁷⁴ 1997	Colon	13/16 (81)	178/190 (94)	13 (7.1-23)	0.20 (0.07-0.56)
Mitchell et al, ⁷⁹ 2004	Colon	9/17 (53)	1015/1020 (99)	108 (40-288)	0.47 (0.29-0.78)
Summary				23 (6.4-81)	0.25 (0.10-0.63)
Kerber and Slattery, ⁷⁴ 1997	Prostate	21/30 (70)	166/176 (94)	12 (6.5-24)	0.32 (0.18-0.55)
Kerber and Slattery, ⁷⁴ 1997	Breast	18/22 (82)	167/184 (91)	8.9 (5.4-15)	0.20 (0.08-0.49)
Kerber and Slattery, ⁷⁴ 1997	Endometrial	1/3 (33)	198/203 (98)	14 (2.2-83)	0.68 (0.31-1.5)
Kerber and Slattery, ⁷⁴ 1997	Ovarian	1/2 (50)	201/204 (99)	34 (5.7-202)	0.51 (0.13-2.1)

Abbreviations: CI, confidence interval; LR, likelihood ratio.

Box 20-2 Selected Web Sites for Cancer Risk Calculators^a

METHODS FOR ESTIMATING CANCER RISK

Various cancer sites:

<http://www.yourdiseaserisk.wustl.edu/>

Breast Cancer Risk Assessment tools:

<http://www.cancer.gov/bcrisktool/>; <http://www.halls.md/breast/risk.htm>

METHODS FOR ESTIMATING THE LIKELIHOOD OF A BRCA MUTATION

BRCA PRO statistical model:

<http://astor.som.jhmi.edu/BayesMendel/>

Mutation prevalence tables:

<http://www.myriadtests.com/provider/brca-mutation-prevalence.htm>

^aAccessed May 29, 2008.

prostate, and endometrial cancers (Table 20-4). Positive predictive values tended to be better in articles concerning first-degree relatives compared with second-degree relatives. Individuals with personal histories of cancer tended to report family histories with a greater positive predictive value, although the number of studies evaluating unaffected individuals was limited.

Common Reasons for False-Positive or False-Negative Reports

In cancers in which patients are likely to be accurate in their report, such as breast cancer, case reports have indicated that false-positive reports are associated with malingering, problems with patient-physician communication, or history of benign breast disease being reported as malignant.³⁶ Other common reasons for false-positive reports of family cancer history result from confusion based on primary vs metastatic disease.^{58,92} This confusion has been described with false reports of primary liver cancer, as well as central nervous system

cancers. Cancers that are frequently overreported include melanoma, which is incorrectly reported in almost half the reports,⁹³ and noncolonic gastrointestinal malignancies.³⁵

Several factors relate to a false-negative report of a family history of cancer. In one study,⁹⁴ older patients and nonwhite respondents were more likely to underreport a family history of cancer. Another study⁷⁴ demonstrated that older patients were more likely to falsely report a negative family history of cancer, whereas patient sex and education level have little effect on the accuracy of reporting. Specific cancers with high rates of false-negative reporting include central nervous system tumors and hematologic malignancies.⁹⁴

Other Means for Collecting Family History Information and Ways to Improve Family History Data Collection

Several barriers exist for the collection of family history information. Patient-specific factors that might result in poor pedigree collection include poor family communication, family myths, or individual spiritual beliefs. For physicians, probably the most significant barrier is time. Although a comprehensive family history assessment can take 15 to 30 minutes,⁹⁵ the average primary care visit lasts only 16 minutes.⁹⁶ Several alternative methods that involve collecting this information outside the context of the clinical visit may facilitate the collection of family history information. These other methods include self-completed patient paper surveys, computer-based tools, and personal visits arranged solely for pedigree collection.

Family history questionnaires offered outside of a clinical visit confer several theoretic advantages to visit-based pedigree assessment.⁹⁷ Besides saving clinic time, patients can consult with family members to check the accuracy of the information, which can then be reviewed and integrated into a clinic appointment when relevant. The data from a questionnaire developed in Switzerland compared with information found within 2 population-based cancer registries exhibited sensitivities of 74% and 85% and specificities of 97%.⁵⁶ Family history assessment tools (Box 20-2) have also been developed to assist physicians in determining which individuals might be candidates for genetic testing.⁹⁸

Table 20-4 Predictive Value of a Positive Report of a Family History of Cancer in a First-Degree or Second-Degree Relative

Source, y	Cancer Type	Positive Predictive Value, No. of Patients/Total (%)			
		Cancer Cases		Healthy Controls	
		First-Degree Relative	Second-Degree Relative	First-Degree Relative	Second-Degree Relative
Kerber and Slattery, ⁷⁴ 1997	Colon	11/21 (52)		13/25 (52)	
Mitchell et al, ⁷⁹ 2004	Colon	33/43 (70)	13/22 (62)	9/14 (63)	10/16 (63)
King et al, ⁷⁷ 2002	Colon	22/24 (92)			
Love et al, ⁶⁸ 1985	Colon	39/42 (93)	31/37 (84)		
Sijmons et al, ⁴⁴ 2000	Colon	30/33 (91)	15/15 (100)		
Theis et al, ⁷⁰ 1994	Colon	13/14 (93)	21/29 (72)		
Ziegas and Anton-Culver, ⁷⁸ 2003	Colon	174/218 (80)	52/70 (74)		
Aitken et al, ⁷¹ 1995	Colon			70/90 (78)	
Summary ^a		81 (77-85)	77 (70-83)	71 (63-78)	63 (39-82)
Kerber and Slattery, ⁷⁴ 1997	Prostate	11/19 (58)		21/31 (68)	
Ziegas and Anton-Culver, ⁷⁸ 2003	Prostate	46/53 (87)	30/40 (75)		
Theis et al, ⁷⁰ 1994	Prostate	11/13 (85)	11/11 (100)		
King et al, ⁷⁷ 2002	Prostate	25/29 (86)			
Summary ^a		85 (78-90)	80 (67-89)	68 (50-82)	
Kerber and Slattery, ⁷⁴ 1997	Breast	11/16 (69)		18/35 (51)	
Parent et al, ⁷² 1995	Breast	67/74 (91)		33/34 (97)	
Ziegas and Anton-Culver, ⁷⁸ 2003	Breast	188/211 (89)	103/115 (90)		
Eerola et al, ⁷⁵ 2000	Breast	94/99 (95)	109/114 (96)		
Sijmons et al, ⁴⁴ 2000	Breast	65/69 (94)	28/31 (90)		
Theis et al, ⁷⁰ 1994	Breast	166/167 (99)	33/39 (85)		
Love et al, ⁶⁸ 1985	Breast	78/83 (94)	65/74 (88)		
Anton-Culver et al, ⁷³ 1996	Breast	54/60 (90)			
Verkooijen et al, ⁸⁰ 2004	Breast	60/62 (97)			
King et al, ⁷⁷ 2002	Breast	38/40 (95)			
Breuer et al, ⁶⁹ 1993	Breast	84/94 (89)			
Summary ^a		93 (91-94)	91 (88-94)	74 (63-83)	
Kerber and Slattery, ⁷⁴ 1997	Endometrial	2/6 (33)		1/6 (17)	
Ziegas and Anton-Culver, ⁷⁸ 2003	Endometrial	10/27 (37)	3/14 (21)		
King et al, ⁷⁷ 2002	Endometrial	2/5 (40)			
Summary ^a		37 (24-53)	21 (7-47)	17 (3-57)	
Kerber and Slattery, ⁷⁴ 1997	Ovarian	2/7 (28)		1/4 (25)	
Ziegas and Anton-Culver, ⁷⁸ 2003	Ovarian	35/46 (76)	15/24 (63)		
Sijmons et al, ⁴⁴ 2000	Ovarian	10/15 (67)			
Verkooijen et al, ⁸⁰ 2004	Ovarian	4/6 (67)			
Theis et al, ⁷⁰ 1994	Ovarian	2/2 (100)			
King et al, ⁷⁷ 2002	Ovarian	2/4 (50)			
Summary ^a		69 (58-78)	63 (43-79)	25 (5-70)	

^aSummary data are presented as likelihood ratio (95% confidence interval).

Computerized genograms can also be effective and convenient tools for both patients and physicians.⁹⁹⁻¹⁰¹ These tools offer the benefits of paper-based systems and, through clinical decision support, educate patients and offer guidance to physicians.¹⁰²⁻¹⁰⁴ Sweet et al³⁷ compared family history information obtained by physicians at a comprehensive cancer clinic with those directly entered by patients into a computer program. Patients were then determined to be “high risk” for cancer

according to pedigree information collected from either the computer program or information recorded within the medical record. Of 362 computer entries, 69% had some form of family history information recorded within their medical record. A total of 101 patients were considered high risk according to their pedigree information collected from the computer program, but only 69 of these patients had information recorded within their medical record to confirm this high risk.

Table 20-5 Posttest Probabilities of Having a Family History of Cancer^a

Cancer Type	Estimated Cancer Family History Prevalence (Pretest Probability), %	Personal History of Cancer	Posttest Probability of Having a Family Cancer History in a First-Degree Relative, LR (95% CI)	
			Positive	Negative
Colon ⁸	9.4	Yes	70 (46-87)	3 (1-7)
		No	70 (40-89)	2.5 (1.0-6.0)
Breast ⁴⁸	22	Yes	92 (90-94)	1.0 (0.8-3.0)
		No	71 (60-81)	5 (2-12)
Ovarian ⁴⁸	3.5	Yes	68 (57-77)	0.7 (0.3-1.0)
		No	55 (17-88)	2.0 (0.5-7.0)
Endometrial ⁴⁶	7.8	Yes	70 (56-80)	4 (3-6)
		No	54 (16-88)	5 (3-11)
Prostate ¹¹	9.5	Yes	72 (24-97)	3 (2-4)
		No	56 (41-72)	3 (2-5)

Abbreviations: CI, confidence interval; LR, likelihood ratio.

^aPosttest probability = $\frac{[(\text{pretest probability}/1 - \text{pretest probability}) \text{LR}]}{[1 + (\text{pretest probability}/1 - \text{pretest probability}) \text{LR}]}$.

Special visits outside of the clinical encounter have also been evaluated as a means to obtain family history information. In one study,¹⁰⁵ patients observed at a single primary care practice were invited to a special visit designed to collect detailed family history information. Ten percent of patients observed in the pedigree clinic had a family history of cancer (breast, colon, melanoma, or thyroid) and some patients were referred for further care according to their pedigree. Patients were less anxious about their family history after the special visit, but this effect was not sustained beyond 12 weeks. A major limitation of the study was the poor attendance to the special clinic; only 16% of invited patients attended.

CLINICAL SCENARIO—RESOLUTION

This patient's pretest probability of having a positive family history of breast cancer before her clinical interview can be estimated at 22%.⁴⁸ After consideration of her self-reported pedigree, her posttest probability for a family history of breast cancer increased to 71% (95% CI, 60%-81%) (Table 20-5). According to the family history information presented, it is likely that this woman does in fact have a positive family history for breast cancer and that seeking confirmatory evidence is unlikely to offer any additional gain. This patient has 2 relatives affected with breast cancer, both diagnosed before 50 years and one a first-degree relative. According to online mutation prevalence tables (Box 20-2), this woman's risk estimate of harboring a deleterious mutation for either *BRCA1* or *BRCA2* was 10%, and a referral was recommended to a specialized cancer risk assessment program for counseling on genetic testing.¹⁰⁶ During pretesting counseling, she learned of the potential limitations of testing on a presymptomatic individual without a known deleterious

mutation. The genetic counselor suggested testing one of her affected relatives first to produce a more informative test. The patient said that her maternal grandmother is no longer alive, and she is not particularly close to her mother and unsure whether her mother would be willing to undergo testing. After pretesting genetic counseling, the patient decided not to be tested for a *BRCA1* or *BRCA2* mutation and planned to broach the subject of testing with her mother. According to this patient's family history, clinical breast examinations every 6 months are prescribed and the patient is shown how to perform monthly self-breast examinations. Mammography screening is initiated and an appointment is scheduled.

THE BOTTOM LINE

Family history assessment is taking on greater importance as high-risk individuals are being offered earlier screening interventions and risk-reducing therapies. Cancer family histories acquired on first-degree relatives for breast and colon cancer are likely to represent true positives and true negatives for the disease and may not require further evaluation to substantiate. However, other cancers with a familial disposition are less accurately reported.

Author Affiliations at the Time of the Original Publication

Division of General Internal Medicine, Vanderbilt University Medical Center, and Department of Veterans Affairs, Tennessee Valley Healthcare System (Dr Murff), and the Sarah Cannon Cancer Center (Dr Spigel), Nashville; Division of Gastroenterology, Brigham and Women's Hospital and Population Sciences Division, Dana-Farber Cancer Institute, Boston, Massachusetts (Dr Syngal).

Acknowledgments

This study was supported in part by grant K07CA08453 from the National Institutes of Health (Dr Syngal).

The National Institutes of Health did not participate in the design and conduct of the study; in the collection, analysis, and interpretation of the data; or in the preparation, review, or approval of the manuscript.

We thank Lori Orlando, MD, Deborah A. Fisher, MD, MHS, and Russell Rothman, MD, MPP, for their thoughtful reviews and comments in preparation of the manuscript.

REFERENCES

- Lynch HT, Brodkey FD, Lynch P, et al. Familial risk and cancer control. *JAMA*. 1976;236(6):582-584.
- Weber W, Gencik A, Muller H. Familial cancer: consequences for the oncological practice. *Cancer Detect Prev*. 1986;9(5-6):455-458.
- Pharoah PD, Day NE, Duffy S, et al. Family history and the risk of breast cancer: a systematic review and meta-analysis. *Int J Cancer*. 1997;71(5):800-809.
- Stratton JF, Pharoah P, Smith SK, Easton D, Ponder BA. A systematic review and meta-analysis of family history and risk of ovarian cancer. *Br J Obstet Gynaecol*. 1998;105(5):493-499.
- Parslov M, Lidegaard O, Klintorp S, et al. Risk factors among young women with endometrial cancer: a Danish case-control study. *Am J Obstet Gynecol*. 2000;182(1 pt 1):23-29.
- Gruber SB, Thompson WD; Cancer and Steroid Hormone Study Group. A population-based study of endometrial cancer and familial risk in younger women. *Cancer Epidemiol Biomarkers Prev*. 1996;5(6):411-417.

7. Nelson CL, Sellers TA, Rich SS, Potter JD, McGovern PG, Kushi LH. Familial clustering of colon, breast, uterine, and ovarian cancers as assessed by family history. *Genet Epidemiol.* 1993;10(4):235-244.
8. Fuchs CS, Giovannucci EL, Colditz GA, Hunter DJ, Speizer FE, Willett WC. A prospective study of family history and the risk of colorectal cancer. *N Engl J Med.* 1994;331(25):1669-1674.
9. St John DJ, McDermott FT, Hopper JL, Debney EA, Johnson WR, Hughes ES. Cancer risk in relatives of patients with common colorectal cancer. *Ann Intern Med.* 1993;118(10):785-790.
10. Weber W. Cancer control by family history. *Anticancer Res.* 1993;13(4): 1197-1201.
11. Kalish LA, McDougal WS, McKinlay JB. Family history and the risk of prostate cancer. *Urology.* 2000;56(5):803-806.
12. Lesko SM, Rosenberg L, Shapiro S. Family history and prostate cancer risk. *Am J Epidemiol.* 1996;144(11):1041-1047.
13. Cerhan JR, Parker AS, Putnam SD, et al. Family history and prostate cancer risk in a population-based cohort of Iowa men. *Cancer Epidemiol Biomarkers Prev.* 1999;8(1):53-60.
14. Vasen HF, Watson P, Mecklin JP, Lynch HT. New clinical criteria for hereditary nonpolyposis colorectal cancer (HNPCC, Lynch syndrome) proposed by the International Collaborative group on HNPCC. *Gastroenterology.* 1999;116(6):1453-1456.
15. Lynch HT, de la Chapelle A. Hereditary colorectal cancer. *N Engl J Med.* 2003;348(10):919-932.
16. Shattuck-Eidens D, Oliphant A, McClure M, et al. BRCA1 sequence analysis in women at high risk for susceptibility mutations: risk factor analysis and implications for genetic testing. *JAMA.* 1997;278(15):1242-1250.
17. American Society of Clinical Oncology. American Society of Clinical Oncology policy statement update: genetic testing for cancer susceptibility. *J Clin Oncol.* 2003;21(12):2397-2406.
18. US Preventive Services Task Force. Screening for breast cancer: recommendations and rationale. *Ann Intern Med.* 2002;137(5 pt 1):344-346.
19. US Preventive Services Task Force. Screening for prostate cancer: recommendation and rationale. *Ann Intern Med.* 2002;137(11):915-916.
20. US Preventive Services Task Force. Screening for colorectal cancer: recommendation and rationale. *Ann Intern Med.* 2002;137(2):129-131.
21. Smith RA, von Eschenbach AC, Wender R, et al. American Cancer Society guidelines for the early detection of cancer. *CA Cancer J Clin.* 2001;51(1):38-75.
22. National Institutes of Health Consensus Development Panel. National Institutes of Health Consensus Development Conference Statement: breast cancer screening for women ages 40-49, January 21-23, 1997. *J Natl Cancer Inst.* 1997;89(14):1015-1026.
23. Ferrini R, Mannino E, Ramsdell E, Hill L. Screening mammography for breast cancer: American College of Preventive Medicine practice policy statement. *Am J Prev Med.* 1996;12(5):340-341.
24. American Academy of Family Physicians. Summary of policy recommendations for periodic health examinations, revision 5.4, August 2003. <http://www.aafp.org/PreBuilt/PHErev54.pdf>. Accessed May 29, 2008.
25. Burt RW, Bishop DT, Lynch HT, Rozen P, Winawer SJ. Risk and surveillance of individuals with heritable factors for colorectal cancer: WHO Collaborating Centre for the Prevention of Colorectal Cancer. *Bull World Health Organ.* 1990;68(5):655-665.
26. Winawer S, Fletcher R, Rex D, et al. Colorectal cancer screening and surveillance: clinical guidelines and rationale—update based on new evidence. *Gastroenterology.* 2003;124(2):544-560.
27. Levin B, Murphy GP. Revision in American Cancer Society recommendations for the early detection of colorectal cancer. *CA Cancer J Clin.* 1992;42(5):296-299.
28. Burke W, Daly M, Garber J, et al; Cancer Genetics Studies Consortium. Recommendations for follow-up care of individuals with an inherited predisposition to cancer, II: BRCA1 and BRCA2. *JAMA.* 1997;277(12):997-1003.
29. Burke W, Petersen G, Lynch P, et al; Cancer Genetics Studies Consortium. Recommendations for follow-up care of individuals with an inherited predisposition to cancer, I: hereditary nonpolyposis colon cancer. *JAMA.* 1997;277(11):915-919.
30. National Institutes of Health Consensus Development Conference Statement. Ovarian cancer: screening, treatment, and follow-up. *Gynecol Oncol.* 1994;55(3 pt 2):S4-S14.
31. Fisher B, Costantino JP, Wickerham DL, et al. Tamoxifen for prevention of breast cancer: report of the National Surgical Adjuvant Breast and Bowel Project P-1 Study. *J Natl Cancer Inst.* 1998;90(18):1371-1388.
32. Cauley JA, Norton L, Lippman ME, et al. Continued breast cancer risk reduction in postmenopausal women treated with raloxifene: 4-year results from the MORE Trial. *Breast Cancer Res Treat.* 2001;65(2):125-134.
33. Hartmann LC, Schaid DJ, Woods JE, et al. Efficacy of bilateral prophylactic mastectomy in women with a family history of breast cancer. *N Engl J Med.* 1999;340(2):77-84.
34. Severin MJ. Hereditary cancer litigation: a status report. *Oncology (Williston Park).* 1996;10(2):211-214, 217.
35. Douglas FS, O'Dair LC, Robinson M, Evans DG, Lynch SA. The accuracy of diagnoses as reported in families with cancer: a retrospective study. *J Med Genet.* 1999;36(4):309-312.
36. Kerr B, Foulkes WD, Cade D, et al. False family history of breast cancer in the family cancer clinic. *Eur J Surg Oncol.* 1998;24(4):275-279.
37. Sweet KM, Bradley TL, Westman JA. Identification and referral of families at high risk for cancer susceptibility. *J Clin Oncol.* 2002;20(2):528-537.
38. Fry A, Campbell H, Gudmundsdottir H, et al. GPs' views on their role in cancer genetics services and current practice. *Fam Pract.* 1999;16(5):468-474.
39. Emery J, Rose P. Expanding the role of the family history in primary care [editorial]. *Br J Gen Pract.* 1999;49(441):260-261.
40. Harris R, Harris H. Genetics in primary care: report on workshop of EC Concerted Action on Genetics Services in Europe (CAGSE) in association with the Royal College of GP Spring Meeting, Blackpool, UK, 28 April 1995. *J Med Genet.* 1996;33(4):346-348.
41. Emery J, Watson E, Rose P, Andermann A. A systematic review of the literature exploring the role of primary care in genetic services. *Fam Pract.* 1999;16(4):426-445.
42. Hunter A, Wright P, Cappelli M, Kasaboski A, Surh L. Physician knowledge and attitudes towards molecular genetic (DNA) testing of their patients. *Clin Genet.* 1998;53(6):447-455.
43. Gollust SE, Hull SC, Wilfond BS. Limitations of direct-to-consumer advertising for clinical genetic testing. *JAMA.* 2002;288(14):1762-1767.
44. Sijmons RH, Boonstra AE, Reehuis J, et al. Accuracy of family history of cancer: clinical genetic implications. *Eur J Hum Genet.* 2000;8(3):181-186.
45. Johnson N, Lancaster T, Fuller A, Hodgson SV. The prevalence of a family history of cancer in general practice. *Fam Pract.* 1995;12(3):287-289.
46. Scheuner MT, Wang SJ, Raffel LJ, Larabell SK, Rotter JI. Family history: a comprehensive genetic risk assessment method for the chronic conditions of adulthood. *Am J Med Genet.* 1997;71(3):315-324.
47. Slattery ML, Kerber RA. A comprehensive evaluation of family history and breast cancer risk: the Utah Population Database. *JAMA.* 1993;270(13): 1563-1568.
48. Hughes KS, Roche C, Campbell CT, et al. Prevalence of family history of breast and ovarian cancer in a single primary care practice using a self-administered questionnaire. *Breast J.* 2003;9(1):19-25.
49. Makinen T, Tammela TL, Stenman UH, et al. Family history and prostate cancer screening with prostate-specific antigen. *J Clin Oncol.* 2002; 20(11):2658-2663.
50. Lynch HT, Kimberling W, Albano WA, et al. Hereditary nonpolyposis colorectal cancer (Lynch syndromes I and II), I: clinical description of resource. *Cancer.* 1985;56(4):934-938.
51. Thompson WD, Schildkraut JM. Family history of gynaecological cancers: relationships to the incidence of breast cancer prior to age 55. *Int J Epidemiol.* 1991;20(3):595-602.
52. Parazzini F, La Vecchia C, Negri E, Franceschi S, Tozzi L. Family history of breast, ovarian and endometrial cancer and risk of breast cancer. *Int J Epidemiol.* 1993;22(4):614-618.
53. Schmidt S, Becher H, Chang-Claude J. Breast cancer risk assessment: use of complete pedigree information and the effect of misspecified ages at diagnosis of affected relatives. *Hum Genet.* 1998;102(3):348-356.
54. Bennett RL, Steinhause KA, Uhlrich SB, et al. Recommendations for standardized human pedigree nomenclature. *Am J Hum Genet.* 1995;56(3):745-752.
55. Bennett RL. *The Practical Guide to the Genetic Family History.* New York, NY: Wiley-Liss; 1999.
56. Mussio P, Weber W, Brunetti D, Stemmermann GN, Torhorst J. Taking a family history in cancer patients with a simple questionnaire. *Anticancer Res.* 1998;18(4B):2811-2814.
57. Zippin C, Lum D, Hankey BF. Completeness of hospital cancer case reporting from the SEER Program of the National Cancer Institute. *Cancer.* 1995;76(11):2343-2350.
58. Hawkins MM, Swerdlow AJ. Completeness of cancer and death follow-up obtained through the National Health Service Central Register for England and Wales. *Br J Cancer.* 1992;66(2):408-413.

59. Penberthy L, McClish D, Pugh A, et al. Using hospital discharge files to enhance cancer surveillance. *Am J Epidemiol.* 2003;158(1):27-34.
60. Korhonen P, Malila N, Pukkala E, et al. The Finnish Cancer Registry as follow-up source of a large trial cohort: accuracy and delay. *Acta Oncol.* 2002;41(4):381-388.
61. Dickinson HO, Salotti JA, Birch PJ, et al. How complete and accurate are cancer registrations notified by the National Health Service Central Register for England and Wales? *J Epidemiol Community Health.* 2001;55(6):414-422.
62. Bickell NA, Chassin MR. Determining the quality of breast cancer care: do tumor registries measure up? *Ann Intern Med.* 2000;132(9):705-710.
63. Bergmann MM, Calle EE, Mervis CA, et al. Validity of self-reported cancers in a prospective cohort study in comparison with data from state cancer registries. *Am J Epidemiol.* 1998;147(6):556-562.
64. Mukherjee AK, Leck I, Langley FA, Ashcroft C. The completeness and accuracy of health authority and cancer registry records according to a study of ovarian neoplasms. *Public Health.* 1991;105(1):69-78.
65. Izquierdo JN, Schoenbach VJ. The potential and limitations of data from population-based state cancer registries. *Am J Public Health.* 2000;90(5):695-698.
66. James DS, Bull AD. Information on death certificates: cause for concern? *J Clin Pathol.* 1996;49(3):213-216.
67. Kircher T, Nelson J, Burdo H. The autopsy as a measure of accuracy of the death certificate. *N Engl J Med.* 1985;313(20):1263-1269.
68. Love RR, Evans AM, Josten DM. The accuracy of patient reports of a family history of cancer. *J Chronic Dis.* 1985;38(4):289-293.
69. Breuer B, Kash KM, Rosenthal G, Diemer K, Osborne MP, Miller DG. Reporting bilaterality status in first-degree relatives with breast cancer: a validity study. *Genet Epidemiol.* 1993;10(4):245-256.
70. Theis B, Boyd N, Lockwood G, Tritchler D. Accuracy of family cancer history in breast cancer patients. *Eur J Cancer Prev.* 1994;3(4):321-327.
71. Aitken J, Bain C, Ward M, Siskind V, MacLennan R. How accurate is self-reported family history of colorectal cancer? *Am J Epidemiol.* 1995;141(9):863-871.
72. Parent ME, Ghadirian P, Lacroix A, Perret C. Accuracy of reports of familial breast cancer in a case-control series. *Epidemiology.* 1995;6(2):184-186.
73. Anton-Culver H, Kurosaki T, Taylor TH, Gildea M, Brunner D, Bringman D. Validation of family history of breast cancer and identification of the BRCA1 and other syndromes using a population-based cancer registry. *Genet Epidemiol.* 1996;13(2):193-205.
74. Kerber RA, Slattery ML. Comparison of self-reported and database-linked family history of cancer data in a case-control study. *Am J Epidemiol.* 1997;146(3):244-248.
75. Eerola H, Blomqvist C, Pukkala E, Pyrhonen S, Nevanlinna H. Familial breast cancer in southern Finland: how prevalent are breast cancer families and can we trust the family history reported by patients? *Eur J Cancer.* 2000;36(9):1143-1148.
76. Katballe N, Juul S, Christensen M, Orntoft TF, Wikman FP, Laurberg S. Patient accuracy of reporting on hereditary non-polyposis colorectal cancer-related malignancy in family members. *Br J Surg.* 2001;88(9):1228-1233.
77. King TM, Tong L, Pack RJ, Spencer C, Amos CI. Accuracy of family history of cancer as reported by men with prostate cancer. *Urology.* 2002;59(4):546-550.
78. Ziogas A, Anton-Culver H. Validation of family history data in cancer family registries. *Am J Prev Med.* 2003;24(2):190-198.
79. Mitchell RJ, Brewster D, Campbell H, et al. Accuracy of reporting of family history of colorectal cancer. *Gut.* 2004;53(2):291-295.
80. Verkooijen HM, Fioretta G, Chappuis PO, et al. Set-up of a population-based familial breast cancer registry in Geneva, Switzerland: validation of first results. *Ann Oncol.* 2004;15(2):350-353.
81. Ivanovich J, Babb S, Goodfellow P, et al. Evaluation of the family history collection process and the accuracy of cancer reporting among a series of women with endometrial cancer. *Clin Cancer Res.* 2002;8(6):1849-1856.
82. Koch M, Hill GB. Problems in establishing accurate family history in patients with ovarian cancer of epithelial origin. *Cancer Detect Prev.* 1987;10(3-4):279-283.
83. Douglas FS, O'Dair LC, Robinson M, et al. The accuracy of diagnosis as reported in families with cancer: a retrospective study. *J Med Genet.* 1999;36(4):309-312.
84. Gierlowski TC, Ron E, Schneider AB. The accuracy of siblings' family history reports of thyroid and other cancers. *Am J Public Health.* 1995;85(3):408-409.
85. Novakovic B, Goldstein AM, Tucker MA. Validation of family history of cancer in deceased members. *J Natl Cancer Inst.* 1996;88(20):1492-1493.
86. Kuijten RR, Bunin GR, Meadows AT. Validation and accuracy of cancer reporting in first-degree relatives [abstract]. *Am J Epidemiol.* 1992;136(8):1019.
87. Airewele G, Adatto P, Cunningham J, et al. Family history of cancer in patients with glioma: a validation study of accuracy. *J Natl Cancer Inst.* 1998;90(7):543-544.
88. Bondy ML, Strom SS, Colopy MW, Brown BW, Strong LC. Accuracy of family history of cancer obtained through interviews with relatives of patients with childhood sarcoma. *J Clin Epidemiol.* 1994;47(1):89-96.
89. Salkind AR, Cuddy PG, Foxworth JW. Is this patient allergic to penicillin? an evidence-based analysis of the likelihood of penicillin allergy. *JAMA.* 2001;285(19):2498-2505.
90. Simel DL, Samsa GP, Matchar DB. Likelihood ratios with confidence: sample size estimation for diagnostic test studies. *J Clin Epidemiol.* 1991;44(8):763-770.
91. Garbers V, Toniolo PG, Taioli E. Changes in self-reported family history of breast cancer with change in case-control status. *Eur J Epidemiol.* 2001;17(6):517-520.
92. Novakovic B, Goldstein AM, Tucker MA. Validation of family history of cancer in deceased family members. *J Natl Cancer Inst.* 1996;88(20):1492-1493.
93. Weinstock MA, Brodsky GL. Bias in the assessment of family history of melanoma and its association with dysplastic nevi in a case-control study. *J Clin Epidemiol.* 1998;51(12):1299-1303.
94. Desai MM, Bruce ML, Desai RA, Druss BG. Validity of self-reported cancer history: a comparison of health interview data and cancer registry records. *Am J Epidemiol.* 2001;153(3):299-306.
95. Loescher LJ. The family history component of cancer genetic risk counseling. *Cancer Nurs.* 1999;22(1):96-102.
96. Blumenthal D, Causino N, Chang YC, et al. The duration of ambulatory visits to physicians. *J Fam Pract.* 1999;48(4):264-271.
97. Cole J, Conneally PM, Hodes ME, Merritt AD. Genetic family history questionnaire. *J Med Genet.* 1978;15(1):10-18.
98. Gilpin CA, Carson N, Hunter AG. A preliminary validation of a family history assessment form to select women at risk for breast or ovarian cancer for referral to a genetics center. *Clin Genet.* 2000;58(4):299-308.
99. Gerson R, McGoldrick M. The computerized genogram. *Prim Care.* 1985;12(3):535-545.
100. Ebell MH, Heaton CJ. Development and evaluation of a computer genogram. *J Fam Pract.* 1988;27(5):536-538.
101. Westman J, Hampel H, Bradley T. Efficacy of a touchscreen computer based family cancer history questionnaire and subsequent cancer risk assessment. *J Med Genet.* 2000;37(5):354-360.
102. Emery J. Computer support for genetic advice in primary care. *Br J Gen Pract.* 1999;49(444):572-575.
103. Green MJ, Fost N. An interactive computer program for educating and counseling patients about genetic susceptibility to breast cancer. *J Cancer Educ.* 1997;12(4):204-208.
104. Green MJ, Fost N. Who should provide genetic education prior to gene testing? computers and other methods for improving patient understanding. *Genet Test.* 1997;1(2):131-136.
105. Rose P, Humm E, Hey K, Jones L, Huson SM. Family history taking and genetic counselling in primary care. *Fam Pract.* 1999;16(1):78-83.
106. Fletcher SW, Elmore JG. Clinical practice: mammographic screening for breast cancer. *N Engl J Med.* 2003;348(17):1672-1680.

UPDATE : Family History of Cancer

20

Prepared by David L. Simel, MD, MHS
Reviewed by Harvey Murff, MD, MPH

CLINICAL SCENARIO

A 48-year-old woman makes an urgent appointment to see you. She is distressed because her 52-year-old sister just returned home from an outpatient colonoscopy procedure and called to tell her that she has cancer. Your patient is healthy, has a normal well-balanced diet, and has no abnormal bowel symptoms. She wants to know what she should do.

UPDATED SUMMARY ON FAMILY HISTORY OF CANCER

Original Review

Murff HJ, Spigel DR, Syngal S. Does this patient have a family history of cancer? an evidence-based analysis of the accuracy of family cancer history. *JAMA*. 2004;292(12):1480-1489.

UPDATED LITERATURE SEARCH

Our literature search replicated the search strategy reported in the original article, limited to 2004-2006. The results yielded 32 titles, for which we reviewed the abstracts. None of these studies reported both the sensitivity and specificity of the family history for cancer as obtained from healthy individuals in the clinic office setting. One study evaluated the sensitivity of the family history among patients who had a first-degree relative with either the Li-Fraumeni syndrome or the breast–ovarian cancer syndrome.¹

NEW FINDINGS

Details of the Update

No new studies assessed the accuracy of the family medical history in an unselected general medical population.

CHANGES IN THE REFERENCE STANDARD

None.

RESULTS OF LITERATURE REVIEW

A study of patients in a genetic screening clinic because they have a first-degree relative with a breast cancer syndrome provides some insight into the factors that might

affect the accuracy of a family history of carcinoma.¹ The accuracy of the patient's report depended on the actual genetic syndrome. Perhaps, not surprisingly, for 2 breast cancer syndromes the history from female first-degree relatives was more accurate than the family history elicited from male first-degree relatives. Those with a college education were more accurate than less-educated persons; first-degree relatives of the affected individual were more accurate than second-degree relatives; however, age did not affect the accuracy. Given the select population for this study, we do not know whether these factors generalize to other populations. The higher specificity of the family history reported by women was validated in a population-based sample of patients.² However, the same population-based study found higher specificity for family histories reported by younger (<50 years) patients and no difference as a function of the consultand or maternal or paternal level of education.

EVIDENCE FROM GUIDELINES

All standard physical examination and clinical textbooks recommend that clinicians elicit a family history. Guidelines for specific cancers depend on accurate family histories.³

CLINICAL SCENARIO—RESOLUTION

Typically, screening for colon cancer begins at aged 50 years. However, you might obtain a colonoscopy now, depending on the family history. This case scenario highlights the importance of confirming the medical history. Had the situation been different in that the patient's sister called a week after the colonoscopy with the report of cancer, the likelihood is high (likelihood ratio, 23) that your patient's report of a family history of colon cancer would be accurate. Although it is certainly possible that the sister's physician told her she had carcinoma from the colonoscopic findings, it would be prudent to wait for confirmation. You should explain to your patient that it is important for you both to understand the exact colonoscopy results (eg, the presence of multiple polyps) and the biopsy results (to confirm the presence of cancer). Once you have those findings, you can discuss with the patient the appropriate timing and approach to screening for colon cancer.

FAMILY HISTORY OF CANCER—MAKE THE DIAGNOSIS

PRIOR PROBABILITY

The prior probability of a family history of any carcinoma depends on the specific cancer. The general rates are as shown in [Table 20-6](#).

Table 20-6 Prevalence of Family History of Some Common Cancers

Cancer	Family History Prevalence, %
Breast	5-22
Colon	2-9.4
Ovarian	1.1-3.5
Endometrial	0.5-1.4
Prostate	4.6-9.5

POPULATION FOR WHOM A FAMILY HISTORY OF CANCER SHOULD BE CONSIDERED

A family history that addresses cancer should be obtained from all patients. However, the field of genetics and personal risk assessment is changing rapidly, and physicians will need to get further education based on new data that describe a myriad of genetic associations with cancer. Online assessment tools can help patient assess their individual risk (<http://www.yourdiseaserisk.wustl.edu>; accessed May 29, 2008). The *BRCA* mutation, a particularly strong risk factor for breast or ovarian cancer, has specific online resources for assessing risk, although all risk assessments depend on accurate information from the patient (<http://astor.som.jhmi.edu/BayesMendel/> or <http://www.myriadtests.com/provider/brca-mutation-prevalence.htm>; accessed May 29, 2008).

DETECTING THE LIKELIHOOD OF A FIRST-DEGREE RELATIVE WITH CANCER

A healthy patient who reports no family history of cancer will most likely be correct. However, even among patients with a personal history of cancer, the accuracy of a positive report of cancer in first-degree relatives may sometimes require confirmation, depending on the specific surveillance or genetic screening plan (see [Tables 20-7](#) and [20-8](#)).

Table 20-7 Likelihood Ratio of a Healthy Patient's Reported Family History for Cancer

Family History of Carcinoma	Healthy Patients	
	LR+ (95% CI)	LR- (95% CI)
Ovarian	34 (5.7-202)	0.51 (0.13-2.1)
Colon	23 (6.4-81)	0.25 (0.10-0.63)
Endometrial	14 (2.2-83)	0.68 (0.31-1.5)
Prostate	12 (6.5-24)	0.32 (0.18-0.63)
Breast	8.9 (5.4-15)	0.20 (0.08-0.49)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

Table 20-8 Likelihood Ratio of an Affected Patient's Reported Family History for Cancer

Family History of Carcinoma	Patient With Personal History of Cancer	
	LR+ (95% CI)	LR- (95% CI)
Ovarian	44 (15-132)	0.21 (0.12-0.37)
Colon	23 (8.1-64)	0.29 (0.13-0.67)
Endometrial	20 (4.3-89)	0.55 (0.35-0.86)
Prostate	24 (2.3-262)	0.25 (0.16-0.39)
Breast	41 (23-75)	0.07 (0.03-0.13)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

REFERENCE STANDARD TESTS

Verification of cancer from the first-degree relative's medical record, physician, population cancer registry, or autopsy.

REFERENCES FOR THE UPDATE

- Schneider KA, DiGianni LM, Patenaude AF, et al. Accuracy of cancer family histories: comparison of two breast cancer syndromes. *Genet Test*. 2004;8(3):222-228.
- Chang ET, Smedby KE, Hjalgrim H, Glimelius B, Adami HO. Reliability of self-reported family history of cancer in a large case-control study of lymphoma. *J Natl Cancer Inst*. 2006;98(1):61-68.
- US Preventive Services Task Force. Genetic risk assessment and *BRCA* mutation testing for breast and ovarian cancer susceptibility: recommendation statement. *Ann Intern Med*. 2005;143(5):355-361.

CHAPTER 21

Does This Patient Have a Goiter?

Kerry Siminoski, MD

CLINICAL SCENARIOS

How Large Are These Thyroid Glands?

For each of the following patients, assessment of thyroid size is an important part of the clinical examination. In case 1, a 32-year-old woman presents with symptoms and findings consistent with hyperthyroidism, but she has no exophthalmos and has always been anxious. In case 2, a 55-year-old man has a diagnosis of Graves disease, and the choice is made for radioactive iodine ablation therapy. In case 3, a 64-year-old man has a goiter that causes discomfort on swallowing, and thyroxine is administered in an attempt to shrink the thyroid gland.

WHY ASSESS THE THYROID GLAND FOR SIZE?

A goiter is simply an enlargement of the thyroid gland and may result from hormonal or immunologic stimulation of gland growth or the presence of inflammatory, proliferative, infiltrative, or metabolic disorders (Table 21-1). A common error among clinicians first learning about the thyroid is to associate thyroid size with function; a goiter, however, can be present in hyperthyroidism, hypothyroidism, or a euthyroid state. Determining whether a thyroid is enlarged can aid in diagnosis, differential diagnosis, and decisions about laboratory testing; in determining specific therapy and therapeutic dosing; and subsequently in monitoring of the clinical course. For example, when a patient presents with symptoms that could be caused by hyperthyroidism, the detection of a goiter increases the likelihood that thyrotoxicosis is present.² If the patient described in the first case had an enlarged thyroid, hyperthyroidism would be a likely diagnosis.² On the other hand, if her gland were of normal size, anxiety might be the explanation for her symptoms. Determination of thyroid size also is useful once a specific disease is diagnosed. In patients with Graves disease, for example, thyroid size may be a factor in determining choice of treatment because patients with smaller glands are more likely to go into immunologic remission during antithyroid drug therapy.³ If radioactive iodine is the chosen treatment, as in the second case, the size of the gland is often used in calculating the dose to be administered.⁴ Finally, responses to various therapies can be monitored clinically by assessing thyroid size, such as the attempt to shrink a large goiter with thyroid hormone administration in the third case.⁵

THE ANATOMIC BASIS OF THYROID EXAMINATION

Landmarks and Relation to Other Structures

The thyroid gland is located in the anterior neck and usually consists of 2 lobes connected at their lower midregions by a

transverse isthmus (Figure 21-1). The most prominent structure in the anterior neck is the thyroid cartilage. Inferior to the thyroid cartilage lies the cricoid cartilage, and inferior to this lies the isthmus of the thyroid gland, which can be as low as the level of the fourth tracheal ring. Each thyroid lobe lies against the sides of the trachea, extending up from the isthmus to the region of the cricoid and thyroid cartilages and downward toward the clavicles. The posterior portion of each lobe lies beneath the belly of the ipsilateral sternocleido-

mastoid muscle. Because the fascial envelope of the thyroid gland is continuous with the pretracheal fascia of the cricoid cartilage and hyoid bone, the thyroid ascends and descends with the laryngeal structures during swallowing.

How Large Is the Normal Thyroid?

The normal thyroid size for a population is largely determined by the supply of iodine in the diet, with a tendency to larger glands in iodine-deficient areas.⁶⁻⁸ Consequently, studies of clinically normal thyroid glands have demonstrated sizes that span an extreme range in euthyroid individuals, differing by geographic location and varying through time within a given region as iodine supplementation has been instituted. Until the middle of this century, most authors considered a typical thyroid gland to be about 20-25 g, and a commonly accepted upper normal size was 35 g.⁸⁻¹¹ More recent studies in iodine-supplemented populations have reported mean weights of 10 g or less and an upper normal size of 20 g.^{12,13} Although a value of 35 g may still apply in iodine-deficient areas, an upper normal weight of 20 g is probably appropriate for most parts of the western world and will be used for this analysis. With this definition, the prevalence of goiter is typically 2% to 5% in iodine-replete regions.^{13,14}

Table 21-1 Conditions That May Present With an Enlarged Thyroid Gland^a

Endemic/iodine deficiency goiter

Multinodular goiter

Graves disease

Hashimoto thyroiditis

Subacute thyroiditis

Painless/postpartum thyroiditis

Familial goiter

Malignancy

Goitrogens

Iodine excess

^aAdapted from Eastham.¹

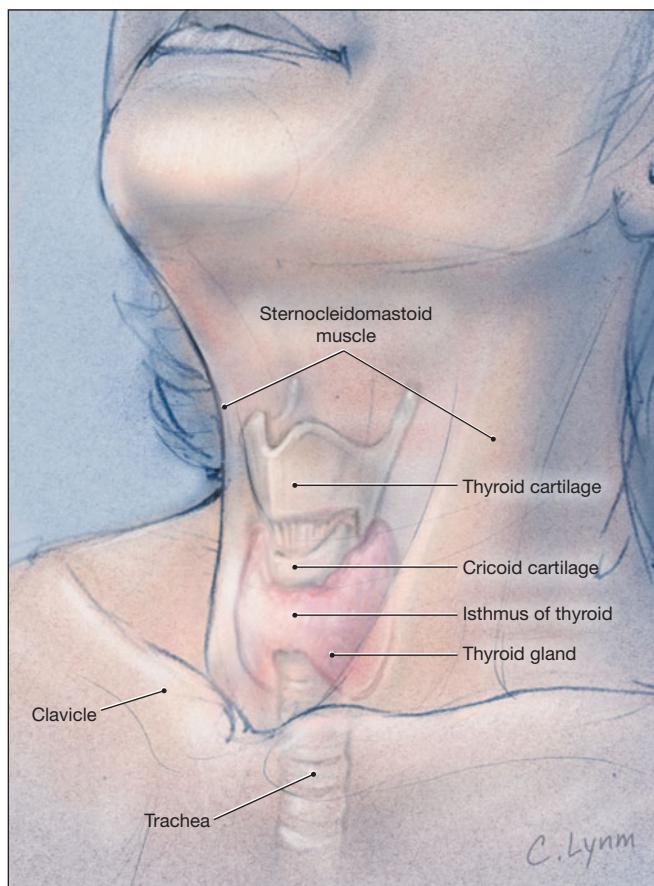


Figure 21-1 The Location of the Thyroid Gland in Relationship to Nearby Structures

HOW TO EXAMINE THE THYROID GLAND TO DETERMINE SIZE

The normal thyroid is rarely visible because of its relatively small size, partial concealment by the sternocleidomastoids, and soft texture, and it may be marginally palpable.^{5,9,15} Enlargement is initially observed as an increase in the size of the lateral lobes to palpation.^{5,8} Further growth results in a gland visible in the anterior side of the neck that can be seen when inspecting from the side^{16,17} and from the front with the patient's neck extended.^{5,7,15,18} With increasing size, the gland becomes even more prominent on inspection from the side and it becomes visible from the front with the patient's head in a normal position. Ultimately, a large goiter is easily palpable, has prominence from the side of greater than 1 cm, and is visible from the front at a distance.^{5,17,18}

As a result of observations on these patterns of enlargement, various systems have been described to size a thyroid gland according to (1) the estimated weight¹⁹⁻²¹; (2) the volume relative to the size of normal glands^{5,8}; (3) the presence or absence of palpable or visible enlargement^{8,18,22}; (4) the degree of visible prominence when the neck is viewed laterally¹⁷; (5) neck circumference determined by tape measure^{23,24}; (6) the surface area of the gland projected onto the skin^{22,25}; and (7) the maximum width of the lower poles, measured with a ruler or calipers.²⁶ Many of these rating scales were developed for epidemiologic studies of goiter in endemic areas and were intended to classify significant goiters rapidly (with examination time in some studies averaging only 18 seconds per subject).⁵ As a result, many are of little use for the smaller thyroid glands observed in regions without significant levels of endemic goiter. Most studies from which data for accuracy and precision of goiter deter-

mination can be derived do not report specifics of thyroid examination technique. Consequently, there is no objective evidence to support the use of one examination method over another.^{23-25,27,28} Many of the variations are minor, so shared features will be described.

The patient should be comfortably positioned, either standing or seated, with the neck in a neutral position or slightly extended. The region of the neck below the thyroid or cricoid cartilage should be observed from the front, with good cross-lighting to accentuate shadows and highlight masses. If an abnormality is suspected, the neck should be moved as appropriate to alter the prominence of the area under suspicion. A particularly useful maneuver is inspection during full extension of the patient's neck. This position stretches superficial tissues over the thyroid gland, which is pressed against the relatively unyielding trachea, and enhances visibility of the gland. Inspection of the neck from the side, looking for a prominence protruding from the normally smooth and straight contour between the cricoid cartilage and the suprasternal notch, can reveal enlargement.¹⁷ The amount of prominence should be measured with a ruler (Figure 21-2). This method requires a certain degree of guesswork in deducing where the normal neck contour would lie, but the measurement can provide information useful for ruling in the presence of a goiter. There is no particular spot to place the ruler; it merely serves as a visual guide in estimating the degree of protrusion.

After inspection, the gland is palpated, and this is where the greatest differences in methods arise. Clinician preference varies about palpation with fingers or thumbs, an approach from the front or from behind the patient, and whether each lobe is palpated by the ipsilateral hand or the opposite hand. In the absence of data to support a specific method, though, examiners should use the approach with which they are most comfortable. Regardless of the technique used, it is often useful to first attempt to locate the thyroid isthmus by palpating between the cricoid cartilage and suprasternal notch. An isthmus may not be felt, but if it is, this can help locate the gland. When palpating the lobes, it is beneficial to relax the sternocleidomastoids. To better feel the left lobe, for example, the neck can be slightly flexed and rotated to the left to relax the left sternocleidomastoid and to make space for the palpating fingers or thumb between the sternocleidomastoid and trachea. There are certain additional maneuvers that may be useful, such as measuring neck circumference or the dimensions of a lobe with calipers, but no information is available to assess accuracy or precision of these techniques. Other elements of the thyroid examination that are carried out concomitantly with size assessment include determining gland texture, gland mobility, tenderness, and the presence of nodularity. Auscultation also may be performed for the presence of bruits. These features have their own implications but are not central to determining the presence of a goiter and so are beyond the scope of this discussion. If no thyroid is detected in the neck, it may be maldescended or intrathoracic. Methods of examining for these variants will not be discussed here, because, again, no information is available to analyze the reported techniques.

Dogma holds that the thyroid examination is improved by having the patient swallow during both inspection and palpation. Indeed, it has been stated that swallowing increases sensitivity of inspection alone to that of inspection combined with palpation.²⁸ No study, however, has actually analyzed whether a swallowing maneuver is of benefit, although most examiners believe it is. The movement resulting from swallowing accomplishes several things. First, it changes the shadowing of any mass, enhancing visual detection of a bulge in the neck contour that may be too subtle to be detected otherwise. Second, movement of the thyroid raises a low-placed gland up from below the sternal notch or lower sternocleidomastoid, making it accessible when it may not have been so previously. Third, as in any palpation technique, movement of the object against the palpating hand increases definition. Finally, because only the larynx, upper trachea, and thyroid gland move with swallowing, this maneuver can aid in anatomic localization.²⁹ The degree of excursion of the thyroid on swallowing is proportional to the size of the bolus swallowed, so the patient should be given a sip of water.³⁰

When the thyroid is examined to determine the presence of a goiter, the goal is to estimate gland size. Most endocrinologists express findings in absolute mass or as relative to an upper limit of a normal-sized gland, such as "normal" or "2 to 3 times normal size." Many nonendocrinologists have some difficulty quantifying thyroid mass, but this ability is crucial in accurately classifying a gland, as will be discussed in the analysis of accuracy.

FALSE-POSITIVE AND FALSE-NEGATIVE GOITER RESULTS

Finding a goiter when one is not present may simply be an error in detection. There are, however, several common causes of a false-positive goiter or pseudogooiter finding. One is simply an easily palpable gland in a thin individual.⁵

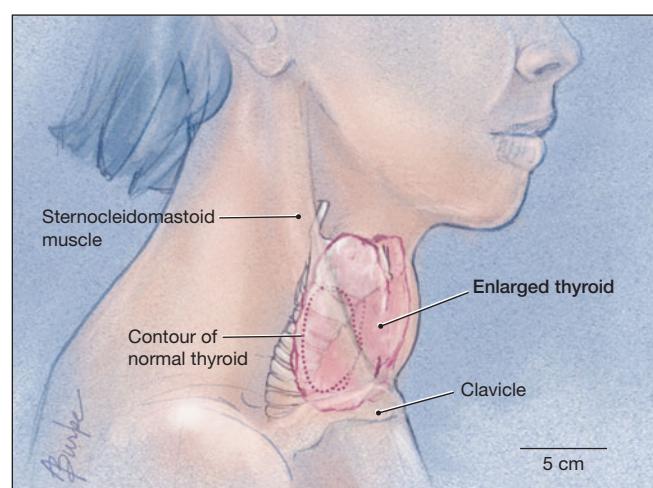


Figure 21-2 Estimating Lateral Thyroid Prominence

When viewed from the side, the normal contour of the thyroid gland is invisible. Enlargement up and out leads to a prominent and visible gland.

Table 21-2 Interobserver Precision in Assessment of Thyroid Size or Presence of Goiter

Reference	Agreement		κ	
	All Categories ^a	Goiter Only ^b	All Categories ^a	Goiter Only ^b
Trotter et al ^{36c}	0.67	0.83	0.48	0.50
Kilpatrick et al ^{38d}	0.86	0.95	0.74	0.77
Dingle et al ^{37e}	0.85	0.87	0.47	0.38
Trowbridge et al ^{7d}	Not available	0.96	Not available	0.58
Combined (95% CI)	0.86 ^f (0.82-0.90)	0.92 (0.90-0.94)	0.70 ^f (0.68-0.72)	0.77 (0.76-0.79)

Abbreviation: CI, confidence interval.

^aAnalysis of all size categories used by authors.^bAnalysis of presence or absence of goiter only, according to authors' definitions.^cAgreement between 2 observers, in 3 categories of staging, after an unspecified time.^dAgreement between one observer and one or two others, in 4 categories of staging, after an unspecified time.^eAgreement between one observer and 2 others, in 4 categories of staging, after 2 years.^fRaw data combined from Kilpatrick et al³⁸ and Dingle et al³⁷ only because they are the only 2 with the same rating scales.**Table 21-3** Comparison of Interobserver Precision for Thyroid Inspection and Palpation

Reference	Agreement		κ	
	Inspection	Palpation	Inspection	Palpation
Kilpatrick et al ³⁸	0.95	0.89	0.77	0.76
Dingle et al ³⁷	0.87	0.89	0.38	0.60
Combined (95% CI)	0.93 (0.90-0.96)	0.89 (0.85-0.92)	0.65 (0.62-0.69)	0.74 (0.67-0.82)

Abbreviation: CI, confidence interval.

Table 21-4 Intraobserver Precision in Assessment of Thyroid Size or Presence of Goiter

Reference	Agreement		κ	
	All Categories ^a	Goiter Only ^b	All Categories ^a	Goiter Only ^b
Hennessy ^{6c}	0.83	0.90	0.70	0.79
MacLennan et al ^{22d}	0.79	0.82	0.41	0.47
Combined (95% CI)	0.81 (0.77-0.84)	0.85 (0.82-0.88)	0.59 (0.52-0.65)	0.65 (0.63-0.67)

Abbreviation: CI, confidence interval.

^aAnalysis of all size categories used by authors.^bPresence or absence of goiter only, according to authors' definitions.^cFour categories of staging, reexamined within 44 days.^dThree categories of staging, examined 12 days later.

Because the entire thyroid is so accessible, the tendency is to interpret this accessibility as being due to an enlarged gland rather than the true reason, a decrease in interfering tissues that normally block access to the gland. A second

cause is a variant of the normal placement of the thyroid gland in the neck. In some individuals, the gland is higher than usual, and this prominence is again attributed to enlargement.³¹ A third anatomic variant has been termed Modigliani syndrome.³² In Modigliani syndrome, the thyroid actually lies in a normal position below the cricoid cartilage, but such individuals possess long, curving necks that enhance the prominence and palpability of the gland. A fourth condition producing pseudogoiiter is a fat pad in the anterior and lateral portion of the neck.²⁴ Although this condition may be more common in obese individuals, it can also be found in those of normal weight, particularly young women. With experience, examiners can learn to differentiate this from true thyroid tissue by the differing textures and shapes and the lack of movement of a fat pad with swallowing. Another cause involves the thyroid being pushed forward by lesions behind it, making it more easily palpable.^{5,33} Finally, any enlargement in the vicinity of the thyroid gland may be mistaken for an enlarged thyroid gland, particularly if it is adherent to the thyroid or larynx and so moves with swallowing.²⁹

There are 3 principal causes of false-negative goiter detection in addition to true misclassification. The first and probably most common cause, of course, is an inadequate physical examination. In some circumstances, an imperfect examination is unavoidable, as when a patient is intubated. In most cases, however, with a little effort, a good examination can be performed on virtually all patients. Second, some individuals, particularly the obese, the elderly, or those with chronic pulmonary disease, have short and thick necks, obscuring the thyroid.^{5,24,34} Some patients also have an atypical thyroid placement, such as a retrosternal location, or lobes that are lateral and obscured by the sternocleidomastoids, making palpation difficult.³⁵

PRECISION OF ESTIMATING THYROID SIZE

Interobserver Variability

Data on interobserver precision in estimating thyroid size are available both for rating scales that attempted to place glands in one of 3 or 4 categories according to palpability and visibility and for simple estimation of the presence or absence of a goiter (Table 21-2). Agreements were good to very good in both cases. When glands were placed in categories, κ ranged from 0.47 to 0.74, with a value from combined data of 0.70 (95% confidence interval [CI], 0.68-0.72).^{7,8,36,37} (The κ statistic and other statistical measures are defined in the introductory article to this series.³⁸) For determination of goiter, κ ranged in these 4 studies from 0.38 to 0.77, with a value for combined data of 0.77 (95% CI, 0.76-0.79). Similar results were reported in another study,³⁹ in which observers determined whether individual lobes were enlarged, with κ from 0.32 to 0.62, and in yet another report⁴⁰ that determined the presence of a goiter, with κ from 0.10 to 0.54. Because of the nature of the rating scales used in 2 of these studies,^{8,37} we can specifically compare interobserver variability for the techniques of

inspection ($\kappa = 0.65$; 95% CI, 0.62-0.69) and palpation ($\kappa = 0.74$; 95% CI, 0.67-0.82). These techniques did not differ significantly in the level of agreement, and both were very good (Table 21-3).

As might be expected, most disagreements between observers involved smaller glands and those near the cutoff for goiter determination, and most disagreed by only 1 stage in classifications.^{7,8,36,37} Agreement may be better between examiners with greater experience than between those with differing levels of training.⁴⁰

Intraobserver Variability

In 2 studies,^{6,22} examiners placed thyroid size in categories of enlargement and repeated the examination on a separate occasion (Table 21-4). These data produced a κ from combined numbers of 0.59 (95% CI, 0.52-0.65) for placement in all categories of the rating scales used by the examiners. For simply determining the presence or absence of goiter, κ ranged from 0.47 to 0.79, with a κ from combined data of 0.65 (95% CI, 0.63-0.67), which is very good. Similar results were reported in a study of patients with various thyroid diseases, in which κ ranged from 0.54 to 0.74.³⁹ Intraobserver agreement was slightly better for the inspection component of the examination ($\kappa = 0.73$; 95% CI, 0.71-0.76) than for palpation ($\kappa = 0.65$; 95% CI, 0.63-0.67) (Table 21-5).

ACCURACY OF ESTIMATING THYROID SIZE

Three criterion standards have been used in assessing the accuracy of thyroid size determination: weight measured after surgical or postmortem removal, ultrasonographic assessment, and nuclear scintigraphy. Ultrasonographic assessments of thyroid weight correlate well with true gland weight as determined after excision ($r = 0.88-1.0$), although there is lack of agreement as to the best formula to use for estimating size.^{18,21,41} Nuclear scan determination is a little less reliable but acceptable ($r = 0.77-0.98$).^{9,42,43} Again, different formulas have been used to translate the scintigraphic profile to thyroid volume.^{9,42,43}

Combining data from 9 studies of detection of goiter by physical examination,^{12,17,18,21,44-48} the sensitivity from combined data was 0.70 (95% CI, 0.68-0.73) with a specificity of 0.82 (95% CI, 0.79-0.85) (Table 21-6). If a goiter was clinically detected, the positive likelihood ratio (LR+) of one being present was 3.8 (95% CI, 3.3-4.5). Conversely, if a goiter was not thought to be clinically present, the negative likelihood ratio was 0.37 (95% CI, 0.33-0.40). These likelihoods are comparable with or better than those for many other physical signs^{49,50} and were not affected by the presence of single or multiple nodules.⁴⁸ Experienced examiners were somewhat more accurate in their assessments than more junior colleagues.⁴⁸

Some authors have defined specific stages of thyroid enlargement according to the usual sequence of changes that occur as the thyroid gland increases in size. Because some of these staging classifications incorporate observations not normally used in simply estimating thyroid mass, they can significantly enhance the predictive abilities of the clinician

Table 21-5 Comparison of Intraobserver Precision for Inspection and Palpation

Reference	Agreement		κ	
	Inspection	Palpation	Inspection	Palpation
Hennessy ⁶	0.93	0.90	0.82	0.79
MacLennan et al ²²	0.95	0.82	0.18	0.47
Combined (95% CI)	0.94 (0.92-0.96)	0.85 (0.82-0.88)	0.73 (0.71-0.76)	0.65 (0.63-0.67)

Abbreviation: CI, confidence interval.

Table 21-6 Accuracy of the Clinical Assessment for the Presence of a Goiter^a

Reference	Sensitivity	Specificity	LR+	LR-
Silink and Reisenauer ^{17b}	0.64	0.89	5.8	0.40
Tannahill et al ^{21c}	0.93	0.75	3.7	0.09
Hegedus et al ^{44d}	0.43	1.00	Infinity	0.57
Hegedus et al ^{45d}	0.60	1.00	Infinity	0.40
Hegedus et al ^{46d}	0.77	0.80	3.9	0.29
Berghout et al ^{18e}	1.00	0.62	2.6	0.00
Perrild et al ^{47f}	0.64	1.00	Infinity	0.36
Hintze et al ^{12g}	0.66	0.74	2.5	0.46
Jarlov et al ^{48c}	0.80	0.80	4.0	0.25
Combined (95% CI)	0.70 (0.68-0.73)	0.82 (0.79-0.85)	3.8 (3.3-4.5)	0.37 (0.33-0.40)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

^aGoiter defined as thyroid gland size greater than 20 g, except in the study by Silink and Reisenauer,¹⁷ in which goiter was defined as gland size greater than 22 g, and in the study by Hintze et al,¹² in which male gland size was greater than 25 g and female gland size was greater than 18 g.

^bGraded degree of lateral prominence, goiter being any prominence, with criterion standard of autopsy weight.

^cDirectly estimated weight, with criterion standard of ultrasonography.

^dGoiter defined as visible or palpable gland, with criterion standard of ultrasonography.

^eGraded 5 stages of thyroid size according to palpability and visibility, with criterion standard of ultrasonography.

^fTwo observers had to agree on the presence of goiter, which was undefined, using ultrasonography as the criterion standard.

^gGraded 5 stages of thyroid size according to palpability, with criterion standard of ultrasonography.

(Table 21-7). In the combined data from 4 studies,^{19-21,48} when a clinician thought that a thyroid gland was of normal size, the LR+ of goiter being present was 0.15 (95% CI, 0.10-0.21). If classified as 1 to 2 times normal size, the LR+ was 1.9 (95% CI, 1.1-3.0), and for greater than 2 times normal, the LR+ was 25 (95% CI, 3.6-175).

Certain staging methods for thyroid enlargement can help clarify the true status of some of the patients with glands thought to be 1 to 2 times normal size after routine inspection and palpation.^{14,17} The amount of prominence of the thyroid on lateral inspection, for example, resulted in a high

likelihood of goiter if it was greater than 2 mm ([Table 21-8](#)). Of further utility was finding that a gland was not visible with the neck extended, a result that effectively ruled out a goiter.

Table 21-7 Accuracy in Assessing Grades of Thyroid Gland Weight

Reference	LR+
Normal Thyroid Size, 0-20 g	
Williams et al ^{19a}	0.00
Smith and Wilson ^{20a}	0.00
Tannahill et al ^{21b}	0.10
Jarlov et al ^{48b}	0.26
Combined (95% CI)	0.15 (0.10-0.21)
Thyroid Size 1-2 Times Normal, 20-40 g	
Williams et al ^{19a}	Infinity
Smith and Wilson ^{20a}	0.32
Tannahill et al ^{21b}	2.2
Jarlov et al ^{48b}	2.6
Combined (95% CI)	1.9 (1.1-3.0)
Thyroid Size > 2 Times Normal, >40 g	
Williams et al ^{19a}	Infinity
Smith and Wilson ^{20a}	Infinity
Tannahill et al ^{21b}	Infinity
Jarlov et al ^{48b}	13
Combined (95% CI)	25 (3.6-175)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio.

^aDirectly estimated thyroid weight, with postsurgical weight as the criterion standard.

^bDirectly estimated thyroid weight, with ultrasonography as the criterion standard.

Table 21-8 Accuracy in Assessing Thyroid Size by Categories

Stage, Size	LR+ (95% CI)
Method of Silink and Reisenauer ^{17a}	
0, not visible	0.41 (0.34-0.49)
1, 0-2 mm	3.4 (1.8-6.3)
2, 2-10 mm	Infinity
3, >10 mm	Infinity
Method of Berghout et al ^{48b}	
OA	0.00
OB	0.00
1	1.00 (0.42-2.4)
2	3.9 (1.8-8.2)
3	Infinity

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio.

^aGraded degree of lateral prominence, with goiter being any prominence, using autopsy weight as a criterion standard.

^bGraded stages 0-3 according to palpability and visibility, with goiter being 1-3, using ultrasonography as a criterion standard: OA indicates lobes smaller than the size of the thumb terminal phalanx, thyroid not visible with neck extended; OB, lobes bigger than the size of the thumb terminal phalanx, thyroid not visible with neck extended; 1, easily palpable, visible with neck extended; 2, visible with neck in normal position; and 3, easily visible.

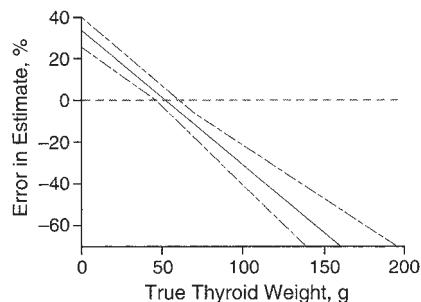


Figure 21-3 Error in Estimating Thyroid Mass

Error in estimating thyroid mass can be described by the following formula: percentage of error = $(-0.656 \times \text{mass}) + 34.8$, where thyroid mass is expressed in grams ($r = 0.41$; $P < .001$). The 95% confidence interval is indicated by the broken lines.

BIAS IN ESTIMATING THYROID SIZE

When the results from 4 studies^{19-21,48} estimating thyroid gland weights were combined, a regression line was produced describing the bias in gland size determination ([Figure 21-3](#)). This clearly shows that sizes of smaller glands are routinely overestimated, whereas those of larger glands are underestimated. The size at which this crossover occurs corresponds to about 2 times normal size. The practical application of this finding is that glands in the 1- to 2-times-normal-size category fall in the range in which size is typically overestimated.

THE BOTTOM LINE

To determine whether a goiter is present, follow these steps:

1. Examine the thyroid gland by inspection and palpation.
2. Categorize thyroid size as normal or goiter. Subcategorize goiter as small goiter (1-2 times normal) or large goiter (greater than 2 times normal).
3. If you placed the thyroid in the small-goiter category, consider whether you overestimated the size; determine whether there is any prominence in the profile of the neck in the region of the thyroid when viewed laterally (classify the prominence as ≥ 2 or > 2 mm), and determine whether the gland is not visible from the front with the neck extended.
4. Place your patient in one of the following categories: “goiter ruled out,” normal thyroid size or thyroid considered to be not visible with neck extended; “goiter ruled in,” large goiter present or lateral prominence greater than 2 mm; or “inconclusive,” all other findings.

Author Affiliations at the Time of the Original Publication

Division of Endocrinology and Metabolism, Department of Medicine, University of Alberta, Edmonton, Alberta, Canada

Acknowledgment

This work was supported by the Alberta Heritage Foundation for Medical Research.

REFERENCES

1. Eastham RD. Thyroid disorders. In: Eastham RD, ed. *A Pocket Guide to Differential Diagnosis*. 2nd ed. Littleton, MA: John Wright-PSG Inc; 1985:220-221.
2. Gurney C, Owen SG, Hall R, Roth M, Harper M, Smart GA. Newcastle thyrotoxicosis index. *Lancet*. 1970;2(7686):1275-1278.
3. Winsa B, Dahlberg PA, Jansson R, Agren H, Karlsson FA. Factors influencing the outcome of thyrostatic drug therapy in Graves' disease. *Acta Endocrinol*. 1990;122(6):722-728.
4. Ladenson PW. Diagnosis of thyrotoxicosis. In: Braverman LE, Utiger RD, eds. *Werner and Ingbar's The Thyroid: A Fundamental and Clinical Text*. 6th ed. Philadelphia, PA: JB Lippincott; 1991:901.
5. Perez C, Scrimshaw NS, Munoz JA. Classification of goitre and technique of endemic goitre surveys. *Bull World Health Organ*. 1958;18(1-2):217-232.
6. Hennessy WB. Goitre prophylaxis in New Guinea with intramuscular injections of iodized oil. *Med J Aust*. 1964;1:505-512.
7. Trowbridge FL, Matovinovic J, McLaren GD, Nichaman MZ. Iodine and goiter in children. *Pediatrics*. 1975;56(1):82-90.
8. Kilpatrick R, Milne JS, Rushbrooke M, Wilson ESB. A survey of thyroid enlargement in two general practices in Great Britain. *BMJ*. 1963;1(5322):29-34.
9. Malamos B, Vagenakis AG, Pandros PG, Papapetrou PD, Sfontouris J, Koutras DA. Comparison of scanning and palpation in the assessment of the weight of the thyroid gland. *Endokrinologie*. 1970;56(2):232-240.
10. Jaffe RH. The variation in weight of the thyroid gland and the frequency of its abnormal enlargement in the region of Chicago. *Arch Pathol*. 1930;10:887-894.
11. Hull OH. Critical analysis of two hundred twenty-one thyroid glands. *Arch Pathol*. 1955;59(3):291-311.
12. Hintze G, Windeler J, Baumert J, Stein H, Kobberling J. Thyroid volume and goitre prevalence in the elderly as determined by ultrasound and their relationships to laboratory indices. *Acta Endocrinol*. 1991;124(1):12-18.
13. Gutekunst R, Smolarek H, Hasenpusch U, et al. Goitre epidemiology: thyroid volume, iodine excretion, thyroglobulin, and thyrotropin in Germany and Sweden. *Acta Endocrinol*. 1986;112(4):494-501.
14. Berghout A, Wiersinga WM, Smits NJ, Touber JL. Determinants of thyroid volume as measured by ultrasonography in healthy adults in a non-iodine deficient area. *Clin Endocrinol*. 1987;26(3):273-280.
15. Tunbridge WMG, Evered DC, Hall R, et al. The spectrum of thyroid disease in a community: the Whickam survey. *Clin Endocrinol*. 1977;7(6):481-493.
16. Crooks J, Tulloch MI, Turnbull AC, Davidsson D, Skulason T, Snaedal G. Comparative incidence of goitre in pregnancy in Iceland and Scotland. *Lancet*. 1967;2(7517):625-627.
17. Silink K, Reisenauer R. Geographical spread of endemic goitre and problems of its mapping. In: Silink K, Cerny K, eds. *Endemic Goiter and Allied Diseases*. Bratislava, Slovakia: Slovak Academy of Science; 1966:33-47.
18. Berghout A, Wiersinga WM, Smits NJ, Touber JL. The value of thyroid volume measured by ultrasonography in the diagnosis of goitre. *Clin Endocrinol*. 1988;28(4):409-414.
19. Williams RH, Towery BT, Jaffe H. Radioiodotherapy. *Am J Med*. 1949;7(6):702-717.
20. Smith RN, Wilson GM. Clinical trial of different doses of ^{131}I in treatment of thyrotoxicosis. *BMJ*. 1967;1(5533):129-132.
21. Tannahill AJ, Hooper MJ, England M, Ferriss JB, Wilson GM. Measurement of thyroid size by ultrasound, palpation, and scintiscan. *Clin Endocrinol*. 1978;8(6):483-486.
22. MacLennan R, Gaitan E, Miller MC. Observer variation in grading and measuring the thyroid in epidemiological surveys. In: Stanbury JB, ed. *Endemic Goiter*. Washington, DC: Pan American Health Organization; 1969:67-77. Publication 193.
23. Edwards CRW. The general examination and external features of disease. In: Munro J, Edwards CRW, eds. *MacLeod's Clinical Examination*. 8th ed. New York, NY: Churchill Livingstone; 1990:66.
24. Daniels GH. Physical examination of the thyroid gland. In: Braverman LE, Utiger RD, eds. *Werner and Ingbar's The Thyroid: A Fundamental and Clinical Text*. 6th ed. Philadelphia, PA: JB Lippincott; 1991:572-577.
25. Greenspan FS. Thyroid gland. In: Greenspan FS, Forsham PH, eds. *Basic and Clinical Endocrinology*. 3rd ed. Englewood Cliffs, NJ: Prentice-Hall International Inc; 1991:213-215.
26. Hollingsworth DR, Hamilton HB, Tamagaki H, Beebe GW. Thyroid disease in Hiroshima, Japan. *Medicine (Baltimore)*. 1963;42:47-71.
27. Shapira JD. *The Art and Science of Bedside Diagnosis*. Baltimore, MD: Urban & Schwarzenberg; 1990:236.
28. Kearns JE. *Treatment of Hyperthyroidism With Radioactive Iodine: A Twenty-Year Review*. Springfield, IL: Charles C Thomas Publisher; 1967:16-17.
29. Siminoski K. Differential movement during swallowing as an aid in the detection of thyroid pseudonodules. *Head Neck*. 1994;16(1):21-24.
30. Cook IJ, Dodds WJ, Dantos RO, et al. Opening mechanisms of the human upper esophageal sphincter. *Am J Physiol*. 1989;257(5 pt 1):G748-G759.
31. Gwinup G, Morton ME. The high lying thyroid: a cause of pseudogoitre. *J Clin Endocrinol Metab*. 1975;40(1):37-42.
32. Mercer RD. Pseudo-goiter: the Modigliani syndrome. *Cleve Clin Q*. 1975;42(4):319-236.
33. Burch WM. *Case Studies in Endocrinology*. Baltimore, MD: Williams & Wilkins; 1987:52-53.
34. Clain A. The thyroid gland. In: *Hamilton Bailey's Demonstrations of Physical Signs in Clinical Surgery*. 17th ed. Bristol, England: IOP Publishing; 1986:155-156.
35. Katlick MR, Wang C, Grillo HC. Substernal goiter. *Ann Thorac Surg*. 1986;39:391-399.
36. Trotter WR, Cochrane AL, Benjamin IT, Miali WE, Exley D. A goitre survey in the vale of Glamorgan. *Br J Prev Soc Med*. 1962;16:16-21.
37. Dingle PR, Ferguson A, Horn DB, Tubmen J, Hall R. The incidence of thyroglobulin antibodies and thyroid enlargement in a general practice in Northeast England. *Clin Exp Immunol*. 1966;1(3):277-284.
38. Sackett DL. A primer on the precision and accuracy of the clinical examination. *JAMA*. 1992;267(19):2638-2644.
39. Brander A, Viikinkoski P, Tuuheja J, Voutilainen L, Kivisaari L. Clinical versus ultrasound examination of the thyroid gland in common clinical practice. *J Clin Ultrasound*. 1992;20(1):37-42.
40. Jarlov AE, Hegedus L, Gjorup T, Hansen JM. Observer variation in the clinical assessment of the thyroid gland. *J Intern Med*. 1991;229(2):159-161.
41. Hegedus L. Thyroid size determined by ultrasound. *Dan Med Bull*. 1990;37(3):249-263.
42. Brown MC, Spencer R. Thyroid volume estimated by use of ultrasound in addition to scintigraphy. *Acta Radiol*. 1978;17(4):337-341.
43. Mandart G, Erbsmann F. Estimation of thyroid weight by scintigraphy. *Int J Nucl Med Biol*. 1975;2(4):185-188.
44. Hegedus L, Karstrup S, Veiergang D, Jacobsen B, Skosted L, Feldt-Rasmussen U. High frequency of goitre in cigarette smokers. *Clin Endocrinol*. 1985;22(3):287-292.
45. Hegedus L, Hansen JM, Luhdorf K, Perrild H, Feldt-Rasmussen U, Kampmann JP. Increased frequency of goitre in epileptic patients on long-term phenytoin or carbamazepine treatment. *Clin Endocrinol*. 1985;23(4):423-429.
46. Hegedus L, Molholm Hansen JE, Veiergang D, Karstrup S. Thyroid size and goitre frequency in hyperthyroidism. *Dan Med Bull*. 1987;34(2):121-123.
47. Perrild H, Hegedus L, Bastrup PC, Kayser L, Kastberg S. Thyroid function and ultrasonically determined thyroid size in patients receiving long-term lithium treatment. *Am J Psychiatry*. 1990;147(11):1518-1521.
48. Jarlov AE, Hegedus L, Gjorup T, Hansen JEM. Accuracy of the clinical assessment of thyroid size. *Dan Med Bull*. 1991;38(1):87-89.
49. Tamayo SG, Rickman LS, Mathews WC, et al. Examiner dependence on physical diagnostic tests for the detection of splenomegaly. *J Gen Intern Med*. 1993;8(2):69-75.
50. Williams JW, Simel DL, Roberts L, Samsa GP. Clinical evaluation for sinusitis: making the diagnosis by history and physical examination. *Ann Intern Med*. 1992;117(9):705-710.

This page intentionally left blank

Prepared by David L. Simel, MD, MHS

Reviewed by Adi Cohen, MD

CLINICAL SCENARIO

A 34-year-old woman had a child about 14 months ago. She had been breast-feeding her newborn but stopped about 2 months before her routine visit with you. She complains that her weight has not gone back to baseline and that her skirts are tight at the waist and her blouses are tight at the neck. Does she simply need to lose weight, or could she have a goiter?

UPDATED SUMMARY ON GOITERS

Original Review

Siminoski K. Does this patient have a goiter? *JAMA*. 1995; 273(10):813-817.

UPDATED LITERATURE SEARCH

Our updated literature search used the parent search strategy for The Rational Clinical Examination series, combined with the subject headings “exp Goiter,” “limited to diagnosis,” “radionuclide imaging,” “epidemiology,” and “ultrasound studies,” published in English from 1994 to 2004. We also crossed the clinical subject headings with “meta-analysis,” “ROC curve,” and the textword “systematic review” in both MEDLINE and the Cochrane databases. The results yielded 135 titles, for which we reviewed the titles and abstracts; 10 were selected for additional review. Two additional articles were selected for review from the references. We included articles that allowed us to calculate the sensitivity and specificity, or the observer variability, of the clinical examination for a goiter.

IMPROVEMENTS IN THE DATA PRESENTED IN THE ORIGINAL PUBLICATION

The actual techniques for palpating the thyroid are described well in the original publication. However, 3 methods for assessing thyroid size from palpation were presented: estimates of thyroid volume (in grams), lateral prominence of the thyroid (in millimeters), and a 5-level ordinal assessment based on palpability and visibility. The

World Health Organization (WHO) proposed a simplified classification for the presence of a goiter: an individual has a goiter when each lateral lobe has a volume greater than the individual's terminal phalanx of the thumb.¹ A grade 1 goiter will be palpably enlarged but not visible when the neck is in a normal position. A grade 2 goiter will be palpably enlarged and visible with the neck in the normal position. Most of the work on establishing these criteria comes from epidemiologic studies of endemic iodine deficiency that used children as the study subjects. The epidemiologic studies use examiners with considerable thyroid examination experience.

The effect of changing the threshold for the clinical screening test (palpation) changes the performance of the test. The interobserver variability when performed by experienced examiners is acceptable with both the 1960 and 1994 criteria. It is important to compare the size of the thyroid to the thumb because the case definition is not “any” palpable thyroid but one that is larger than the distal thumb. One study in a high-prevalence area found that defining a palpable thyroid as enlarged, without comparing the size to the thumb, increased the clinical goiter rate by 20%.²

CHANGES IN THE REFERENCE STANDARD

The reference standard for thyroid enlargement remains ultrasonography. A goiter is defined as a thyroid gland of increased volume. However, the appropriate threshold for identifying the patient as having enlargement vs not having enlargement is evolving. WHO recognizes endemic iodine deficiency as a global health problem. The prevalence rate of goiters in school-age children defines regions as having endemic iodine deficiency vs normal iodine status. The definitions of normality for children may be different from those of adults because thyroid volume depends on body surface area (it also varies by sex). Areas of endemic iodine deficiency may be severely affected by malnourishment, and this in turn affects the size of thyroid glands.

The older 1960 WHO standard for thyromegaly required that a lobe of the thyroid have greater volume than the terminal phalanx of the child's thumb. These criteria, established

before ultrasonography, defined endemic iodine deficiency as a population with a greater than 10% prevalence of goiter. Palpation was the only method for assessing thyroid volume, and the use of the child's thumb as a comparative standard would seemingly account for both the child's sex and body surface area. In 1994, a newer threshold was proposed for epidemiologic research that used ultrasonography and normative thyroid volume adjusted for body surface area. The newer threshold decreased the prevalence level to more than 5% to define iodine deficiency areas but also simplified the clinical criteria for a goiter. A key question is whether a universal normative standard should be used for thyroid volume (eg, a universal threshold volume above which defines a goiter, or above a percentile for the universe of patients) or whether local reference standards should be established (eg, thresholds developed within a defined geographic region).³

RESULTS OF LITERATURE REVIEW

When palpating the thyroid, compare the results of each lobe to the subject's distal thumb. A thyroid with both lobes larger than the patient's distal thumb is considered palpably enlarged (see Table 21-9).

Table 21-9 Likelihood Ratios for a Palpable Thyroid Gland Indicating a Goiter

Finding	LR+ (95% CI)	LR- (95% CI)
Palpable thyroid, children (1994 criteria) ⁴	3.0 (2.5-3.5)	0.30 (0.24-0.37)
Palpable thyroid, pregnancy (1994 criteria) ⁵	4.7 (3.6-6.0)	0.08 (0.02-0.27)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

EVIDENCE FROM GUIDELINES

The US Preventive Health Services Task Force evaluated the role of thyroid-stimulating hormone (TSH) screening

in healthy adults and observed that men have a lower prevalence of unrecognized and unsuspected thyroid disease compared with women.⁶ High risk patients for thyroid disorder include the elderly, postpartum women, those with high levels of radiation exposure (>20 mGy), and patients with Down syndrome. However, the task force concluded that the data are inconclusive for recommending TSH screening. The task force does not address clinical screening with palpation. Health Canada guidelines came to similar conclusions as the US Health Services Task Force.⁷

CLINICAL SCENARIO—RESOLUTION

Returning to prenatal weight is a postpartum problem for many women. This patient has an unusual complaint of clothing feeling tight around the neck, so you feel obligated to palpate for a thyroid. Despite the enlargement, many patients do not recognize that they have a goiter. Goiters are more common in women, especially during pregnancy and in lactating mothers. When you palpate her thyroid, you need to use proper technique and make sure that any palpable thyroid tissue moves upward when she swallows. If you feel thyroid tissue, decide whether the volume of the palpable tissue in both lobes is greater than the volume of her distal thumb. Although this approach of assessing volume has been validated in children and not in adults, inexperienced examiners may have difficulty deciding whether the thyroid volume is normal compared to endocrinologists who assess the size compared with a normal gland (eg, 1.5 times normal or 2 times normal). If you are uncertain whether the gland is normal, ultrasonography would confirm the presence or absence of a goiter. You should also assess more fully for signs and symptoms of thyroid dysfunction. Although the most common cause for inability to lose weight postpartum may be lack of exercise, a sensitive TSH assay would be required to make sure she does not have hypothyroidism.

GOITER—MAKE THE DIAGNOSIS

PRIOR PROBABILITY

The prior probability of a goiter is affected by many variables, including the patient's body surface area, sex, and regional variations associated with the endemic iodine deficiency. Two recent European studies of thyroid volume among community samples of healthy adults give us insight into the prevalence of goiter in the non-iodine-deficient area: 4% of patients in Spain (95% confidence interval [CI], 3%-6%)⁸ and 10% of patients in France (95% CI, 9%-11%)⁹ had palpable goiters. Unfortunately, the thyroid volume was not confirmed for patients with palpable goiters. Nonetheless, we can make some inferences that give us good starting points. The WHO defines an iodine-deficient area by the prevalence of goiter in school-aged children. According to normative population values, children who live in a non-iodine-deficient area should have a goiter prevalence of less than 5%.¹ Adults might have palpable thyroid glands for reasons other than iodine deficiency, so prevalence values slightly higher make sense. A starting point of 5% to 10% for healthy adults makes sense for the prior probability of a palpable thyroid.

POPULATION FOR WHOM A GOITER DISEASE SHOULD BE CONSIDERED

- Symptoms of hyperthyroidism or hypothyroidism
- Children, especially those in endemic iodine deficiency locales
- Pregnant and lactating women
- Elderly patients
- Patients with excessive radiation exposure
- Patients with Down syndrome

DETECTING THE LIKELIHOOD OF A GOITER

Because examining children is different from examining pregnant women for thyroid disease, we cannot combine the data (see Table 21-10). The techniques for examination, however, are similar. We have no data for the results of thyroid palpation in non-pregnant adults because epidemiologic studies of normal adults' thyroid volume exclude those with palpable enlargement.

Table 21-10 Likelihood Ratios for a Palpable Thyroid Gland Indicating a Goiter

Palpable Thyroid With Both Lobes > the Volume of the Subject's Distal Thumb (1994 criteria) vs Not Palpable	LR+ (95% CI)	LR- (95% CI)
Children	3.0 (2.5-3.5)	0.30 (0.24-0.37)
Pregnancy	4.7 (3.6-6.0)	0.08 (0.02-0.27)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

Palpating thyroid tissue in both lobes of a volume greater than the volume of the patient's distal thumb phalanx increases the likelihood of a goiter, but there will be false-positive results.

REFERENCE STANDARD TESTS

Ultrasonography.

In epidemiologic research, urinary iodine studies are evaluated along with thyroid palpation.

REFERENCES FOR THE UPDATE

1. World Health Organization. *Assessment of Iodine Deficiency Disorders and Monitoring Their Elimination*. 3rd ed. Geneva, Switzerland: World Health Organization; 2007. Document WHO/NHD/01.1.35-37. http://whqlibdoc.who.int/publications/2007/9789241595827_eng.pdf. Accessed April 27, 2008.
2. Peterson S, Sanga A, Eklöf H, et al. Classification of thyroid size by palpation and ultrasonography in field surveys. *Lancet*. 2000;355(9198):106-110.^a
3. Delange F. What do we call a goiter? *Eur J Endocrinol*. 1999;140(6):486-488.
4. Zimmermann M, Saad A, Hess S, Torresani, Chaouki N. Thyroid ultrasound compared with World Health Organization 1960 and 1994 palpation criteria for determination of goiter prevalence in regions of mild and severe iodine deficiency. *Eur J Endocrinol*. 2000;143(6):727-731.^a
5. Castañeda R, Lechuga D, Ramos RI, Magos C, Orozco M, Martínez H. Endemic goiter in pregnant women: utility of the simplified classification of thyroid size by palpation and urinary iodine as screening tests. *BJOG*. 2002;109(12):1366-1372.^a
6. US Preventive Services Task Force. *Screening Thyroid Disease: Recommendation Statement*. Rockville, MD: Agency for Healthcare Research and Quality; 2004. <http://www.ahrq.gov/clinic/3rduspstf/thyroid/thyrrs.htm>. Accessed April 27, 2008.
7. Beaulieu MD. Screening for thyroid disorders and thyroid cancer in asymptomatic adults. In: Canadian Task Force on the Periodic Health Examination, ed. *Canadian Guide to Clinical Preventive Health Care*. Ottawa, Ontario, Canada: Health Canada; 1994:612-618.
8. Gómez JM, Maravall J, Gómez N, Gumà A, Soler J. Determinants of thyroid volume as measured by ultrasonography in healthy adults randomly selected. *Clin Endocrinol*. 2000;53(5):629-634.
9. Barrère X, Valeix P, Preziosi P, et al. Determinants of thyroid volume in healthy French adults participating in the SU.VI.MAX cohort. *Clin Endocrinol*. 2000;52(3):273-278.

^aFor the Evidence to Support the Update for this topic, see <http://www.JAMAEvidence.com>.

This page intentionally left blank

EVIDENCE TO SUPPORT THE UPDATE:

21

Goiter

TITLE Endemic Goiter in Pregnant Women: Utility of the Simplified Classification of Thyroid Size by Palpation and Urinary Iodine as Screening Tests.

AUTHORS Castañeda R, Lechuga D, Ramos RI, Magos C, Orozco M, Martínez H.

CITATION BJOG. 2002;109(12):1366-1372.

QUESTION Do the simplified World Health Organization (WHO) criteria for goiter work well for pregnant women?

DESIGN Prospective, cross-sectional survey of patients who underwent independent clinical examinations and ultrasonography.

SETTING Three communities in Mexico. One region had endemic iodine deficiency, one had a low prevalence of goiter, and one was an urban area not expected to have a high prevalence of iodine deficiency.

PATIENTS Pregnant women who showed up for delivery in each of the 3 referral hospitals for the region.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

Thyroid size by the WHO 2001 criteria: grade 0, normal; grade 1, both lobes larger than the distal phalanx of the thumb, but the gland is not visible; grade 2, both lobes palpably enlarged but also visible. The examination techniques are well described and the examiners had their reliability confirmed.

Thyroid size was confirmed by ultrasonography. Patients also provided urine samples for urinary iodine.

MAIN OUTCOME MEASURE

Thyroid size. Values below the 90th percentile for the regions were considered not enlarged.

Table 21-11 Likelihood Ratio of a Palpable Thyroid for Thyromegaly

Test	Sensitivity	Specificity	LR+ (95% CI)	LR- (95% CI)
Palpable thyroid	0.94	0.80	4.7 (3.6-6.0)	0.08 (0.02-0.27)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

MAIN RESULTS

The 2 endocrinologist examiners had a κ of 0.70 for their agreement on the scoring scheme. The criteria have good discriminative properties for identifying patients with goiter (see Table 21-11).

CONCLUSIONS

LEVEL OF EVIDENCE Level 1.

STRENGTHS Study population included sample of low-risk and higher-risk patients. The examiners had their reliability assessed. The examination techniques are well described.

LIMITATIONS Examination done among pregnant patients only. The reference standard was a threshold that some might consider too low; a value greater than the 90th percentile was considered as goitrogenous.

Palpating pregnant women's thyroid glands may be easier than palpating the thyroid gland of nongravid subjects. However, it is possible that the simpler grading scheme and training of the examiners led to excellent reliability. The low negative likelihood ratio is impressive, suggesting that the finding of a nonpalpable gland during pregnancy rules out thyromegaly. Of course, patients can have substernal goiters, so we know that there will be some false-negative results (but not many).

Because the lower threshold for defining a goiter was used (the 1960 WHO criteria of 10th percentile rather than the currently recommended 5th percentile), we would expect the specificity and the positive likelihood ratio to be worse. However, compared with the results for using a 5-level scheme, the results are promising.

Reviewed by David L. Simel, MD, MHS

TITLE Classification of Thyroid Size by Palpation and Ultrasonography in Field Surveys.

AUTHORS Peterson S, Snaga A, Eklöf H, et al.

CITATION *Lancet.* 2000;355(9198):106-110.

QUESTION What are the effects on observer variability of the World Health Organization 1994 palpation system vs the 1960 system?

DESIGN Cross-sectional sample, convenience sample. Independence of examiners (3) and radiologist not specified.

SETTING Area of high prevalence of goiter from endemic iodine deficiency in Tanzania.

PATIENTS Schoolchildren.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

Three examiners examined each child. One of the examiners was described as “experienced,” 1 was an experienced paramedic, and 1 was an inexperienced “expatriate physician.” Each examiner evaluated each child in the morning and then again in the afternoon. Although they were not given their morning results, the examiners were not blinded to the child. The ultrasonographic testing for each child was also repeated.

MAIN OUTCOME MEASURES

Interobserver and intraobserver variability.

MAIN RESULTS

Seventy-five percent of the children had goiter by the reference standard ultrasonogram.

The inexperienced physician had a low intraobserver variability with both the 1960 and 1994 criteria (κ of 0.36 and 0.44, respectively). The intraobserver variability for the experienced examiners was similar for both criteria (κ of 0.57-0.58 for the 1960 criteria and 0.53-0.60 for the 1994 criteria). The performance of the inexperienced observer improved over time (κ of 0.26 during the first 3 days of the study compared with 0.56 for the last 3 days, using the 1960 criteria).

The ultrasonographer had an intraobserver variability that resulted in a reclassification of 14% of patients from morning to afternoon examinations (κ of 0.63).

CONCLUSIONS

LEVEL OF EVIDENCE Level 3.

STRENGTHS Three examiners of various levels of experience. Intraobserver variability was assessed.

LIMITATIONS The ultrasonographer had poor precision, making the quality of the reference standard doubtful.¹ There was a lack of independence in the examination by each clinician. The study was done in an exceedingly-high-prevalence area.

We include a review of this article for several reasons. First, it seems clear that the intraobserver variability is better for experienced examiners. Second, this study demonstrated that the inexperienced observer’s precision improved during the course of the study, which allows us to infer that practice is helpful. Third, it is important to apply the clinical criteria as they are currently specified. “Any” palpable enlargement does not qualify the patient has having a goiter, because each lobe must be of greater volume than the distal phalanx of the thumb. Finally, the reliability of the ultrasonographic reference standard was probably too low. For this reason, we do not show the sensitivity and specificity of the individual examiners.

Reviewed by David L. Simel, MD, MHS

REFERENCE FOR THE EVIDENCE

- Zimmermann M. Assessing goiter prevalence. *Lancet.* 2000;355(9219):1995-1996; author reply 1996-1997.

TITLE Thyroid Ultrasound Compared With World Health Organization 1960 and 1994 Palpation Criteria for Determination of Goiter Prevalence in Regions of Mild and Severe Iodine Deficiency.

AUTHORS Zimmermann M, Saad A, Hess S, Torresani T, Chaouki N.

CITATION *Eur J Endocrinol.* 2000;143(6):727-731.

QUESTION How do the 1960 World Health Organization (WHO) criteria for goiter compare with the simplified 1994 criteria?¹

DESIGN Prospective, independent, cross-sectional sample.

SETTING Two mountainous regions of Morocco. One was an area of WHO-defined mild endemic iodine deficiency disease; the other had goiter prevalence compatible with severe iodine deficiency.

PATIENTS Schoolchildren, 200 from each village (n = 400 total).

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

Two examiners evaluated each child, blinded to each other’s result. An ultrasonographer evaluated each subject, blinded to the clinicians’ examinations. The examiners and ultra-

Table 21-12 WHO 1994 Criteria for Goiter

Grade 0	No palpable or visible goiter.
Grade 1	Palpable but not visible neck mass consistent with the thyroid when the neck is in the normal position. The gland moves up when the patient swallows.
Grade 2	A swelling in the neck that is visible when the neck is in the normal position and is consistent with thyroid when palpated.

sonographer were all experienced in goiter epidemiologic studies. The clinicians recorded their findings according to the WHO 1960 criteria for goiter and the 1994 criteria (**Table 21-12**). The WHO upper limit of the thyroid volume, adjusted by the subject's sex and body surface, was used as the reference standard.

MAIN OUTCOME MEASURES

Sensitivity, specificity, and κ values.

MAIN RESULTS

In the community with mild iodine deficiency, κ was 0.47 between examiners for the 1960 vs 1994 criteria and 0.53 for the 1994 criteria. In the severe iodine deficiency site, κ was 0.67 between examiners for both the 1960 and 1994 criteria.

In the high-prevalence village, the accuracy of the clinical examination was similar for the 1960 and 1994 criteria (see **Table 21-13**). In the low-prevalence village, the clinicians estimated a prevalence of 20% to 21% with the 1960 criteria and 25% to 26% with the 1994 criteria; however, the actual prevalence was 12%.

CONCLUSIONS

LEVEL OF EVIDENCE Level 1.

STRENGTHS Large sample size of a community-based population for whom it was reasonable to screen for thyroid disease. The study subjects were not enrolled because of a suspicion for disease.

LIMITATIONS The sampling frame is not specified and the study enrolled no adults. All the clinicians were experienced examiners, which limits generalizability.

Table 21-13 Likelihood Ratio for Thyroid Palpation for 2 Different Examiners, Based on the 1994 vs 1960 WHO Criteria

Test	Examiner	LR+ (95% CI)	LR- (95% CI)
WHO 1994	A	2.9 (2.3-3.6)	0.32 (0.24-0.43)
	B	3.1 (2.5-3.9)	0.27 (0.19-0.38)
	Combined	3.0 (2.5-3.5)	0.30 (0.24-0.37)
WHO 1960	A	3.3 (2.6-4.2)	0.34 (0.26-0.45)
	B	3.3 (2.6-4.2)	0.31 (0.23-0.42)
	Combined	3.3 (2.8-3.9)	0.33 (0.27-0.40)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

The 1994 revised criteria for goiter simplified the scale from 5 to 3 levels. Overall, the performance between the 2 criteria appears similar. With fewer choices, the 1994 criteria ought to be more reliable, especially when used by less experienced examiners. The newer criteria required only that the thyroid be palpable to be considered clinically enlarged. In areas of low prevalence, this would lead to overestimates of ultrasonographically proven thyromegaly. A 2001 revision of the criteria clarified that “palpable” meant an enlargement of both lobes to a volume greater than the distal phalanx of the subject’s thumb.

It is disappointing that the experienced thyroid examiners did not have higher diagnostic accuracy. Other studies have found that the clinical examination overestimates the volume of the thyroid in schoolchildren.² A partial explanation may relate to problems with using a worldwide WHO fixed cut point for ultrasonographic size rather than with using local reference values adjusted for sex and age.³

Reviewed by David L. Simel, MD, MHS

REFERENCES FOR THE EVIDENCE

- World Health Organization. *Indicators for Assessing Iodine Deficiency Disorders and Their Control Programmes: Report of a Joint WHO/UNICEF/ICCIDD Consultation, 3-5 November 1992 (WHO/NUT/93.1).* Geneva, Switzerland: World Health Organization; 1993. Review version, September 1993.
- Lisboa HRK, Gross JL, Orsolini A, Fuchs S. Clinical examination is not an accurate method of defining the presence of goiter in schoolchildren. *Clin Endocrinol.* 1996;45(4):471-475.
- Foo LC, Zulfiqar A, Fafikudin M, Fadzil MT, Asmah ASA. Local versus WHO/International Council for control of iodine deficiency disorders—recommended thyroid volume reference in the assessment of iodine deficiency disorders. *Eur J Endocrinol.* 1999;140(6):491-497.

This page intentionally left blank

CHAPTER 22

Does This Patient Have Hepatomegaly?

C. David Naylor, MD, DPhil, FRCPC

CLINICAL SCENARIO

The patient in your examining room is new to the practice. He is 52 years old, emigrated from Southeast Asia about 10 years ago, and has no specific complaints except fatigue. On examination you find little of note except that his liver edge is firm, is easily felt, and extends about 6 cm below the costal margin across much of the right upper quadrant. The span, by light percussion, is 17 or 18 cm. Should you be concerned? What does the research literature tell us about the meaning of these findings?

WHY IS THE CLINICAL EXAMINATION IMPORTANT?

Ideally, the clinical meaning of physical examination findings should be established in research studies that account for the overall context, including other signs and details from the medical history. This approach is difficult in liver disease because the physical manifestations of hepatic dysfunction are protean, and many multisystem diseases affect the liver. Our focus, therefore, is on physical examination of the liver itself. This means, however, that we implicitly depend on the clinician's ability to make a baseline estimate of the likelihood of liver disease according to the medical history or other physical findings.

Although many maneuvers recommended in liver examination are unproven, there is reasonable evidence that the presence or absence of hepatomegaly can be determined with moderate accuracy on physical examination. Descriptive studies suggest that other qualitative findings may help in clinical assessment of patients with possible liver disease. Liver examination, like most physical diagnosis maneuvers, is not dissimilar to a screening test; it may support or refute hypotheses generated by the medical history and generate further hypotheses itself, allowing more selective use of imaging techniques and laboratory tests as tools to confirm the suspected diagnoses.¹

TOPOGRAPHY

Situated intraperitoneally in the right upper quadrant, the liver seldom extends more than 5 to 6 cm across the midline into the left upper quadrant. The upper surface is convex and nestles under the diaphragm, typically at the level of the fifth or sixth anterior rib in quiet respiration. The lower surface tends to be concave, with the gallbladder in it. Although the fundus of the gallbladder may project below and anteriorly to the lower liver edge, it is not felt in healthy persons.

The bulk of the liver sits posteriorly, where it cannot be assessed from behind because of intervening retroperitoneal contents, ribs, and lumbar musculature. Anteriorly, the liver sits partly above the costal margin, with ribs and lung supervening, and partly below it. The portion extending below or inferior to the costal margin varies and typically runs parallel

to the costal margin. However, physicians working in modern imaging departments, like generations of surgeons and anatomists before them, can attest to the degree of variability in the shape of the organ, including the extent to which the lower edge parallels the costal margin and the degree of extension beyond the midline into the left upper quadrant ([Figure 22-1](#)). To some extent, the vertical liver span (ie, the linear distance from the top of the liver dome down to the lower edge) is a function of where in the right upper quadrant the liver edge is palpated or percussed (Figures 22-1, [22-2](#), and [22-3](#)). The falciform ligament joins the midanterior surface of the liver to the diaphragm and anterior abdominal wall. With respiration, diaphragmatic contraction drives the liver downward, and the anterior surface of the organ rotates slightly to the right. In quiet inspiration and expiration, the excursion is approximately 2 to 3 cm.

A SUGGESTED APPROACH TO LIVER EXAMINATION

We assume that, as part of the general abdominal examination, you have already inspected the abdomen, including the right upper quadrant, looking for obvious irregularities or deformities. Then, in adults without a history or physical findings suggestive of potential liver disease, palpate for the lower liver edge. Start with gentle pressure in the right lower quadrant; ask the patient to breathe in gently and slowly to bring the liver edge down to the examining fingertips. At each exhalation, move the fingers up about 2 cm. If the edge is not felt, no further examination is suggested.

If the edge is felt, confirm that you are palpating roughly in the middle of the right portion of the abdomen, that is, corresponding to the midthoracic line or so-called midclavicular line (MCL). Mark the lower edge. Then, in the same approximate plane, percuss down from about the level of the third rib, with the pleximeter finger (the finger that you strike with the per-

cussing finger) laid horizontally. Typical lung field resonance will be heard. Move one rib space at a time until the tone changes because of the interposition of the dome of the liver behind the air-filled lung. There will be a gradation with increasing dullness as you move caudally and the volume of the air-filled lung overlying the liver is diminished ([Figure 22-3](#)).

To confirm increased dullness, spread 2 or 3 pleximeter fingers over adjacent rib spaces and percuss quickly a number of times from greater to lesser resonance. If doubts persist, have the patient take a deeper breath and hold it; then percuss to confirm an unequivocal increase in resonance at that rib space. Determination of a level for the upper edge of liver dullness is sometimes helped by placing the middle finger over the likely level for initial tone change and laying the second and ring fingers on adjacent rib spaces. Again, percuss back and forth. The percussion tone over the top finger should be resonant; the lower finger, unequivocally dull; and the middle finger, resonance between that of the other fingers.

Try to ensure that the lower and upper borders are marked either in quiet respiration or, if deep breaths are taken, in the same phase of respiration.

In instances during which you have other evidence to suggest liver disease, but the liver edge was not palpable, attempt to locate the lower edge by gentle percussion in the right lower quadrant, following the plane of the MCL and again working from resonance to dullness. Tricks similar to these (eg, multiple pleximeter fingers and manipulating level of dullness with changes in depth of respiration) may help confirm the finding. If there is no definite tone change up to the costal margin—a not uncommon finding—end the attempt to define liver size.

Determination of vertical liver span in the MCL can be done in 2 ways. We recommend gentle percussion for locating the upper liver border and palpation or gentle percussion to locate the lower border. An alternative is to use firm per-

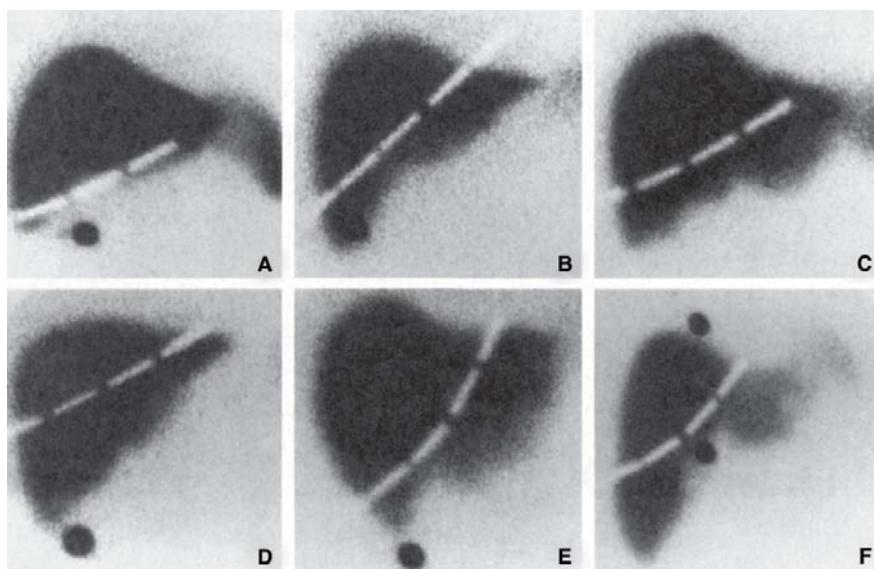


Figure 22-1 Radioisotope Scans of the Liver Showing Variability in Organ Shape

Note the costal margin markers as white broken lines and the other 2 dark point markers for research purposes; respiratory excursion blurs and expands the point markers, a limitation on the precision of any study done with reference to scintigraphic standards. A-D, Variation in alignment with the costal margin. E, F, Prominence of the left (caudate) and right (pyramidal) lobes, respectively.

cussion, deliberately ignoring whether or not the lower edge is palpable.

Liver size correlates with body size, and liver shape correlates with habitus. Liver span is greater in men than women and in tall vs short persons. However, as a rough guide, an MCL span of less than 12 to 13 cm with gentle percussion alone or gentle percussion combined with palpation makes hepatomegaly unlikely. Ranges of normal have been established for firm percussion (Table 22-1) but will vary among clinicians, depending on percussion techniques. Enlargement suggested by percussive span alone is weaker evidence for hepatomegaly than span based on palpation of the lower liver edge.

Apart perhaps from the situation of fulminant hepatic failure, observing reduction in liver span is of limited use because many other features of chronic liver failure will be present in situations in which reduction in parenchymal mass has occurred.

When the liver edge is palpable, tracing the edge and defining its characteristics qualitatively are recommended primarily in persons who are strongly suspected of having liver disease. Auscultation is seldom helpful. Once you have a high index of suspicion about liver disease, biochemical tests and biopsy are the main events; the more esoteric findings on physical examination become a sideshow for impressing referring physicians or trainees.

EVIDENTIARY BASIS FOR THE APPROACH

Inspection

Visualization of infracostal extension of the liver is occasionally possible when malnutrition or cachexia thin out the overlying tissues or when there is massive hepatomegaly. No studies, to our knowledge, describe the yield from inspection of the liver outline in the abdomen, but clear-cut abnormalities should at least be specific, thereby ruling in hepatomegaly and underlying disease.

Auscultation

Friction rubs may occur with primary and metastatic malignancies, after liver biopsies, with infective and inflammatory conditions, and with or without concomitant hepatomegaly. Rubs, although always abnormal, are rare and nonspecific; even with careful examination of patients with liver tumors, no more than 10% of patients have a rub.²⁻⁴

A detailed review of abdominal auscultation is provided by Sapira,⁵ including bruits and hums occurring in and around the right upper quadrant. Considerable time can be spent on auscultation, but there is no evidence that these findings are helpful in routine examination. Features reputed to help separate bruits of arterial and venous sources are described in Table 22-2. Venous hums occur in portal venous hypertension of any cause. The hum, a low-pitched murmur with systolic and diastolic components, arises from communication between the umbilical or paraumbilical veins and abdominal wall veins. The responses of venous hums to the Valsalva maneuver, splenic pressure, or ingestion of meals are inconsistent.^{6,7} Other causes of true continuous murmurs, such as arteriovenous fistula in the splanchnic circulation or hepatic

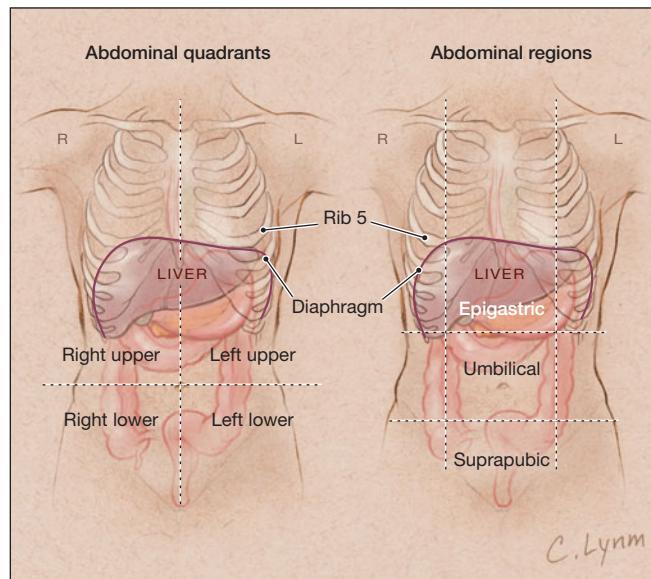


Figure 22-2 The Surface Anatomy of the Abdomen Can Be Divided Into Quadrants or Regions

The edge of the liver will typically be felt in the right upper quadrant.

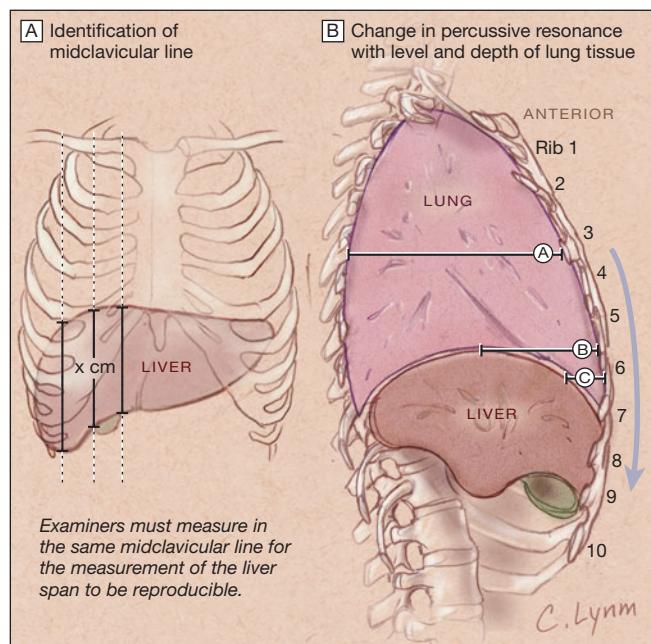


Figure 22-3 (A) Surface Landmarks for (B) Percussing the Liver Creates Resonance Dependent on the Underlying Structures

A, Variation in liver span according to the vertical plane of examination. Because there is variability in where clinicians determine the midclavicular line to be, the inevitable consequence is that liver span may also vary, even if multiple observers are perfectly accurate in measuring it. B, Percussive resonance varies with the thickness of interposed air-filled lung tissue. The percussion note changes with decreasing resonance caudally as less air-filled lung tissue is interposed between the liver and ribs. However, the site of a change from obvious resonance over the lung (point A) to less resonance (points B and C) may be difficult to judge.

Table 22-1 Normal Liver Span in the Midclavicular Line and the Midsternal Line, as Determined With Heavy Percussion Alone³⁷

Height, cm	MCL, cm		MSL, cm	
	Men	Women	Men	Women
150	8.25	6	6	4
157.5	9	6.75	6.5	4
165	9.75	7.5	7	5
172.5	10.25	8	7.5	...
180	11	8.75	8	...
187.5	11.75	9.5	8.5	...

Abbreviations: MCL, midclavicular line; MSL, midsternal line.

^aEllipses indicate data not available.**Table 22-2** Potential but Unproven Means to Differentiate Venous Hums and Arterial Bruits

Feature	Venous Hum	Arterial Bruit
Pitch	Lower	Higher
Volume	Soft	May be loud
Timing	Continuous	Systolic Systolic and diastolic
Systolic accentuation	Yes	Yes
Localized	No	Yes
Change with position	Yes	Sometimes ^a
Change with inspiration	Louder	May decrease
Stethoscope pressure	Diminishes	Unchanged

^aAlthough positional change in arterial bruits would not be expected in an arterial bruit caused by tumor vascularity, positional change may occur if a bruit is caused by pressure on the abdominal aorta from the enlarged left lobe of the liver.^{4,5,9}

hemangioma, are uncommon, and arterial bruits rarely have such lengthy diastolic spillover that they sound continuous.^{4,8}

Arterial bruits over the liver or in the epigastrium have been described with most liver tumors, as well as alcoholic hepatitis.^{2,4,8-11} However, among patients with liver disease in general (eg, a convenience sample of cirrhosis, alcoholic hepatitis, and malignancy), the prevalence of bruits has been reported at less than 3%.⁸ Both clinically⁹ and with phonoangiographic enhancement,⁸ the murmurs associated with alcoholic hepatitis and malignancy cannot be distinguished from one another, although the former resolve if and when the condition improves. The prevalence of clinically audible bruits in patients with confirmed liver cancer varies from 10%¹¹ to as much as 56%.¹² Kingston et al¹³ reported a diastolic component to most bruits heard in their patients with hepatoma. About 1% to 2% of unselected patients on a general medical service will have abdominal bruits of some kind,⁴ and the ability of clinicians to distinguish hepatic from other arterial bruits has never been assessed.

Auscultation over the liver should be considered only when medical history and other physical findings are suggestive of

hepatic disease; even then, the findings should be interpreted cautiously.

A PALPABLE LIVER EDGE: WHAT DOES IT MEAN?

Cirrhosis or infiltrative disorders increase the firmness of the liver edge and the likelihood of its being felt independent of effect on organ size.¹⁴ Among gastroenterologists, agreement on the presence of a palpable liver edge is about 50% greater than expected by chance alone.¹⁵ More interobserver disagreement would be expected in ordinary practice.

There is a paucity of data on the prevalence of palpable livers in the general population. One study¹⁶ has reported data on palpability of the liver among 1000 military personnel (717 men and 283 women) undergoing routine examination; 852 subjects were 40 years of age or younger. Palmer,¹⁶ the author and sole examiner, excluded any persons in whom liver disease was suspected or who were difficult to examine. In 57% of subjects, the liver was either not palpable in the right upper quadrant or felt just at the costal margin. An additional 28% descended only 1 to 2 cm below the costal margin. Findings were similar for both sexes. The proportion of palpable livers was inflated by 2 factors. First, all subjects were examined in deep, held inspiration. Second, as Palmer¹⁶ himself cautioned, “There is no question but that many of the potentially palpable livers would have been overlooked if this had not been a specially directed study.”

Ability to palpate the liver is not closely correlated with liver size in studies using reference standards such as scintigraphy or ultrasonography.¹⁷⁻²⁰ (Although many published studies use scintigraphy as a reference standard, it does have the drawback of motion artifact in conventional applications.) Patients undergoing liver scintiscan are preselected, and a high proportion of palpable livers might be expected. However, studies from nuclear medicine departments show that although the majority of patients scanned have some infracostal extension of the liver, less than half of these patients had palpable livers.^{14,18-21} In one study, Rosenfield et al²² chose 100 scintiscans at random and compared the findings with the clinical records. Among patients without definite evidence for liver disease in the medical records, mean scintigraphic vertical span in the right MCL was similar among those with palpable (12.9 cm) and nonpalpable organs (12.5 cm), as were the proportions in each category (45% vs 55%). Overall, the chance that a patient with a palpable liver also had liver disease was 63% (36 of 57 patients; 95% confidence interval [CI], 49%-76%), but the chances of a palpable liver meeting scintigraphic criteria for enlargement were only 46% (24/52 patients; 95% CI, 32%-61%). Studies^{14,22,23} on palpability and hepatomegaly are summarized in Table 22-3. This distinction between abnormal and enlarged livers is a recurrent problem because livers may be abnormal yet not enlarged.

What of the converse proposition, that is, that a nonpalpable liver is not enlarged? Because normal livers usually extend below the costal margin yet may not be palpable, this proposition rests on an assumption that enlarged livers

will be diseased, abnormally hard, and therefore much more easily felt. As summarized in Table 22-3, a nonpalpable liver does reduce the probability of hepatomegaly, even though a palpable organ has less than a 50% chance of being enlarged. These figures are influenced by the pooled prevalence of hepatomegaly, 23% in these studies. As a prevalence-free characteristic, we can report that the pooled likelihood ratio²⁴ (LR) for hepatomegaly, given a palpable liver (positive likelihood ratio [LR+]), is 2.5. The LR in the absence of palpable hepatomegaly (negative LR) for the presence of an enlarged liver detected by scanning is 0.45. However, there will likely be an evaluation bias in these figures as a result of preferential referral of patients with palpable livers for scintigraphy. This bias would arguably lead to a slight overestimate of sensitivity and still larger underestimate of specificity. If specificity were higher, the LR+ would be stronger. In any event, an LR approach is most useful if you know the previous odds of hepatomegaly for representative cohorts of patients with various diseases, a set of numbers that are currently unknown and should be the subject of research in the future.

In sum, a palpable liver is not necessarily enlarged or diseased but does increase the likelihood of hepatomegaly. The vertical liver span and overall clinical context must also be considered. Conversely, a nonpalpable liver edge does not rule out hepatomegaly but does reduce its likelihood. This is particularly relevant in those settings of low prior probability of liver disease, in which further examination is likely to have little yield if the liver cannot be felt.

WHAT ELSE CAN BE LEARNED FROM PALPATION?

Da Costa²⁵ wrote 93 years ago, “Tactile sense decides the questions of hepatic tenderness, pulsation, friction, and thrills, and determines the consistence and the contour of its anterior and lower surfaces.” However, there are few data on the reliability and accuracy of these qualitative judgments about liver edge characteristics.

A pulsatile liver edge is well documented in tricuspid valvular disease.²⁶⁻²⁸ Although this sign may be present clinically in the majority of cases,²⁹ no modern studies adequately document the frequency of the association and its relationship to differing degrees of tricuspid valvular dysfunction. Unequivocal pulsatile hepatomegaly is also reported in 35 of 55 consecutive patients (64%; 95% CI, 50%-76%) with confirmed constrictive pericarditis accumulated in 2 case series.^{30,31} The low false-negative rates give this sign some potential value in a setting in which constrictive pericarditis is already suspected. Unfortunately, as Osler³² observed more than a century ago, there is a need to distinguish between an expansile liver edge and transmitted aortic or right ventricular impulses that are commonly present. There are no data on examination maneuvers to make such a distinction, although inspiratory increase in the magnitude of the pulsation has been reported anecdotally with tricuspid insufficiency.²⁶ Detection of differential timing of hepatic pulsations has been described (eg, A vs V waves) but is rare and doubtless difficult to pinpoint.²⁷

Table 22-3 Probability of Hepatomegaly if a Liver Is Palpable or Not and Related Likelihood Ratios

Liver Palpability	Hepatomegaly		LR (95% CI) ^a
	Yes	No	
Peternel et al ^{14b}	12	12	LR+, 1.7 (1.0-2.8)
	4	15	LR-, 0.45 (0.18-1.1)
Rosenfield et al ^{22c}	24	28	LR+, 1.4 (1.0-2.1)
	13	35	LR-, 0.63 (0.39-1.0)
Walk ^{23d}	195	263	LR+, 2.6 (2.3-3.0)
	95	768	LR-, 0.44 (0.37-0.52)
Pooled	231	303	LR+, 2.5 (2.2-2.8)
	112	818	LR-, 0.45 (0.38-0.52)

Abbreviations: CI, confidence interval; LR, likelihood ratio; LR+, positive LR; LR-, negative LR.

^aCI values on LRs were determined using the method of Simel et al.²⁴

^bScintigraphic span of 16.5 cm or more; this reflects an arbitrary interpretation based on the bigger-than-usual span among clinically normal persons reported by Peternel et al.¹⁴

^cScintigraphic MCL span of 15.5 cm or more.

^dVolume greater than 1100 mL/m², in which a volume of 900 mL/m² usually signifies enlargement.

Palpation for an expansile liver edge should be limited to cases of suspected tricuspid valve disease or constrictive pericarditis.

The other qualitative characteristics are consistency, nodularity, and tenderness of a palpable liver edge. Among multiple expert observers examining variously alcoholic or jaundiced patients, κ statistics for chance-corrected agreement were 11% for abnormal consistency of a palpable liver edge¹⁵ and 26%¹⁵ or 29%³³ for presence of nodularity. Only agreement on tenderness of the liver edge was within a useful range, at 49%.³³

Palpation to describe the liver edge qualitatively or to detect isolated enlargement of the left (caudate) lobe of the liver should therefore be considered primarily if there is other evidence of organ disease or concern about liver tumor and even then is optional.

ASSESSING VERTICAL LIVER SPAN

Unequivocal reduction in liver size should be detectable in fulminant hepatic failure. However, no evidence was located to support the common belief that a substantial proportion of persons with chronic cirrhosis have detectably small livers by physical examination. The focus herein is accordingly on hepatomegaly.

Because half of all palpable livers are not enlarged, measurement of vertical liver span in some plane is required. The usual reference point is the MCL. However, unless care is taken in examination, the MCL can be “a wandering landmark,” with documented interobserver variation as much as 10 cm.³⁴ Variation in the MCL will inevitably lead to imprecision in

Table 22-4 Match of Clinically Measured Midclavicular Line Span and Imaged Span

Authors and Procedure	No. (%) of Patients With MCL Proportion Within 2 cm/No. of Total Patients	No. of Observers	Imaging Method
Sullivan et al ¹⁹			
Scratch test	15/36 (42)	1	Scintigraphy, MCL, not matched
Percussion alone	19/47 (40)		
Palpation alone (where applicable)	17/32 (53)		
Fuller et al ³⁶			
Palpation or percussion and scratch test	31/40 (78)	3 or 4	Ultrasonographic MCL matched ^a
Palpation or percussion alone	16/36 (44)		
Petersen et al ¹⁴			
Percussion and palpation	18/43 (42)	2 or 3 ^b	Scintiscan, MCL not matched
Naylor et al ³⁵			
Percussion or palpation			
Observer 1	20/39 (51)	1	Scintiscan, MCL matched
Observer 2	13/34 (38)	1	Scintiscan, MCL matched

Abbreviation: MCL, midclavicular line.

^aSpan from costal margin only.

^bMean MCL span used when observers' results differed.

liver span assessments (Figure 22-2). Vertical span could be an accurate predictor of liver mass only if the organ were more or less cuboid rather than irregular.

Palpation should, in theory, be the most reliable and accurate method of locating the lower border of the liver to measure organ span. Two studies^{12,33} report specialists' ability to agree on distance from the costal margin to a palpable liver edge, an approach that overstates accuracy by eliminating the largest source of error location of the upper border of the liver.^{14,19,35} Meyhoff et al¹² further controlled interobserver disagreement by having all measurements made at a predetermined MCL. Mean maximum interobserver difference of distance from the costal margin was 6.1 cm (SD, 2.7 cm) in the MCL. Intraobserver variation was smaller, with differences not greater than 2 cm in 60% to 80% of MCL measurements. There was no clear relationship to liver size, a finding that underscores the need to measure span from the upper border of the liver, not the costal margin. Theodossi et al³³ performed a similar experiment without marking the MCL. The intraclass correlation coefficient was 0.66, analogous to a weighted κ of more than 60%. However, agreement beyond chance on whether the liver was truly enlarged was only 30%.

What are the alternatives to localizing the lower liver edge by palpation? The scratch test is performed by placing the diaphragm of the stethoscope at the xiphisternum or over the

liver just above the costal margin in the MCL. Starting low in the abdomen, a finger is moved up the abdomen, scratching gently. The intensity becomes greatly enhanced once the finger is over the lower border of the liver.³⁶ The other major alternative is percussion.

Comparative studies are summarized in Table 22-4.^{19,36} Two caveats are in order. Both studies involved limited numbers of observers and patients. Furthermore, the overall accuracy in the report by Fuller et al³⁶ is greatly exaggerated on 2 scores. First, the ultrasonographic measurement was made in a plane defined by the observers. In actual practice, the MCL of the clinical observer varies from that of the scintigrapher or ultrasonographer,^{34,35} a situation that was applicable for the patients examined by Sullivan et al.¹⁹ Second, Fuller et al³⁶ took their measurements from the costal margin.

The scratch test may be a useful adjunct to percussion or palpation in locating the lower edge of the liver. However, more studies are needed before it can be recommended for routine use.

Also shown in Table 22-4 are the results of other studies in which the authors used percussion or palpation to locate the lower liver edge. Excluded is one outlying study in which 100% of measurements were accurate within 2 cm of scintigraphic MCL span and exact agreement at the 0.1-cm level is claimed for several observations.¹⁷ We also exclude a study using direct percussion without a pleximeter finger³⁷; this study related mean clinical liver span to ultrasonographic span but lacked measures of either case-by-case absolute span discrepancies or categorical agreement on organ normalcy.

Once the span has been determined, clinicians must still decide whether the liver is enlarged or not. Blendis et al³⁸ reported that among 28 patients with blood dyscrasias or liver diseases examined by 4 observers, 3 of 4 observers agreed in 93% of cases about the presence or absence of hepatic enlargement, but the data do not permit a κ correction. Theodossi et al³³ with 5 observers and a structured medical history and physical examination on 20 jaundiced patients, reported a κ for presence or absence of hepatomegaly of 30%. Moreover, agreement among the qualitative judgments of clinicians and an external reference standard is modest. For example, Blendis et al³⁸ found that in the cases in which at least 3 clinicians agreed on hepatomegaly, concordant assessments of radiologic liver surface area were found in only 48% of cases. Halpern et al²⁰ compared judgments recorded in medical charts with a convenience sample of 214 scintigraphic images with 16 cm as the cut point. Accuracy was 66%, slightly higher than in the study by Blendis et al.³⁸ However, when corrected for agreement expected according to chance alone, the resulting κ statistic was only 32%. Naylor et al³⁵ used 15 cm as a cut point for scintigraphic hepatomegaly and, with 2 observers, found that the accuracy of clinical examination ranged from 67% to 82%, depending on the observer and choice of clinical threshold value for determining the presence of hepatomegaly. Correcting for chance agreement, the κ statistics ranged from 28% to 55%. Overall, it appears that combinations of palpation and percussion yield modest accuracy greater than expected by chance alone in determining whether the liver is enlarged or not.

Castell et al³⁹ suggested measuring span by percussion alone. They examined 116 healthy subjects to establish a range of nor-

mal for percussive span in the MCL and midsternal line. Because the goal was to establish a clinical range of normal, there was no reason to validate the measurements against a reference standard. Percussive span correlated positively with height and differed between men and women, as would be expected from autopsy studies (Table 22-1). Formulas to predict span were derived that incorporated height and weight. The MCL liver dullness for men (cm) = {[0.032 × weight (lb)] + [0.18 × height (in)]} – 7.86 and MCL liver dullness for women (cm) = {[0.027 × weight (lb)] + [0.22 × height (in)]} – 10.75.³⁹

The advantages of percussion alone are that observers may not agree on the presence of a palpable liver, and palpable livers will often be felt below the point at which the percussion note changes. The latter occurs because the thin lower liver edge may not cause dullness. Thus, you must rely on palpation in a variable proportion of subjects because not all livers are palpable, and these subjects will tend to have somewhat larger liver spans. However, clinical MCL span compared with technetium scintigraphic span is less accurate when the lower border is nonpalpable,^{14,19} and errors are always greatest in the upper border that can only be approached by percussion.^{14,19,35} It therefore seems counterintuitive to propose examining liver span by percussion alone. Also, the forcefulness of percussion greatly modifies the measured span.^{19,39,40} Use of percussive span therefore demands that each observer double check his or her own range of normal against the established norms to ensure that strength of percussion is not a confounder.

Another group used the percussive span technique⁴¹ to examine 46 patients with liver disease. There was significant disagreement among the 6 examining clinicians, presumably because of strength and plane of percussion. Interobserver agreement on the appraisal of the organ as “small,” “normal,” or “enlarged” was excellent only for massively enlarged livers. If moderately enlarged organs (ultrasonographic volumes between 2000 and 2700 mL) are included, the probability of any 2 randomly chosen observers agreeing on the presence of hepatomegaly was between 40% and 75%.

This limited performance is perhaps understandable because the concept of percussive span rests on the questionable assumption that it consistently underestimates liver span, allowing for reliable demarcation of abnormally sized livers. Nonetheless, Castell et al³⁹ are the only group to establish a range of normal for clinical liver span that reflects the known variability of span with height, weight, and sex.

Use of percussion alone to determine span, independent of whether or where the lower liver edge is felt, remains feasible. However, clinicians should standardize their percussion technique and compare their typical findings in normal subjects with published normal ranges. Future research should evaluate the clinical use of the percussive method compared with methods using percussion and palpation, with and without the “scratch test.”

PHYSICAL FINDINGS IN CONTEXT

In the foregoing studies, accuracy is generally defined against a single reference standard such as ultrasonography

or scintigraphy. This procedure contrasts with studies such as the one by Rosenfield et al,²² in which measured span and palpability were compared with evidence in the clinical record for any liver disease. The latter study has the advantage of capturing the fact that although all truly enlarged livers are diseased, not all normal-sized livers are free of disease.

A further problem with many studies is the extent of blinding. Some studies blind observers to all details of the patients’ medical history and other physical findings. Others ask observers to perform a structured medical history and physical examination or set inclusion criteria (eg, jaundiced or alcoholic patients) that will affect clinicians’ judgments. The nature and extent of confounding from this variable are unknown, but it seems probable that the extent of interobserver agreement, and even the match between clinical judgments and reference standards, will be affected by the amount of information available to the examiner.

Finally, few studies try to place liver findings in the overall context of clinical decision making. Sapira¹ observed that clinical liver span assessments need not match closely ultrasonographic or scintigraphic measures because the “clinical worth” of a sign is its potential contribution to clinical decision making. Of interest, Espinoza et al¹⁵ used stepwise discriminant analysis to assess the ability of a variety of physical findings to distinguish among 50 consecutive alcoholic patients presenting variously with cirrhosis, noncirrhotic alcoholic liver disease, or no clinical/biochemical evidence of liver disease. Three variables—spider nevi, splenomegaly, and abdominal wall collateral veins—appeared useful; liver examination findings were not significant contributors to the differential diagnostic exercise. Similarly, Theodossi et al⁴² and Theodossi⁴³ examined the ability of a large array of symptoms and signs to differentiate between medical and surgical causes of jaundice. They found that descent of the liver edge greater than 2 cm below the costal margin was more common with surgical causes of jaundice ($P < .01$), but the independent contribution of this sign to the overall diagnostic process was unclear.

Both studies started with populations that had liver disease and determined whether physical diagnosis helped in categorizing the type of disease. Neither addresses whether the physical examination was helpful in deciding which patients had liver disease in the first place. Little is known about the real contribution of liver examination findings to the overall clinical diagnostic and management process. This topic should be a research priority.

WHAT CAN YOU DO TO GET BETTER AT EXAMINING THE LIVER?

No educational studies, to our knowledge, have tested methods to improve your accuracy and precision in examining the liver, but a few suggestions can be hazarded. First, once you are comfortable examining the liver, pursue the various shortcuts recommended herein. Early on, however, it is useful to check the liver span by percussion, even in persons with a low probability of liver disease and a nonpalpable

organ. This method can help you begin to understand what your own range of normal is likely to be. Second, check your reliability by reexamining stable patients and comparing your follow-up assessment with your first impressions. Third, both learners and experts should quantitatively and qualitatively benchmark their physical examinations of the liver against findings on nuclear examination or ultrasonography. Try to determine how you are doing in assessing vertical liver span or extent of descent of the edge below the costal margin or in “calling” the presence of hepatomegaly. Fourth, consider the potential errors in locating the MCL. If sequential clinical span assessments are being made (eg, fulminant hepatic failure or treatment of hepatic metastases), it may help to record a reference plane such as 10 cm from the midline or where the lateral edge of the rectus abdominis crosses the costal margin.⁴⁴

THE BOTTOM LINE

Once historical data and other physical signs have been elicited, the additional value of a detailed physical examination of the liver remains uncertain. Moreover, just as diagnostic tests yield little at the extremes of prior probability so also would you expect less yield from liver examination in persons who are not suspected of having liver disease or who obviously have some hepatobiliary complaint.

A selective approach to physical examination of the liver is therefore suggested. Palpate to locate the lower liver border in the MCL in situations of low probability of liver disease. If the liver is not palpable, one can defensibly forgo any further examination in patients without reasons to suspect liver disease. However, because palpation of the abdomen is difficult in some subjects, light percussion remains an option to confirm lack of extension of the liver edge below the costal margin or guide further palpation. With a palpable lower edge, MCL span can be ascertained by light percussion of the upper border. A span of less than 13 cm reduces the probability of hepatomegaly. In persons with an impalpable liver and a high probability of liver disease, measuring span by percussion alone may also be worthwhile; tables of norms have been published, although these apply to moderate or heavy percussion methods. Palpation specifically to assess the quality of the liver edge is recommended only if there are signs of liver disease, including unequivocal hepatomegaly. Auscultation over the liver has a limited role in examination.

Author Affiliations at the Time of the Original Publication

Clinical Epidemiology Unit and Division of General Internal Medicine, Sunnybrook Health Science Centre, Toronto, Ontario, Canada; and Departments of Medicine and Surgery and the Graduate Department of Community Health, University of Toronto, Toronto, Ontario, Canada.

Acknowledgments

The author has learned from the comments of David L. Sackett, MD, and David L. Simel, MD, MHS, who offered constructive editorial advice, and from helpful hepatologic

colleagues in Toronto, namely, Jenny Heathcote, MD, and Laurence Blendis, MD. Joseph D. Sapira, MD, also shared his wisdom as a reviewer but frequently disagreed with what has been written here. These persons should share any credit and be spared any blame for errors or omissions.

REFERENCES

1. Sapira JD. Why perform a routine history and physical examination? *South Med J*. 1989;82(3):364-365.
2. Zuidema PJ, Van der Meer C. Systolic liver murmurs in primary carcinoma of the liver. *Trop Geogr Med*. 1966;18(4):265-271.
3. Fenster LF, Klatskin G. Manifestations of metastatic tumors of the liver. *Am J Med*. 1961;31:238-248.
4. Sherman HI, Hardison JE. The importance of a coexistent hepatic rub and bruit. *JAMA*. 1979;241(14):1495.
5. Sapira JD. *The Art and Science of Bedside Diagnosis*. Baltimore, MD: Urban & Schwarzenberg; 1990:377-384.
6. Ramakrishnan T. Venous hum of the Cruveilhier-Baumgarten syndrome. *Arch Intern Med*. 1978;138(5):826.
7. Hardison JR. Venous hum of the Cruveilhier-Baumgarten syndrome. *Arch Intern Med*. 1977;137(11):1623-1624.
8. Zoneraich S, Zoneraich O. Diagnostic significance of abdominal arterial murmurs in liver and pancreatic disease: a photoarteriographic study. *Angiology*. 1971;22(4):197-205.
9. Clain D, Wartnaby K, Sherlock S. Abdominal arterial murmurs in liver disease. *Lancet*. 1966;2(7462):516-519.
10. Motoki T, Hayashi T, Katoh Y, Sakamoto T, Takeda T, Murao S. Hepatic bruits in liver tumours. *Lancet*. 1978;2:259.
11. Motoki T, Hayashi T, Katoh Y, Sakamoto T, Takeda T, Murao S. Hepatic bruits in malignant liver tumors. *Am J Gastroenterol*. 1979;71(6):582-586.
12. Meyhoff HH, Roder O, Andersen B. Palpatory estimation of liver size. *Acta Chir Scand*. 1979;145(7):479-481.
13. Kingston M, Ali MA, Lewall D. Hepatic tumors in Saudi Arabia: a practical approach to diagnosis. *Cancer*. 1985;55(7):1579-1585.
14. Peternel WW, Schaefer JW, Schiff L. Clinical evaluation of size and hepatic scintiscan. *Am J Dig Dis*. 1966;11(5):346-350.
15. Espinoza P, Ducot B, Pelletier G, et al. Interobserver agreement in the physical diagnosis of alcoholic liver disease. *Dig Dis Sci*. 1987;32(3):244-247.
16. Palmer ED. Palpability of the liver edge in healthy adults. *US Armed Forces Med J*. 1958;9(12):1685-1690.
17. Naftalis J, Leevy CM. Clinical examination of liver size. *Am J Dig Dis*. 1963;8:236-243.
18. Ralphs DNL, Venn G, Kan O, Palmer JG, Cameron DE, Hobsley M. Is the undeniably palpable liver ever “normal”? *Ann R Coll Surg Engl*. 1983;65(3):159-160.
19. Sullivan S, Krasner N, Williams R. The clinical estimation of liver size: a comparison of techniques and an analysis of the source of error. *BMJ*. 1976;2(6043):1042-1043.
20. Halpern S, Coel M, Ashburn W, et al. Correlation of liver and spleen size: determinations by nuclear medicine studies and physical examination. *Arch Intern Med*. 1974;134(1):123-124.
21. Ariel IM, Briceno M. The disparity of the size of the liver as determined by physical examination and by hepatic gamma-scanning in 504 patients. *Med Pediatr Oncol*. 1976;2(1):69-73.
22. Rosenfield AT, Laufer I, Schneider PB. The significance of a palpable liver. *Am J Roentgenol Radiat Therapy Nucl Med*. 1974;122(2):313-317.
23. Walk L. Liver size and its post-mortem change. *Radiologe*. 1983;23(4):189-191.
24. Simel DL, Samsa GP, Marchar DB. Likelihood ratios with confidence: sample size estimation for diagnostic test studies. *J Clin Epidemiol*. 1991;44(8):763-770.
25. Da Costa JC Jr. *Principles and Practice of Physical Diagnosis*. 3rd ed. Philadelphia, PA: WB Saunders Co; 1915:542.
26. Bokhari SI, DeSa'Neto A, Desser KB, Benchimol A. Inspiratory augmentation of hepatic pulsation in tricuspid insufficiency. *Chest*. 1979;76(1):89-90.
27. George DS, Marsh C, Leier CV. A-wave: liver. *Clin Cardiol*. 1988;11(5):349-350.

28. Pelteret RM, de Moor MM. Unusual cause of pulsatile liver: a case report. *S Afr Med J*. 1988;74(2):74-75.
29. Muller O, Shillingford J. Tricuspid incompetence. *Br Heart J*. 1954;16(2):195-207.
30. Bhatia ML, Tewari HL. Hepatic pulsations in constrictive pericarditis. *Indian Heart J*. 1974;26(suppl):165-170.
31. Manga P, Vythilingum S, Mitha AS. Pulsatile hepatomegaly in constrictive pericarditis. *Br Heart J*. 1984;52(4):465-467.
32. Osler W. *The Principles and Practice of Medicine*. New York, NY: Appleton & Co; 1893:428-456.
33. Theodossi A, Knill-Jones RP, Skene A, et al. Inter-observer variation of symptoms and signs in jaundice. *Liver*. 1981;1(1):21-32.
34. Naylor CD, McCormack DG, Sullivan SN. The midclavicular line: a wandering landmark. *CMAJ*. 1987;136(1):48-50.
35. Naylor CD, McCormack DG, Driedger AA, Cradduck TD. Assessing liver size: clinical span versus scintigraphic span and tomographic volume. *Clin Invest Med*. 1988;11:C90-R573.
36. Fuller GN, Hargreaves MR, King DM. Scratch test in clinical examination of the liver. *Lancet*. 1988;1(8578):181.
37. Skrainka B, Stahlhut J, Knight F, Holmes RA, Butt JH. Measuring liver span: bedside examination versus ultrasound and scintiscan. *J Clin Gastroenterol*. 1986;8(3 pt 1):267-270.
38. Blendis LM, McNeilly WJ, Sheppard L, Williams R, Laws JW. Observer variation in the clinical and radiological assessment of hepatosplenomegaly. *BMJ*. 1970;1(5698):727-730.
39. Castell DO, O'Brien KD, Muench H, Chalmers TC. Estimation of liver size by percussion in normal individuals. *Ann Intern Med*. 1969;70(6):1183-1189.
40. Sapira JD, Williamson DL. How big is the normal liver? *Arch Intern Med*. 1979;139(9):971-973.
41. Malchow-Møller A, Rasmussen SN, Jensen AM, Keilding N, Skovgaard LT, Juhl E. Clinical estimation of liver size. *Dan Med Bull*. 1984;31(1):63-67.
42. Theodossi A, Spigelhalter DJ, Portmann B, Eddleston ALWF, Williams R. The value of clinical, biochemical, ultrasound and liver biopsy data in assessing patients with liver disease. *Liver*. 1983;3(5):315-326.
43. Theodossi A. The value of symptoms and signs in the assessment of jaundiced patients. *Clin Gastroenterol*. 1985;14(3):545-557.
44. Harding SR, Cowan DHC. The midclavicular line: a wandering landmark. *CMAJ*. 1987;136(9):921.

This page intentionally left blank

Prepared by David L. Simel, MD, MHS

Reviewed by Marisa D'Silva, MD

CLINICAL SCENARIO

A 21-year-old college student with a flulike illness for 2 days presents to the student health clinic. You suspect influenza and examine her oropharynx, neck, chest, and abdomen. When you feel her liver edge about 2 cm below the costal margin, you inquire about abdominal discomfort, nausea, vomiting, and anorexia other than with the current illness. Her skin is not jaundiced. She has no history of liver disease or illnesses associated with liver enlargement. You reexamine the abdomen to confirm the presence of the liver edge and additionally find no evidence for splenomegaly.

UPDATED SUMMARY ON HEPATOMEGALY

Original Review

Naylor CD. The rational clinical examination: physical examination of the liver. *JAMA*. 1994;271(23):1859-1865.

UPDATED LITERATURE SEARCH

Our literature search used the parent search strategy for The Rational Clinical Examination series, combined with the subject “hepatomegaly/di,” published in English from 1993 to 2004. The results yielded 71 titles, for which we reviewed the titles and abstracts; 13 were selected for additional review. These articles were reviewed to identify articles that assessed the sensitivity and specificity of the medical history or physical examination for hepatomegaly. Two articles were identified for inclusion.

NEW FINDINGS

- Clinicians should stop assessing the liver span by “scratching” the abdomen.

Details of the Update

A nonsystematic review was published at about the same time as the original Rational Clinical Examination article.¹ The conclu-

sions in the 2 articles were similar in observing that palpation of the liver edge occurs commonly in healthy patients.

The scratch test was suggested as a method for determining the distance below the right costal margin. A study with methodologic flaws, which should have enhanced the accuracy of this method, found no correlation between the distance of the edge below the costal margin and the total liver span identified by ultrasonography.²

A second study confirmed the relationship between the liver edge identified by percussion and liver span (confirmed by ultrasonography), but only in patients with cirrhosis.³

IMPROVEMENTS IN THE DATA PRESENTED IN THE ORIGINAL PUBLICATION

The original publication did not use meta-analytic techniques for assessing the pooled likelihood ratios (LRs) for palpating the liver edge. The data were reanalyzed, and they showed that the presence of a palpated liver edge is not as good as previously reported (LR, 2.0 vs previously reported 2.5) for identifying a patient with an increased liver span or volume (see Table 22-5).

CHANGES IN THE REFERENCE STANDARD

None.

RESULTS OF LITERATURE REVIEW

The scratch test for identifying the edge of the liver below the costal margin yields a result with no correlation to the actual liver span ($r = 0.04$).

Percussion of the liver in cirrhotic patients agrees with the total liver span measured by ultrasonography ($\kappa = 0.93$).

Table 22-5 Likelihood Ratio for a Palpable Liver Edge to Identify Hepatomegaly

Finding (No. of Studies)	LR+ (95% CI)	LR- (95% CI)
Palpable liver edge (3)	2.0 (1.5-2.8)	0.41 (0.3-0.55)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

EVIDENCE FROM GUIDELINES

There are no guidelines addressing an assessment for hepatomegaly in the general population.

CLINICAL SCENARIO—RESOLUTION

Your previous suspicion of liver disease is low, and identifying the liver edge in this young patient is likely a normal finding. You should consider screening for excessive alcohol use because she could have a fatty liver unrelated to the current illness. Although you might consider infectious mononucleosis as the current underlying illness, hepatomegaly is not a common presentation (as opposed to splenomegaly). Additional testing for liver enlargement is not indicated unless there are other suggestions that she might have liver disease.

HEPATOMEGLY—MAKE THE DIAGNOSIS

PRIOR PROBABILITY

The probability that the liver edge can be felt below the right costal margin is about 50%. However, this does not correlate with the liver span in normal patients. Thus, the prior probability of hepatomegaly depends entirely on the possible underlying disease states.

POPULATION FOR WHOM HEPATOMEGLY SHOULD BE CONSIDERED

- Known or suspected liver disorders
- Malignancy
- Congestive heart failure

DETECTING THE LIKELIHOOD OF HEPATOMEGLY

Palpating a liver edge below the right costal margin correlates poorly with the actual liver span, although it does increase the likelihood that the patient will have an enlarged liver (positive LR, 2.0). Likewise, the failure to identify the liver edge does not rule out the presence of an increased liver span (negative LR, 0.41). The effect of this finding depends on the previous suspicion of liver disease.

When there is a suspicion of liver disease, we recommend that clinicians forgo the “scratch” test and use percussion to estimate the liver span (>15 cm = enlargement). Liver ultrasonography will be required to confirm the clinical findings.

REFERENCE STANDARD TESTS

Ultrasonography or scintigrams.

REFERENCES FOR THE UPDATE

1. Meidl EJ, Ende J. Evaluation of liver size by physical examination. *J Gen Intern Med*. 1993;8(11):635-637.
2. Tucker WN, Saab S, Rickman LS, Mathews WC. The scratch test is unreliable for detecting the liver edge. *J Clin Gastroenterol*. 1997;25(2):410-414.^a

3. Zoli M, Magalotti D, Grimaldi M, Gueli C, Marchesini G, Pisi E. Physical examination of the liver: is it worth it? *Am J Gastroenterol*. 1995;90(9):1428-1432.^a

^aFor the Evidence to Support the Update for this topic,
see <http://www.JAMAevidence.com>.

EVIDENCE TO SUPPORT THE UPDATE:

22

Hepatomegaly

TITLE The Scratch Test Is Unreliable for Detecting the Liver Edge.

AUTHORS Tucker WN, Saab S, Rickman LS, Mathews WC.

CITATION *J Clin Gastroenterol.* 1997;25(2):410-414.

QUESTION What is the interobserver variability of the scratch test for measuring the liver span below the right costal margin?

DESIGN Multiple independent examinations. The ultrasonography was performed before the physical examination and thus was blinded to the scratch test. It is not clear whether the examiners knew the results of the ultrasonography.

SETTING Patients were identified from a list of those undergoing abdominal ultrasonography. The examiners included attending physicians (2), a gastroenterologist and an infectious disease specialist, gastroenterology fellows (2), chief medical residents (2), a medicine resident, senior medical students (3), and a nurse practitioner.

PATIENTS Inpatients ($n = 22$). Most patients had normal body habitus, although 3 had ascites and 1 was greater than ideal body weight.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

A scratch test with the results recorded as the location of the liver edge in centimeters below the right costal margin.

MAIN OUTCOME MEASURE

Liver span measured by ultrasonography.

MAIN RESULTS

Of 22 patients, 18 (80%) had hepatomegaly (liver span > 15.5 cm). There was no correlation in the ultrasonographically measured liver span and the span of the liver below the right costal margin by the scratch test ($r = 0.05$). The pairwise reliability coefficient ranged from -0.32 to 0.74, with a mean of 0.26 (poor correlation).

CONCLUSIONS

LEVEL OF EVIDENCE Level 4.

STRENGTHS Many examiners with different training levels.

LIMITATIONS There is no certainty that the examiners were blinded to the ultrasonographic results. However, if they knew the results of the ultrasonography, the bias would have been toward an improved correlation. There was a high prevalence of patients with known liver disease, although this is the population for whom measuring the liver span would be most applicable.

The biases in this study should have enhanced the correlation between scratch test determination of the liver edge and the actual liver edge by ultrasonography. Despite the significant limitations in study population, there was no correlation. From this study, we can conclude that physicians should stop scratching the abdomen to identify the liver edge for patients with suspected liver disease.

Reviewed by David L. Simel, MD, MHS

TITLE Physical Examination of the Liver: Is It Still Worth It?

AUTHORS Zoli M, Magalotti D, Grimaldi M, Gueli C, Marchesini G, Pisi E.

CITATION *Am J Gastroenterol.* 1995;90(9):1428-1432.

QUESTION What is the correlation between palpating the lower liver edge and the actual size of the liver?

DESIGN Independent, prospective, nonconsecutive evaluation of case patients and control patients.

SETTING Gastroenterology clinic. The examiners were one of 2 clinical hepatologists.

PATIENTS Cases were patients with cirrhosis and control patients were healthy.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

Total liver span in the midclavicular line and the presence of a palpable liver below the right costal margin. The examiners used percussion to identify the liver edge.

MAIN OUTCOME MEASURE

Palpation was compared with ultrasonography results. The clinicians were not informed of whether the patient was a case or a control. The ultrasonographer did not have the physical examination results.

MAIN RESULTS

Forty-seven of 100 control patients had a liver edge palpated below the costal margin vs 78 of 100 patients with cirrhosis.

When the liver edge was palpated below the right costal margin, the likelihood that the ultrasonography would identify the edge below the margin was 39 (95% confidence interval [CI], 4.6-373). When the edge was not palpated, the likelihood ratio (LR) for ultrasonography identifying the edge below the margin was 0.28 (95% CI, 0.22-0.36).

The investigators compared the measured distance from the costal margin to the liver edge vs the distance measured with

ultrasonography. There was poor agreement for control subjects ($\kappa = 0.13$) and excellent agreement for cirrhotic patients ($\kappa = 0.93$).

However, identifying the edge below the margin is not the same as identifying a large liver. The liver span was more than 15 cm in 18 control and 8 case patients. Among patients with cirrhosis, the clinical estimation correlated with the liver span ($r = 0.59$). There was no statistically significant correlation for control subjects ($r = 0.20$). The investigators also compared liver span by palpation to liver volume by ultrasonography; the correlation was good for case patients ($r = 0.65$ for cirrhosis patients) but not for healthy patients ($r = 0.33$). The same correlation results were found when liver span below the costal margin was compared with the liver volume.

Data are presented that allow us to estimate the sensitivity of the pattern of the liver margin as an indicator for cirrhosis. The positive LR for a thickened liver margin for cirrhosis is 2.4 (95 % CI, 1.4-4.4). The LR for finding a sharp edge is 0.62 (95% CI, 0.46-0.81).

CONCLUSIONS

LEVEL OF EVIDENCE Level 4.

STRENGTHS Independent comparison, although the examiners likely knew from clinical observations that some of the case patients had cirrhosis. The interobserver variability of the ultrasonographers was reported as 3 mm or less.

LIMITATIONS Nonconsecutive patients in whom the examiners knew that half of the patients had cirrhosis and others did not. The data are not presented in a fashion that allows us to extract the sensitivity and specificity for percussion of the liver edge.

A large number of healthy patients will have their liver edge palpated below the right costal margin, which indicates neither disease nor the presence of hepatic enlargement. These results support the suggestion that physicians should specifically assess the liver edge only when there is a suspicion of liver disease.

The study design (high prevalence of cirrhosis) does not allow us to extrapolate the results for the pattern of the liver edge to a population of patients with a suspicion of other liver diseases ("thick" indicating cirrhosis and "sharp" indicating normal liver).

Reviewed by David L. Simel, MD, MHS

CHAPTER 23

Does This Patient Have Hypertension?

Richard A. Reeves, MD, FRCPC

CLINICAL SCENARIO

Is This Patient's Blood Pressure Really Elevated?

A 46-year-old man who has recently moved to your neighborhood presents with a painful ankle sprain. Before he leaves, you decide to check his blood pressure (BP) and obtain an initial reading of 164/102 mm Hg. He denies having high BP previously.

WHY IS ACCURATE BP MEASUREMENT IMPORTANT?

Elevated arterial BP, or hypertension, is important because it is common, it is clinically silent, it leads to cardiovascular disease (CVD), and it decreases life expectancy. Because surveys find that approximately 20%¹⁻³ of North American adults have an elevated BP (systolic BP [SBP] \geq 140 mm Hg or diastolic BP [DBP] \geq 90 mm Hg) or are taking antihypertensive medication, physicians are advised to check all patients periodically for BP elevation.³⁻⁷ On the other hand, overestimation of BP can erroneously label people as hypertensive and potentially result in unnecessary dietary restrictions, exposure to adverse effects from drug treatment, medication expense, and adverse socioeconomic effects.^{8,9} Fortunately, measuring BP is an easy and safe diagnostic procedure that, when followed by appropriate antihypertensive drug treatment, can lead to reduced CVD and mortality.^{10,11}

STANDARDS FOR MEASURING BP

The gold standard for instantaneous BP measurement is the intra-arterial or direct BP (determined by a rigid-walled catheter). The standard for clinical practice is the so-called casual cuff or indirect BP.

Guidelines for Diagnosing Hypertension

Cardiovascular disease risk increases monotonically with BP, revealing no cut point below which risk is minimal and above which CVD will definitely occur. Terms used to indicate the degree of BP elevation now emphasize the importance of what was previously termed mild hypertension and the long-recognized greater predictive value of elevated SBP¹² (Table 23-1). Risk for future CVD is predicted by even a single careful BP reading.¹³⁻¹⁵ However, BP is rather variable and often decreases with observation so that, in accord with statistical expectations, risk relates more closely to mean BP during several visits¹³ (although brief, severe BP elevation can also be catastrophic, eg, with cocaine overdose). Therefore, we could define the “treatable BP level” as that mean clinical BP above which treatment has been shown in randomized controlled trials to do more good than harm. The largest of these trials used drug treatment vs placebo after finding a consistent or average entry BP from 2 to 3 visits of greater than or equal to 160 mm Hg SBP (tested only in the elderly) with

or without DBP elevation,¹⁶ or greater than or equal to 90 mm Hg DBP (tested in the young and in the elderly).¹¹ In the future, individualized assessments of absolute risk incorporating

Table 23-1 Classification of Blood Pressure for Adults Aged 18 Years and Older^a

Category	Systolic, mm Hg	Diastolic, mm Hg
Normal	<130	<85
High normal	130-139	85-89
Hypertension ^b		
Stage 1 (mild)	140-159	90-99
Stage 2 (moderate)	160-179	100-109
Stage 3 (severe)	180-209	110-119
Stage 4 (very severe)	≥210	≥120

^aAdapted from the fifth report of the Joint National Committee on Detection, Evaluation, and Treatment of High Blood Pressure.³

^bBased on the average of 2 or more readings taken at each of 2 or more visits after an initial screening.

other relevant information, such as age, sex, concomitant risk factors, and coexisting target organ damage, along with the patient's tolerance for risk and history of drug adverse effects, may replace arbitrary cut points in determining when BP elevation becomes treatable.⁷ At present, a diagnosis of hypertension reflects a consensus regarding the office BP level above which CVD risk worsens significantly, about 140/90 mm Hg.

A detailed conceptual analysis of hypertension is beyond the scope of this article but has been addressed thoughtfully by Jennings and Netsky.¹⁷

How to Measure Clinical BP

Meticulous technique in indirect auscultatory BP measurement is mandatory for research, diagnosis, and optimal clinical care of hypertensive patients. Published procedural guidelines show general uniformity that, if followed, should improve the accuracy and reliability of BP measurement (Table 23-2; Figure 23-1).^{3-5,18-20} BP is customarily measured after obtaining the medical history as part of the “vital sign”

Table 23-2 Techniques for Measuring Blood Pressure^a

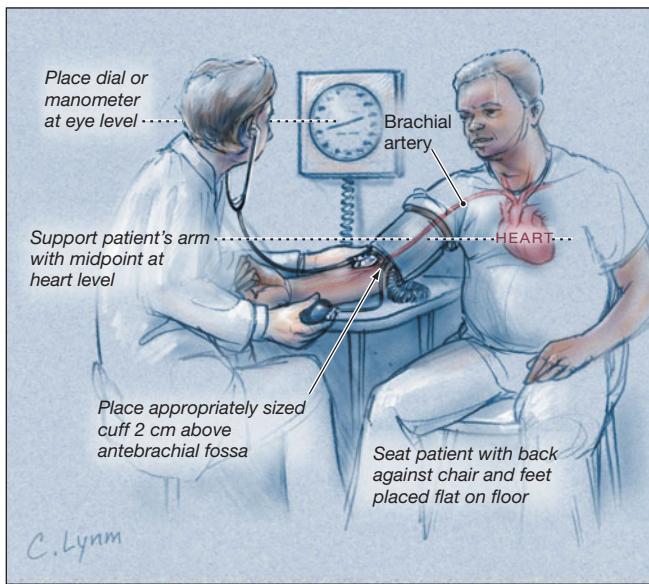
The intent and purpose of the measurement should be explained to the subject in a reassuring manner and every effort made to put the subject at ease [including a 5-min rest before the first measurement].

The sequential steps for measuring BP in the upper extremity, as for routine screening and monitoring purposes, should include the following:

1. Have paper and pen at hand for immediate recording of the pressure.
2. Seat the subject in a quiet, calm environment [with feet flat on the floor, back against the chair] with his or her bared arm resting on a standard table or other support so the midpoint of the upper arm is at the level of the heart.
3. Estimate by inspection or measure with a tape the circumference of the bare upper arm at the midpoint between the acromion and olecranon process and select an appropriately sized cuff. The bladder inside the cuff should encircle 80% of the arm in adults and 100% of the arm in children < 13 years. If in doubt, use a larger cuff. If the available cuff is too small, this should be noted.
4. Palpate the brachial artery and place the cuff so that the midline of the bladder is over the arterial pulsation and then wrap and secure the cuff snugly around the subject's bare upper arm. Avoid rolling up the sleeve in such a manner that it forms a tight tourniquet around the upper arm. Loose application of the cuff results in overestimation of the pressure. The lower edge of the cuff should be 1 in [2 cm] above the antecubital fossa where the head of the stethoscope is to be placed.
5. Place the manometer so the center of the mercury column or aneroid dial is at eye level [except for tilted-column floor models] and easily visible to the observer and the tubing from the cuff is unobstructed.
6. Inflate the cuff rapidly to 70 mm Hg, and increase by 10 mm Hg increments while palpating the radial pulse. Note the level of pressure at which the pulse disappears and subsequently reappears during deflation. This procedure, the palpatory method, provides a necessary preliminary approximation of the SBP to ensure an adequate level of inflation when the actual, auscultatory measurement is made. The palpatory method is particularly useful to avoid underinflation of the cuff in patients with an auscultatory gap and overinflation in those with very low BP.
7. Place the earpieces of the stethoscope into the ear canals, angled forward to fit snugly. Switch the stethoscope head to the low-frequency position (bell). The setting can be confirmed by listening as the stethoscope head [ie, the bell orifice] is tapped gently.
8. Place the head of the stethoscope over the brachial artery pulsation, just above and medial to the antecubital fossa but below the edge of the cuff, and hold it firmly [but not too tightly²¹] in place, making sure that the head makes contact with the skin around its entire circumference. Wedging the head of the stethoscope under the edge of the cuff may free up one hand but results in considerable extraneous noise [and is nearly impossible with the bell in any event].
9. Inflate the bladder rapidly and steadily to a pressure 20-30 mm Hg above the level previously determined by palpation and then partially unscrew [open] the valve and deflate the bladder at 2 mm [Hg]/s while listening for the appearance of the Korotkoff sounds.
10. As the pressure in the bladder decreases, note the level of the pressure on the manometer at the first appearance of repetitive sounds [phase I] and at the muffling of these sounds [phase IV] and when they disappear [phase V]. During the period the Korotkoff sounds are audible, the rate of deflation should be no more than 2 mm per pulse beat, thereby compensating for both rapid and slow heart rates.
11. After the Korotkoff sound is heard, the cuff should be deflated slowly for at least another 10 mm Hg to ensure that no further sounds are audible and then rapidly and completely deflated, and the subject should be allowed to rest for at least 30 s.
12. The systolic [phase I] and diastolic [phase V] pressures should be immediately recorded, rounded off [upward] to the nearest 2 mm Hg. In children, and when sounds are heard nearly to a level of 0 mm Hg, the phase IV pressure should also be recorded [eg, 108/64/56 mm Hg]. All values should be recorded together with the name of the subject, the date and time of the measurement, the arm on which the measurement was made, the subject's position, and the cuff size [when a nonstandard size is used].
13. The measurement should be repeated after at least 30 s and the 2 readings averaged. In clinical situations, additional measurements can be made in the same or opposite arm, in the same or an alternative position.

Abbreviations: BP, blood pressure; SBP, systolic blood pressure.

^aReproduced with permission from Perloff et al.¹⁸ Copyright © 1993 American Heart Association. Text in brackets added by the author.

**Figure 23-1 Clinical Measurement of Indirect Blood Pressure**

See Table 23-2 for appropriate cuff sizing.

determination at the beginning of the physical examination. At each visit, 2 or more readings should be obtained and averaged from the same arm, with the patient supine or seated. As a practical approach to variability, taking additional readings until a stable level is reached has been suggested when the first 2 differ by more than 5 mm Hg diastolic.²⁰ BP in both arms should be measured at the first visit, and the arm with the higher pressure should be used thereafter.¹⁸

Careful technique guarantees maximum accuracy. We have compiled information from a number of sources regarding factors that increase, decrease, or have no effect on BP²¹⁻⁷⁴ (Table 23-3). However, if all serious errors that can underestimate BP are avoided, finding the BP in any setting, position, or time to be within the normal range makes a more careful measurement at that visit unlikely to be high. Assuming BP is checked routinely in all adults, the efficient practitioner can reasonably reserve the “proper” method for the 10% to 20% of patients who have known or newly detected elevated BP (as in our clinical scenario), cardiovascular target-organ damage, or other risk factors or who are receiving antihypertensive therapy.

Variation in BP Measurement

Sources of clinical variability include the patient, equipment, examiner, and procedure. For BP, a major proportion of random fluctuation over time arises from the examinee. Intraarterial monitoring clearly reveals that SBP and DBP differ with every heartbeat and with the respiratory cycle.⁴⁵ Blood pressure also varies minute to minute, with a standard deviation of about 4 mm Hg systolic and 2 to 3 mm Hg diastolic,^{59,75} as well as during hours^{76,77}; short-term variability in SBP is increased with impaired baroreflexes.^{77,78} Day-to-day variation is even greater. With 2 or more cuff readings at each visit, the standard deviation between visits is approximately 5 to 12 mm Hg systolic and 6 to 8 mm Hg diastolic.^{13,59,60,79,80}

Table 23-3 Factors Affecting the Immediate Accuracy of Office Blood Pressure

Factor	Magnitude, SBP/DBP, mm Hg	Reference
Increases Recorded BP		
Examinee		
Soft Korotkoff sounds	DBP	Assumed
Missed auscultatory gap	DBP (rare, huge)	22
Pseudohypertension	2-98/3-49	23-25
“White coat” reaction		
To physician	11-28/3-15	26-30
To nonphysician	1-12/2-7	27, 31, 32
Paretic arm (caused by stroke)	2/5	33
Pain, anxiety	May be large	22
Acute smoking	6/5	34
Acute caffeine	11/5	35
Acute ethanol ingestion	8/8	36
Distended bladder	15/10	37
Talking, signing	7/8	38, 39
Setting, equipment		
Environmental noise	DBP	Assumed
Leaky bulb valve	≥2 DBP	40
Blocked manometer vents	2-10	41
Cold hands or stethoscope	Not stated	22
Examiner		
Expectation bias	Probably < 10	In theory
Impaired hearing	DBP	22
Examination		
Cuff too narrow	-8 to +10/2-8	42-44
Cuff not centered	4/3	45
Cuff over clothing	5-50	46
Elbow too low	6	47
Cuff too loose	Not stated	48
Too short rest period	Varied estimates	22
Back unsupported	6-10	49, 50
Arm unsupported	1-7/5-11	51
Too slow deflation	-1 to +2/5-6	52, 53
Too fast deflation	DBP only	52, 53
Parallax error	2-4	By author
Using phase IV (adult)	6 DBP	45
Too rapid remeasure	1/1	52, 54
Cold (vs warm) season	6/3-10	55-57
Decreases BP		
Examinee		
Soft Korotkoff sounds	SBP	Assumed
Recent meal	-1 to 1/1-4	58
Missed auscultatory gap	10-50 SBP	45
High stroke volume	Phase V can = 0	45
Habituation	0-7/2-12	59-61
Shock (additional pseudohypotension)	33 SBP	62
Setting, equipment		
Noisy environs	SBP	Assumed

(continued)

Table 23-3 Factors Affecting the Immediate Accuracy of Office Blood Pressure (*Continued*)

Factor	Magnitude, SBP/DBP, mm Hg	Reference
Faulty aneroid device	Can be > 10	63
Low mercury level	Varies	22
Leaky bulb	≥2 SBP	40
Examiner		
Reading to next lowest 5-10 mm Hg, or expectation bias	Probably ≤ 10	64
Impaired hearing	SBP only	22
Examination		
Noisy environs	SBP	Assumed
Left vs right arm	1/1	65
Resting for too long (25 min)	10/0	66
Elbow too high	5/5	47
Too rapid deflation	SBP only	40
Excess bell pressure	≥9 DBP	21
Parallax error (aneroid)	2-4	By author
No Effect on BP		
Examinee		
Menstrual phase		67, 68
Chronic caffeine ingestion		69
Phenylephrine nasal spray		70
Cuff self-inflation		71
Examinee and examiner		
Discordance in sex or race		72, 73
Examination		
Thin shirtsleeve under cuff		74
Bell vs diaphragm		49
Cuff inflation per se		29
Hour of day (during work hours)		54
Room temperature		54

Abbreviations: BP, blood pressure; DBP, diastolic blood pressure; SBP, systolic blood pressure.

This variability explains why 2 BP measurements in a patient often differ, but it also suggests that a repeated visit's measurements could be as much as 15/12 mm Hg higher or lower than today's result about 5% of the time. The greater magnitude of the between-visit vs within-visit variability is why more visits are recommended to achieve greater diagnostic precision rather than more replications at a visit. In reality, the return BP reading in our clinical scenario will likely be lower, possibly much lower, because of our patient's present distress, unfamiliarity with the physician and the physician's office procedures, and "regression to the mean" (discussed herein).

Arrhythmias, particularly atrial fibrillation, cause beat-to-beat cardiac output to vary substantially and increase inter-observer variation in measured BP.⁸¹ With atrial fibrillation, probably the best one can do is to deflate the cuff slowly while attempting to ascertain when most of the contractions are resulting in audible Korotkoff sounds (the approximate

SBP) and when the sounds have all but ceased yet still occur infrequently (the approximate DBP), or one can average several readings.^{18,20} Because Korotkoff sounds generated by occasional premature beats (and the subsequent beat) are unrepresentative of the day's mean BP level, they should be ignored.²²

Examiners can introduce random errors. Under ideal conditions, simultaneous BP readings by independent observers typically correlate above $r = 0.95$ with mean absolute differences of less than 2 mm Hg systolic and less than 1 mm Hg diastolic.⁸² However, even in research settings, careful BP readings obtained just a few minutes apart show distressingly high variation (eg, SD of 7 mm Hg systolic and 5 mm Hg diastolic^{13,54,61}). In routine medical practice, physicians and nurses often measure BP far less carefully: differences of 10/8 mm Hg are common.^{83,84} White et al⁸⁵ performed intra-arterial BP recording in 48 hypertensive patients and found the humbling result that 2 auscultatory automatic monitors showed less overall discrepancy and fewer widely discrepant readings compared with those generated by experienced clinicians using the standard method.

Environmental problems (eg, noise from construction work next door) or deficient equipment (eg, an inadequately damped, "bouncy" mercury column, remedied by tightening the knurled nut at the column's top²²) may also be expected to decrease precision.

Accuracy of BP Measurement

Accuracy, or validity, refers to agreement with the truth and requires not only precision but also freedom from systematic error (ie, bias). In clinical BP measurement, we look through a series of dark glasses, further considered herein: (1) the indirect BP may not reflect the concurrent intra-arterial BP; (2) the cuff technique may be incorrectly performed; and (3) a perfectly executed indirect (or even direct) BP reading at a particular moment may not represent the patient's average clinic BP or the average BP throughout the day's activities, as addressed in the section on ambulatory BP monitoring. Finally, to interpret even a perfect BP reading requires consideration of the whole patient because factors other than BP strongly influence the risk for CVD events.

Indirect BP vs Direct BP

Indirect auscultatory BP correlates well with the simultaneous intra-arterial value ($r = 0.94\text{-}0.98$).⁸⁶ However, the Korotkoff phase I sounds do not appear until 15 to 4 mm Hg below the direct SBP, whereas, at phase V, the sounds disappear 3 to 6 mm Hg above the true DBP in adults.^{45,85,86}

If these technical differences applied equally to all patients, they would be merely academic; clinical importance arises when an individual patient possesses an unusual discrepancy. Such patients are often elderly (in which false elevation is termed "pseudohypertension"^{23,24,87-89}) or obese,⁴⁶ but otherwise unexplained extreme false elevations in cuff BP may also occur.⁹⁰ Pseudohypertension might seem at first glance to be a variant of normal BP. However, most patients actually have chronic hypertension,⁹¹ on which is superimposed a further false BP elevation. Although it has been claimed that pseudohypertension can be suspected in an older person if "Osler's

sign" (while feeling the radial pulse, occlude the brachial artery by cuff inflation or by direct pressure using the other thumb; if the radial artery remains palpable as a firm "tube," the sign is positive) is present,²⁴ the test's usefulness remains debatable.^{25,91,92} For example, in 65 geriatric patients unanimously classified "Osler positive" or "Osler negative" by 3 observers, 6 other physicians demonstrated moderate intraobserver consistency ($\kappa = 0.49$) and only modest interobserver agreement ($\kappa = 0.37$).⁹³ Retaining "Osler-equivocal" patients in the study would almost certainly have further reduced agreement. Confirming pseudohypertension requires an intra-arterial BP measurement; fortunately, the condition is uncommon, affecting less than 2% of an otherwise healthy elderly group.²⁵

Technical Inaccuracies of Indirect BP

Examiner biases include end-digit preference (ie, the tendency to overrecord certain numbers, particularly 0 and 5^{64,75,94,95}), recording lower values at critical diagnostic cut points⁶⁴ presumably to avoid institution of long-term drug treatment, and probably other analogous unconscious processes (eg, "observing" a hoped-for BP reduction consequent to instituting therapy). Physicians may also differ when labeling patients as hypertensive. A group of British general practitioners diagnosed hypertension after only 1 BP measurement in 58% of patients despite previously agreeing to use 3 readings as part of the group's uniform diagnostic criteria.⁹⁶ Contrary to their local expert guidelines, about 37% of German out-of-hospital⁹⁷ physicians and British hospital clinicians⁹⁸ record phase IV (muffling) rather than the more accurate Korotkoff phase V. Perhaps the most common technical error is failure to use a sufficiently large cuff; indeed, in 1 survey, only 25% of primary care physicians even owned a large cuff.⁶³

Interestingly, even when an automatic BP recorder is used, systematic differences between operators in the BP values obtained may remain,⁹⁹ suggesting differing examinee reactions to different examiners, as was seen in one careful study in children.¹⁰⁰

Directional equipment errors can occur. Aneroid instruments often go out of adjustment, usually downward.⁵² One survey found that 34% of practitioners used only aneroid units, of which 30% were off by 10 mm Hg or more.⁶³ A mercury unit can yield biased readings if the meniscus does not rest at 0 or if the mercury's descent is impeded by clogged internal air vents.⁴¹ The stethoscope type seems relatively unimportant.⁴⁹ Although the recommended bell amplifies the Korotkoff sounds' low frequencies in comparison with the diaphragm,¹⁰¹ the risk of exerting excessive pressure and obtaining a falsely low DBP when using the bell²¹ may outweigh the benefit of amplification, particularly in thin patients (try a small bell with a rubber rim).

Examination errors are legion (Table 23-3); most overestimate the true BP. Confirming an apparent difference between arms is not simple because it requires taking the averages of several alternating measurements from both sides or simultaneous measurements by 2 observers who then switch sides and remeasure.¹⁰²

Office BP vs Usual BP

Shortly after patients enter the office, their SBP decreases by several millimeters of mercury, whereas DBP remains rela-

tively constant.^{53,59,60,66,100,103} BP remains fairly steady throughout the customary working daytime hours,⁵⁴ decreases in the evening (ie, at home),^{104,105} and finally decreases another 10% to 20% during sleep.^{106,107} In some patients, BP in a physician's office is notably and consistently higher than daytime ambulatory BP.

This phenomenon, termed *office* or *white coat* hypertension,¹⁰⁸ can even occur during self-measurement of cuff BP in the presence of a physician.²⁶ Approximately 10% to 40% of untreated and nominally borderline hypertensive patients show an appreciable white coat effect,^{27,109} and many treated patients will also show differences of greater than 20/10 mm Hg.^{109,110} The phenomenon may depend in part on patient factors such as sex, age, and BP level.¹¹¹ For example, one group of elderly patients showed an increase in BP of 17/7 mm Hg on entering the physician's office; women showed a greater SBP increase than men.^{28,112} Who wears the white coat seems to matter, because nurses (who, along with technicians, have generally performed the BP measurements used for entry to the large clinical trials) seem to evoke a smaller BP increase than physicians.^{29,113}

THE ISSUE OF PREDICTION

Blood Pressure Now vs Blood Pressure Later

Systematic (and therefore at least partially predictable) changes in BP between visits occur for several reasons. As examinees (volunteers or patients) become more familiar with the examiner, environment, and procedure (including BP self-measurement²⁶), BP decreases by 0 to 7 mm Hg systolic and 2 to 12 mm Hg diastolic.⁵⁹⁻⁶¹ This habituation may be more marked in patients with anxiety trait.¹¹⁴ An additional and probably more important factor,¹¹⁵ regression to the mean, represents the tendency for any unusually high (or low) reading to fall closer to the population mean when repeated. These phenomena are distinguishable from a true "placebo" effect because they can occur in the absence of placebo treatment.^{13,59,116} Some BP changes likely represent currently unappreciated systematic influences; for example, a systematic reduction in BP of about 6 mm Hg occurs during warm vs cold seasons.⁵⁵⁻⁵⁷

Major outcome trials of antihypertensive pharmacotherapy have used 2 to 3 BP readings taken at each of 2 or more visits not only to increase precision (by "averaging out" minute-to-minute and between-day random fluctuations) but also to partially control for regression to the mean and habituation. In practice, following the same multivisit protocol helps ensure that published trial results will be applicable to individual patients. A further refinement, used naturally by many experienced clinicians, is to conduct further follow-up visits when the BP is hovering near a diagnostic cut point. Patients whose true values are far from this threshold (above or below) logically need fewer visits for confident classification.¹¹⁷ In practice, the interval between visits should take into account both the BP level and the patient's clinical status. The Joint National Committee³ recommends remeasurement within 1 month for BP initially 160 to 179 mm Hg

systolic or 100 to 109 mm Hg diastolic (ie, stage 2), within 2 months for stage 1, within 1 week for stage 3, and immediately for stage 4.

Relative Risk of Casual BP Elevation for Persistent Hypertension

Given high random variation, how well does the finding of a single elevated BP predict later definite hypertension? Casual BP, particularly SBP at 1 visit, is predictive of later BP elevation in young men,¹¹⁸ medical students,¹¹⁹ adults,¹³ and children.¹²⁰ (Tracking correlations vary widely, eg, $r = 0.2\text{--}0.7$, depending on the population, technique, and follow-up interval.) In a large prospective study,¹²¹ 1 DBP reading of 90 mm Hg or higher predicted a later definite diagnosis of hypertension in 69% of men and 49% of women; any BP elevation warrants careful follow-up. Looked at the other way, however, about one-third to one-half of subjects with initially elevated BP will ultimately prove not to have hypertension. In practice, regression to the mean guarantees that many individuals with initially elevated BP are actually normotensive.⁷⁵ For example, among subjects with 4 DBP measurements at 2 entry visits averaging between 95 and 104 mm Hg in a mild hypertension trial in Australia,¹²² 28% proved to have an average DBP of less than 90 mm Hg during the next 4 years while receiving placebo. In a careful screening program, similar DBP reductions were observed in the 105 to 114 mm Hg stratum from the first to second screen, and approximately 10% of subjects with DBP greater than or equal to 115 mm Hg were normotensive (< 90 mm Hg) at the next visit.¹¹⁶ Therefore, using the mean of several visits' BP readings improves the ability to predict not only future hypertension¹²⁰ but also CVD sequelae.¹³ Because he may be normotensive, the patient in our case scenario should not be told that he is hypertensive at this initial visit,⁴ but he should be carefully followed up.

Is a High BP Value Ever Normal?

In normotensive subjects, aerobic exercise, which is generally accepted to be good for health, causes SBP to increase moderately, whereas DBP changes little.^{123,124} Because increased BP forms part of the “fight or flight” response, pain (eg, a lacerated finger) and other stresses (eg, pulmonary edema) predictably increase BP, sometimes to extreme values. These reactive elevations of BP do not indicate the presence of “hypertension” if the BP returns to normal levels at rest.

How Do I Improve My Technique?

Checking one's equipment periodically is mandatory to preserve accuracy.¹⁹ Aneroid devices should be recalibrated according to the manufacturer's recommendations. Although one can measure arm circumference in each patient to select an appropriately sized cuff, one can more efficiently mark the limit of arm circumference directly on each cuff by drawing a line in indelible ink at a distance from the free bladder end equal to twice the measured bladder width.

Tape recordings can help standardize observers' identification of Korotkoff sounds.^{84,125} Alternatively, locate a 2-headed

stethoscope (and a second set of ears attached to a willing expert brain) for hands-on training. Initial formal training in the technique of BP measurement is necessary, but in addition, periodic review of technique and retraining as needed are recommended.⁴ Retraining can increase accuracy⁸³ but may be needed every 1 to 2 months for optimal effect,¹²⁶ a frequency probably practical only in research settings. Atrial fibrillation requires a modified technique (discussed earlier). When faced with soft Korotkoff sounds, have the subject elevate the arm and then open and close the fist several times; inflate the cuff, lower the arm (with further inflation as needed), and listen again. In this situation, as permitted by some guidelines,⁴ more rapid deflation after determination of the SBP until the vicinity of the DBP will minimize attenuation of the Korotkoff sounds arising from venous congestion without altering the measured BP.⁵³ Applying the cuff with its tubing emerging at the top¹⁹ will eliminate extraneous noises generated if tubing contacts the stethoscope.

For research purposes, random-zero sphygmomanometers will reduce but still not eliminate observer bias. Fully automatic devices, if otherwise technically accurate, should eliminate certain human foibles (eg, end-digit preference⁹⁴ and selective recording of “desirable” readings). Statistical monitoring⁵⁴ to detect end-digit preference or excessive variability followed by mandatory retraining should be helpful.

In practice, the grossest error, not checking BP at all, remains a common failing even among cardiovascular subspecialists.¹²⁷ Most measurement errors could be obviated if practitioners would only follow the published recommendations^{19,128-132}; alas, many do not.^{63,96-98}

Other Ways to Measure BP

If you cannot hear properly, both SBP and DBP can be determined by palpation to within about 10 mm Hg.¹³³ Palpated SBP is about 7 mm Hg lower than the auscultatory value.¹³⁴

Potential Improvements in the Diagnosis of Hypertension

Elevated BP during aerobic exercise testing in subjects normotensive at rest has some predictive value for subsequent definite hypertension (relative risk from 2.3 to approximately 7).^{123,124,135} Because BP is so variable during daily activity, ambulatory BP monitoring¹³⁶⁻¹³⁸ ought to provide a better estimate of whole-day target organ exposure. Ambulatory BP monitoring correlates better with coexisting target organ damage¹³⁸ and a retrospective follow-up study suggested improved prediction for subsequent CVD.¹³⁹ However, some patients cannot tolerate ambulatory BP monitoring and accurate measurements are not always possible (eg, with marked arrhythmia or obesity).¹⁴⁰ Although appropriate studies have begun,¹⁴¹ no data yet exist to show that adding ambulatory BP monitoring results in clinical benefit, and issues such as cost-effectiveness remain.¹³⁶

Self-measurement of BP¹⁴² is also under active study. Concurrent accuracy,⁸² the meaning of differences in measurements at home and at work,¹⁴³ and concerns about selective reporting remain. When patients bring in their own, usually

lower, home BP readings, be certain to explain that only anti-hypertensive treatment of resting BP readings is of proven value, that daytime BP is routinely higher than evening BP, and that cardiac involvement may relate more closely to work time BP.¹⁴⁴ Although the appeal of self-monitoring includes potentially desirable psychological and compliance effects, any benefit remains questionable¹⁴²; a 1-year trial of home BP monitoring found no difference in treatment, attained BP, or risk factor reduction.¹⁴⁵

THE BOTTOM LINE

Hypertension remains one of the most prevalent and most important public health problems. Measurement of BP has won its place in the recommended periodic health examination because hypertension is common, clinically silent, dangerous, and treatable. Accurately measuring BP by the indirect method requires minimal equipment, combined with a willingness to make the effort; all health care practitioners should read and follow published guidelines.¹⁸ Attention to proper technique plus an appreciation of the inherent variability of BP should yield an accurate diagnosis in most patients. Occasional patients with suspected pseudohypertension or white coat syndrome may benefit from ancillary technology such as echocardiography or ambulatory BP monitoring to optimize diagnostic decision making. Conversely, in the far more common, otherwise low-risk patient, yearly BP readings will suffice to rule out the presence of severe or longstanding untreated hypertension. The patient in our clinical scenario would be served well by a return visit in a few weeks for repeated BP measurement,³ whereas immediately labeling him as hypertensive would be incorrect and, by causing him unnecessary concern, could be an immediate disservice.

Following expert treatment guidelines constitutes the physician's final responsibility, tying a proper diagnosis and proven therapy together to benefit the patient.

Author Affiliations at the Time of the Original Publication

Division of Clinical Pharmacology, Department of Medicine, Clinical Epidemiology Unit, Sunnybrook Health Sciences Centre, University of Toronto, Toronto, Ontario, Canada. Dr Reeves is now with Bristol-Myers Squibb Pharmaceutical Research Institute, Princeton, New Jersey.

Acknowledgment

Dr Reeves was supported by a Career Award from the Pharmaceutical Manufacturers Association of Canada/Medical Research Council of Canada Combined Program.

REFERENCES

- Department of National Health and Welfare. *Main Findings Report of the Canadian Blood Pressure Survey*. Ottawa, Ontario, Canada: Minister of National Health and Welfare; 1989:1.
- Hypertension Detection and Follow-up Program. The Hypertension Detection and Follow-up Program: a progress report. *Circ Res*. 1977;40(suppl I):I-106-I-109.
- Joint National Committee. The fifth report of the Joint National Committee on Detection, Evaluation, and Treatment of High Blood Pressure (JNC V). *Arch Intern Med*. 1993;153(2):154-183.
- Haynes RB, Lacourciere Y, Rabkin SW, et al. Report of the Canadian Hypertension Society Consensus Conference, 2: diagnosis of hypertension in adults. *CMAJ*. 1993;149(4):409-418.
- Zanchetti A, Chalmers JP, Arakawa K, et al. The 1993 guidelines for the management of mild hypertension: memorandum from a WHO/ISH meeting. *Hypertension*. 1993;22:392-403.
- Swales JD, Ramsay LE, Cope JR, et al. Treating mild hypertension: report of the British Hypertension Society Working Party. *BMJ*. 1989;298:694-698.
- Jackson R, Barham P, Bills J, et al. Management of raised blood pressure in New Zealand: a discussion document. *BMJ*. 1993;307(6896):107-110.
- Haynes RB, Sackett DL, Taylor DW, Gibson ES, Johnson AL. Increased absenteeism from work after detection and labeling of hypertensive patients. *N Engl J Med*. 1978;299(14):741-744.
- Lefebvre RC, Hursey KG, Carleton RA. Labeling of participants in high blood pressure screening programs: implications for blood cholesterol screenings. *Arch Intern Med*. 1988;148(9):1993-1997.
- MacMahon SW, Cutler JA, Furberg CD, Payne GH. The effects of drug treatment for hypertension on morbidity and mortality from cardiovascular disease: a review of randomized controlled trials. *Prog Cardiovasc Dis*. 1986;29(3 suppl):99-118.
- Collins R, Petro R, MacMahon S, et al. Blood pressure, stroke, and coronary heart disease, part 2: short-term reductions in blood pressure: overview of randomized drug trials in their epidemiological context. *Lancet*. 1990;335(8693):827-838.
- Kannel WB. Some lessons in cardiovascular epidemiology from Framingham. *Am J Cardiol*. 1976;37(2):269-282.
- Gordon T, Sorlie P, Kannel WB. Problems in the assessment of blood pressure: the Framingham Study. *Int J Epidemiol*. 1976;5(4):327-334.
- Waaler HT. Specificity and sensitivity of blood pressure measurements. *J Epidemiol Community Health*. 1980;34(1):52-58.
- Fujii I, Ueda K, Omae T, et al. Natural history of borderline hypertension in the Hisayama community, Japan, I: the relative prognostic importance of transient variability in blood pressure. *J Chronic Dis*. 1984;37(12):895-902.
- Beard K, Bulpitt C, Mascie-Taylor H, O'Malley K, Sever P, Webb S. Management of elderly patients with sustained hypertension. *BMJ*. 1992;304(6824):412-416.
- Jennings D, Netsky MG. Essential hypertension: a sign in search of a disease. *CMAJ*. 1991;144(8):973-979.
- Perloff D, Grim C, Flack J, et al. Human blood pressure determination by sphygmomanometry. *Circulation*. 1993;88(5 pt 1):2460-2470.
- Petrie JC, O'Brien ET, Littler WA, de Swiet M. Recommendations on blood pressure measurement. *BMJ*. 1986;293(6547):611-615.
- American Society of Hypertension. Recommendations for routine blood pressure measurement by indirect cuff sphygmomanometry. *Am J Hypertens*. 1992;5(4 pt 1):207-209.
- Londe S, Klitzner TS. Auscultatory blood pressure measurement: effect of pressure on the head of the stethoscope. *West J Med*. 1984;141(2):193-195.
- Frohlich ED, Grim C, Labarthe DR, Maxwell MH, Perloff D, Weidman WH. Recommendations for human blood pressure determination by sphygmomanometers: report of a special task force appointed by the steering committee, AHA. *Hypertension*. 1988;11:210a-221a.
- Spence JD, Sibbald WJ, Cape RD. Pseudohypertension in the elderly. *Clin Sci Mol Med*. 1978;55(suppl 4):399s-402s.
- Messerli FH, Ventura HO, Amodeo C. Osler's maneuver and pseudohypertension. *N Engl J Med*. 1985;312(24):1548-1551.
- Kuwajima I, Hoh E, Suzuki Y, Matsushita S, Kuramoto K. Pseudohypertension in the elderly. *J Hypertens*. 1990;8(5):429-432.
- Mengden T, Bättig B, Edmonds D, et al. Self-measured blood pressures at home and during consulting hours: are there any differences? *J Hypertens Suppl*. 1990;8(4):S15-S19.
- Pickering TG, James GD, Boddie C, Harshfield GA, Blank S, Laragh JH. How common is white coat hypertension? *JAMA*. 1988;259(2):225-228.
- Shimada K, Ogura H, Kawamoto A, Matsubayashi K, Ishida H, Ozawa T. Noninvasive ambulatory blood pressure monitoring during clinic visit in elderly hypertensive patients. *Clin Exp Hypertens*. 1990;12(2):151-170.
- Mancia G, Parati G, Pomidossi G, Grassi G, Casadei R, Zanchetti A. Alerting reaction and rise in blood pressure during measurement by physician and nurse. *Hypertension*. 1987;9(2):209-215.
- Kenny RA, Brennan M, O'Malley K, O'Brien E. Blood pressure measurements in borderline hypertension. *J Hypertens*. 1987;5(suppl):483-485.

31. Laughlin KD, Sherrard DJ, Fisher L. Comparison of clinic and home blood pressure levels in essential hypertension and variables associated with clinic-home differences. *J Chronic Dis.* 1980;33(4):197-206.
32. Porchet M, Bussien JP, Waeber B, Nussberger J, Brunner HR. Unpredictability of blood pressures recorded outside the clinic in the treated hypertensive patient. *J Cardiovasc Pharmacol.* 1986;8(2):332-335.
33. Yagi S, Ichikawa S, Sakamaki T, Takayama Y, Murata K. Blood pressure in the paretic arms of patients with stroke. *N Engl J Med.* 1986;315(13):836.
34. Benowitz NL, Kuyt F, Jacob PI. Influence of nicotine on cardiovascular and hormonal effects of cigarette smoking. *Clin Pharmacol Ther.* 1984;36(1):74-81.
35. Robertson D, Wade D, Workman R, Woosley RL. Tolerance to the humoral and hemodynamic effects of caffeine in man. *J Clin Invest.* 1981;67(4):1111-1117.
36. Potter JE, Watson RDS, Skan W, Beevers G. The pressor and metabolic effects of alcohol in normotensive subjects. *Hypertension.* 1986;8(7):625-631.
37. Fagius J, Karhuvaara S. Sympathetic activity and blood pressure increases with bladder distension in humans. *Hypertension.* 1989;14(5):511-517.
38. Lynch JJ, Thomas SA, Long JM, Malinow KL, Friedmann E, Katcher AH. Blood pressure changes while talking. *Isr J Med Sci.* 1982;18(5):575-579.
39. Malinow KL, Lynch JJ, Foreman PJ, Friedmann E, Thomas SA. Blood pressure increases while signing in a deaf population. *Psychosom Med.* 1986;48(1-2):95-101.
40. Conceicao S, Ward MK, Kerr DNS. Defects in sphygmomanometers: an important source of error in blood pressure recording. *BMJ.* 1976;1 (6014):886-888.
41. Shaw A, Deehan C, Lenihan JMA. Sphygmomanometers: errors due to blocked vents. *BMJ.* 1979;1(6166):789-790.
42. Steinfeld L, Alexander H, Cohen ML. Updating sphygmomanometry. *Am J Cardiol.* 1974;33(1):107-110.
43. Russell AE, Wing LMH, Smith SA, et al. Optimal size of cuff bladder for indirect measurement of arterial pressure in adults. *J Hypertens.* 1989;7(8):607-613.
44. Maxwell MH, Waks AU, Schroth PC, Karam M, Dornfeld LP. Error in blood-pressure measurement due to incorrect cuff size in obese patients. *Lancet.* 1982;2(8288):33-36.
45. Webb CH. The measurement of blood pressure and its interpretation. *Prim Care.* 1980;7(4):637-651.
46. Trout KW, Bertrand CA, Williams MH. Measurement of blood pressure in obese persons. *JAMA.* 1956;162(10):970-971.
47. Mitchell PL, Parlin RW, Blackburn H. Effect of vertical displacement of the arm on indirect blood-pressure measurement. *N Engl J Med.* 1964;271:72-74.
48. Neusse WF. The importance of a tight blood pressure cuff. *Am Heart J.* 1956;52(6):905-907.
49. Cushman WC, Cooper KM, Horne RA, Meydreich EF. Effect of back support and stethoscope head on seated blood pressure determinations. *Am J Hypertens.* 1990;3(3):240-241.
50. Viol GW, Goebel M, Lorenz GJ, Ing TS. Seating as a variable in clinical blood pressure measurement. *Am Heart J.* 1979;98(6):813-814.
51. Silverberg DS, Shemesh E, Iaina A. The unsupported arm: a cause of falsely raised blood pressure readings. *BMJ.* 1977;2(6098):1331.
52. Thulin T, Andersson G, Schersten B. Measurement of blood pressure: a routine test in need of standardization. *Postgrad Med J.* 1975;51(596):390-395.
53. King GE. Influence of rate of cuff inflation and deflation on observed blood pressure by sphygmomanometry. *Am Heart J.* 1963;65:303-306.
54. Canner PL, Borhani NO, Oberman A, et al. The Hypertension Prevention Trial: assessment of the quality of blood pressure measurements. *Am J Epidemiol.* 1991;134(4):379-392.
55. Izzo JL, Larrabee PS, Sander E, Lillis LM. Hemodynamics of seasonal adaptation. *Am J Hypertens.* 1990;3(5 pt 1):405-407.
56. Kunes J, Tremblay J, Bellavance F, Hamet P. Influence of environmental temperature on the blood pressure of hypertensive patients in Montreal. *Am J Hypertens.* 1991;4(5 pt 1):422-426.
57. Giacconi S, Palombo C, Genovesi-Ebert A, Marabotti C, Volterrani D, Ghione S. Long-term reproducibility and evaluation of seasonal influences on blood pressure monitoring. *J Hypertens.* 1988;6(suppl 4):S64-S66.
58. Mader SL. Effects of meals and time of day on postural blood pressure responses in young and elderly subjects. *Arch Intern Med.* 1989;149(12): 2757-2760.
59. Armitage P, Rose GA. The variability of measurements of casual blood pressure, I: a laboratory study. *Clin Sci.* 1966;30(2):325-335.
60. Watson RDS, Lumb R, Young MA, Stallard TJ, Davies P, Littler WA. Variation in cuff blood pressure in untreated outpatients with mild hypertension: implications for initiating antihypertensive treatment. *J Hypertens.* 1987;5(2):207-211.
61. Shepard DS. Reliability of blood pressure measurements: implications for designing and evaluating programs to control hypertension. *J Chronic Dis.* 1981;34(5):191-209.
62. Cohn JN. Blood pressure measurement in shock: mechanism of inaccuracy in auscultatory and palpitory methods. *JAMA.* 1967;199(13):118-122.
63. McKay DW, Campbell NR, Parab LS, Chockalingam A, Fodor JG. Clinical assessment of blood pressure. *J Hum Hypertens.* 1990;4(6):639-645.
64. Patterson HR. Sources of error in recording the blood pressure of patients with hypertension in general practice. *BMJ.* 1984;289(6459):1661-1664.
65. Hashimoto F, Hunt WC, Hardy L. Differences between right and left arm blood pressures in the elderly. *West J Med.* 1984;141(2):189-192.
66. van Loo JM, Peer PG, Thien TA. Twenty-five minutes between blood pressure readings: the influence on prevalence rates of isolated systolic hypertension. *J Hypertens.* 1986;4(5):631-635.
67. von Eiff AW, Plotz EJ, Beck KJ, Czernik A. The effect of estrogens and progestins on blood pressure regulation of normotensive women. *Am J Obstet Gynecol.* 1971;109(6):887-892.
68. Polefrone JM, Manuck SB. Effects of menstrual phase and parental history of hypertension on cardiovascular response to cognitive challenge. *Psychosom Med.* 1988;50(1):23-36.
69. Myers MG, Reeves RA. The effect of caffeine on daytime ambulatory blood pressure. *Am J Hypertens.* 1991;4(5 pt 1):427-431.
70. Myers MG, Iazzetta JJ. Intranasally administered phenylephrine and blood pressure. *CMAJ.* 1982;127(5):365-368.
71. Veerman DP, van Montfrans GA, Karemker JM, Wieling W. Inflating one's own cuff does not increase self-recorded blood pressure. *J Hypertens.* 1992;6(suppl 4):S77-S78.
72. Kraus JE, Conley A, Hardy R, Sexton M, Sweezy Z. Relationship of demographic characteristics of interviewers to blood pressure measurements. *J Community Health.* 1982;8(1):3-12.
73. Reeves RA, Lewis JE, Hebert PC, Kerneran Y, Revah A. Gender, status and "white coat" phenomenon in family practice. *Clin Invest Med.* 1991;14(suppl A):A69.
74. Holleman DR, Westman EC, McCrory DC, Simel DL. The effect of sleeved arms on oscillometric blood pressure measurement. *J Gen Intern Med.* 1993;8(6):325-326.
75. Hypertension Detection and Follow-up Program. Variability of blood pressure and the results of screening in the Hypertension Detection and Follow-up Program. *J Chronic Dis.* 1978;31(11):651-667.
76. Di Renzo M, Parati G, Pomidossi G, Veniani M, Pedotti A, Mancia G. Blood pressure monitoring over short day and night times cannot predict 24-hour average blood pressure. *J Hypertens.* 1985;3(4):343-349.
77. Mancia G, Parati G, Pomidossi G, Casadei R, Di Renzo M, Zanchetti A. Arterial baroreflexes and blood pressure and heart rate variabilities in humans. *Hypertension.* 1986;8(2):147-153.
78. Watson RDS, Stallard TJ, Flinn RM, Littler WA. Factors determining direct arterial pressure and its variability in hypertensive man. *Hypertension.* 1980;2(3):333-341.
79. Reeves RA. A review of the stability of ambulatory blood pressure: implications for diagnosis of hypertension. *Clin Invest Med.* 1991;14(3):251-255.
80. Reeves RA, Leenen FHH, Joyner CD. Reproducibility of nurse-measured, exercise and ambulatory blood pressure and echocardiographic left ventricular mass in borderline hypertension. *J Hypertens.* 1992;10(10):1249-1256.
81. Sykes D, Dewar R, Mohanaruban K, et al. Measuring blood pressure in the elderly: does atrial fibrillation increase observer variability? *BMJ.* 1990;300(6718):162-163.
82. O'Brien E, Mee F, Atkins N, O'Malley K. Inaccuracy of seven popular sphygmomanometers for home measurement of blood pressure. *J Hypertens.* 1990;8(7):621-634.
83. Scherwitz LW, Evans LA, Hennrikus DJ, Vallbona C. Procedures and discrepancies of blood pressure measurements in two community health centers. *Med Care.* 1982;20(7):727-738.
84. Neufeld PD, Johnson DL. Observer error in blood pressure measurement. *CMAJ.* 1986;135(6):633-637.
85. White WB, Lund-Johansen P, Omvik P. Assessment of four ambulatory blood pressure monitors and measurements by clinicians versus intra-arterial blood pressure at rest and during exercise. *Am J Cardiol.* 1990;65(1):60-66.

86. Blank SG, West JE, Müller FB, et al. Wideband external pulse recording during cuff deflation: a new technique for evaluation of the arterial pressure pulse and measurement of blood pressure. *Circulation*. 1988;77(6):1297-1305.
87. Taguchi JT, Suwangoor P. Pipe-stem brachial arteries: a cause of pseudohypertension. *JAMA*. 1974;228(6):733.
88. Wallace CT, Carpenter FA, Ewins SC, Mahaffey JE. Acute pseudohypertensive crisis. *Anesthesiology*. 1975;43(5):588-589.
89. Keenan WF. Pseudohypertension mimicking a hypertensive emergency. *JAMA*. 1981;246(10):1088.
90. Saklayen MG. Pseudohypertension in a young woman. *Am J Med*. 1988;84(4):794-795.
91. O'Rourke M, Kelly R. Osler's maneuver and pseudohypertension. *N Engl J Med*. 1985;313:1300.
92. Tiffet CP. Are the days of the sphygmomanometer past? *Arch Intern Med*. 1988;148(3):518-519.
93. Hla KM, Samsa GP, Stoneking HT, Feussner JR. Observer variability of Osler's maneuver in detection of pseudohypertension. *J Clin Epidemiol*. 1991;44(6):513-518.
94. Hla KM, Vokaty KA, Feussner JR. Observer error in systolic blood pressure measurement in the elderly: a case for automatic recorders? *Arch Intern Med*. 1986;146(12):2373-2376.
95. Padfield PL, Jyothinagaram SG, Watson DM, Donald P, McGinley IM. Problems in the measurement of blood pressure. *J Hum Hypertens*. 1990;4(suppl 2):3-7.
96. Smith TD, Clayton D. Individual variation between general practitioners in labelling of hypertension. *BMJ*. 1990;300(6717):74-75.
97. Weiland SK, Keil U, Spelsberg A, et al. Diagnosis and management of hypertension by physicians in the Federal Republic of Germany. *J Hypertens*. 1991;9(2):131-134.
98. Wilkinson LS, Perry IJ, Shinton RA, Beevers DG. An emerging consensus among clinicians on treating mild hypertension but persistent uncertainty as to how blood pressure should be measured. *J R Coll Physicians Lond*. 1991;25(2):116-119.
99. Bruce NG, Cook DG, Shaper AG. Differences between observers in blood pressure measurement with an automatic oscillometric recorder. *J Hypertens Suppl*. 1990;8(4):S11-S13.
100. Burke GL, Webber LS, Shear CL, Zinkgraf SA, Smoak CG, Berenson GS. Sources of error in measurement of children's blood pressure in a large epidemiologic study: Bogalusa Heart Study. *J Chronic Dis*. 1987;40(1):83-89.
101. Abella M, Formolo J, Penney DG. Comparison of the acoustic properties of six popular stethoscopes. *J Acoust Soc Am*. 1992;91(4 pt 1):2224-2228.
102. Sapira JD. The vital signs. In: Sapira JD, ed. *The Art and Science of Bedside Diagnosis*. Baltimore, MD: Urban & Schwarzenberg; 1990:85-104.
103. Wietlisbach V, Rickenbach M, Burnand B, Hausser D, Gutzwiller F. Combining repeated blood pressure measurements to obtain prevalences of high blood pressure. *Acta Med Scand Suppl*. 1988;728:165-168.
104. Millar-Craig MW, Bishop CN, Raftery EB. Circadian variation of blood pressure. *Lancet*. 1978;1(8068):795-797.
105. Gould BA, Hornung RS, Cashman PMM, Raftery EB. Ambulatory blood pressure direct and indirect. In: Weber MA, Drayer JIM, eds. *Ambulatory Blood Pressure Monitoring*. New York, NY: Springer-Verlag NY Inc; 1984:9-19.
106. Littler WA, Honour AJ, Carter RD, Sleight P. Sleep and blood pressure. *BMJ*. 1975;3(5979):346-348.
107. Millar-Craig MW, Mann S, Balasubramanian V, Raftery EB. Blood pressure circadian rhythm in essential hypertension. *Clin Sci Mol Med*. 1978;55:391s-393s.
108. Kleinert HD, Harshfield GA, Pickering TG, et al. What is the value of home blood pressure measurement in patients with mild hypertension? *Hypertension*. 1984;6(4):574-578.
109. Ornstein S, Markert G, Litchfield L, Zemp L. Evaluation of the DINAMAP blood pressure monitor in an ambulatory primary care setting. *J Fam Pract*. 1988;26(5):517-521.
110. Myers MG, Reeves RA. White coat phenomenon in patients receiving antihypertensive therapy. *Am J Hypertens*. 1991;4(10 pt 1):844-849.
111. Silagy CA, McNeil JJ, McGrath BP. Isolated systolic hypertension: does it really exist on ambulatory blood pressure monitoring? *Clin Exp Pharmacol Physiol*. 1990;17(3):203-206.
112. Imai Y, Nakatsuka H, Ikeda M, et al. A cross-sectional survey of home blood pressure in a rural community in northern Japan. *Clin Exp Hypertens*. 1990;12(6):1095-1106.
113. Ellis PJ, Marshall E, Ellis SJ. Which blood pressure? *Lancet*. 1988;2(8616):902-903.
114. McGrady A, Higgins JT Jr. Effect of repeated measurements of blood pressure on blood pressure in essential hypertension: role of anxiety. *J Behav Med*. 1990;13(1):93-101.
115. Haynes RB, Logan AG, Flanagan PT, Milne BJ. The effect of patient "familiarity" with blood pressure assessment on the accuracy of follow-up readings. *J Hypertens*. 1983;1(1):91-94.
116. Hypertension Detection and Follow-up Program Cooperative Group. Blood pressure studies in 14 communities: a two-stage screen for hypertension. *JAMA*. 1977;237(22):2385-2391.
117. Perry HM Jr, Miller JP. Difficulties in diagnosing hypertension: implications and alternatives. *J Hypertens*. 1992;10(8):887-896.
118. Rabkin SW, Mathewson FA, Tate RB. Relationship of blood pressure in 20- to 39-year-old men to subsequent blood pressure and incidence of hypertension over a 30-year observation period. *Circulation*. 1982;65(2):291-300.
119. Thomas CB, Duszynski KR. Blood pressure levels in young adulthood as predictors of hypertension and the fate of the cold pressor test. *Johns Hopkins Med J*. 1982;151(3):93-100.
120. Shear CL, Burke GL, Freedman DS, Berenson GS. Value of childhood blood pressure measurements and family history in predicting future blood pressure status: results from 8 years of follow-up in the Bogalusa Heart Study. *Pediatrics*. 1986;77(6):862-869.
121. Trembath CR, Hickner JM, Bishop SW. Incidental blood pressure elevations: a MIRNET project. *J Fam Pract*. 1991;32(4):378-381.
122. Management Committee. The Australian therapeutic trial in mild hypertension. *Lancet*. 1980;1:1261-1267.
123. Dlin RA, Hanne N, Silverberg DS, Bar-Or O. Follow-up of normotensive men with exaggerated blood pressure response to exercise. *Am Heart J*. 1983;106(2):316-320.
124. Guerrera G, Melina D, Colivicchi F, Santoliquido A, Folli G. Abnormal blood pressure response to exercise in borderline hypertension: a two-year follow-up study. *Am J Hypertens*. 1991;4(3 pt 1):271-273.
125. Rose GA. Standardisation of observers in blood-pressure measurement. *Lancet*. 1965;1(7387):673-674.
126. Bruce NG, Shaper AG, Walker M, Wannamethee G. Observer bias in blood pressure studies. *J Hypertens*. 1988;6(5):375-380.
127. Warshaw LJ. Physicians: check that blood pressure. *JAMA*. 1989;262(13):1775-1776.
128. Campbell NR, Chockalingam A, Fodor JG, McKay DW. Accurate, reproducible measurement of blood pressure. *CMAJ*. 1990;143(1):19-24.
129. O'Brien ET, O'Malley K. ABC of blood pressure measurement: the observer. *BMJ*. 1979;2(6193):775-776.
130. O'Brien ET, O'Malley K. ABC of blood pressure measurement: the sphygmomanometer. *BMJ*. 1979;2(6194):851-853.
131. O'Brien ET, O'Malley K. ABC of blood pressure measurement: the patient. *BMJ*. 1979;2(6195):920-921.
132. O'Brien ET, O'Malley K. ABC of blood pressure measurement: technique. *BMJ*. 1979;2(6196):982-984.
133. Segall HN. A note on the measurement of diastolic and systolic blood pressure by the palpation of arterial vibrations (sounds) over the brachial artery. *CMAJ*. 1940;42:311-313.
134. Scordo K. Using radial artery palpation to monitor blood pressure. *Dimens Crit Care Nurs*. 1983;2(2):75-79.
135. Wilson NV, Meyer BM. Early prediction of hypertension using exercise blood pressure. *Prev Med*. 1981;10(1):62-68.
136. Reeves RA, Myers MG. Ambulatory blood pressure monitoring: an emerging technology. *Clin Invest Med*. 1991;14(3):198-201.
137. Health and Public Policy Committee, American College of Physicians. Automated ambulatory blood pressure monitoring. *Ann Intern Med*. 1986;104(2):275-278.
138. National High Blood Pressure Education Program Coordinating Committee. National High Blood Pressure Education Program Working Group report on ambulatory blood pressure monitoring. *Arch Intern Med*. 1990;150:2270-2280.
139. Perloff D, Sokolow M, Cowan R. The prognostic value of ambulatory blood pressures. *JAMA*. 1983;249(20):2792-2798.

140. Reeves RA. Patient and environmental factors affecting ambulatory blood pressure monitoring. *Clin Invest Med.* 1991;14(3):218-223.
141. Clement DL; Steering Committee. Office versus ambulatory recordings of blood pressure (OvA): a European multicenter study. *J Hypertens Suppl.* 1990;8(6):S39-S41.
142. Canadian Coalition for High Blood Pressure Prevention and Control. Recommendations on self-measurement of blood pressure. *CMAJ.* 1988;138(12):1093-1096.
143. Pickering TG, James GD. Some implications of the differences between home, clinic and ambulatory blood pressure in normotensive and hypertensive patients. *J Hypertens Suppl.* 1989;7(3):S65-S72.
144. Devereux RB, Pickering TG, Harshfield GA, et al. Left ventricular hypertrophy in patients with hypertension: importance of blood pressure response to regularly recurring stress. *Circulation.* 1983;68(3):470-476.
145. Midanik LT, Resnick B, Hurley LB, Smith EJ, McCarthy M. Home blood pressure monitoring for mild hypertensives. *Public Health Rep.* 1991;106(1):85-89.

Prepared by David L. Simel, MD, MHS
Reviewed by Ben Stern, MS, DPT, and Sheri Keitz, MD, PhD

CLINICAL SCENARIO

A 47-year-old woman with a strong family history of heart disease reports that her blood pressure (BP) measured on commercial store devices is sometimes, but not always, as high as 140 to 145 mm Hg systolic. She feels well and is physically active. Your nurse obtained her BP of 147/82 mm Hg with your office automated oscillometric device. Given the values she reports and your office measure, have you diagnosed her as having stage 1 hypertension?

UPDATED SUMMARY ON BP MEASUREMENT TO DETECT HYPERTENSION IN ADULTS

Original Review

Reeves RA. Does this patient have hypertension? how to measure blood pressure. *JAMA*. 1995;273(15):1211-1218.

UPDATED LITERATURE SEARCH

Because we were aware of systematic reviews on BP diagnosis and management, our literature search focused on formal systematic reviews of adult hypertension published since 2000. Initially, we crossed the search terms “blood pressure determination/methods” and “hypertension/diagnosis” filtered for “human,” English articles that arose from “consensus development conferences” or were “academic reviews.” The results yielded 36 titles but contained obvious omissions. We next used the SUMSearch strategy (<http://sumsearch.uthscsa.edu/>; accessed May 30, 2008) in PubMed for the search term *blood pressure*, limited to physical examination from 2000-2004; this yielded 111 articles and included the publications from well-known US, Canadian, European, and British consensus groups for the evaluation and management of hypertension. Each of the consensus groups used a formal systematic approach to evaluate the evidence for BP measurement techniques. The consensus groups all published updates during 2003-2004; therefore, we focused our review on these 4 groups and the references from those reports that specifically addressed BP measurement. The independent groups used high methodologic standards, and the recommendations are similar. Therefore, we present the summary

data from these without individual reviews of each society’s recommendations.

NEW FINDINGS

- The classification of BP has now been changed to “normal,” “prehypertension” or “high normal,” stage 1 hypertension, and stage 2 hypertension (Tables 23-4 and 23-5).
- Aneroid sphygmomanometers are accurate, but only when they are calibrated at least once yearly and the examiners use proper measurement techniques. Sphygmomanometers should be calibrated to their manufacturer’s specifications.
- Self-measured BP can be used as part of the patient’s medical history, but the thresholds for diagnosing hypertension are lower (> 135 mm Hg systolic or > 85 mm Hg diastolic).

Details of the Update

The techniques for measuring BP are presented well in the original article and are unchanged. A systematic review of BP measurement was presented in a series of articles published in the *British Medical Journal*¹; one article highlighted the errors in measurement that were also reviewed in The Rational Clinical Examination article.²

IMPROVEMENTS IN THE DATA PRESENTED IN THE ORIGINAL PUBLICATION

Self-measurement of BP had not been adequately studied when The Rational Clinical Examination article was pub-

Table 23-4 Classification of Blood Pressure (JNC-VII)³

	SBP, mm Hg	DBP, mm Hg
Normal	<120	<80
Prehypertension	120-139	Or 80-89
Stage 1 hypertension	140-159	Or 90-99
Stage 2 hypertension	≥160	Or ≥100

Abbreviations: DBP, diastolic blood pressure; JNC-VII, seventh report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure; SBP, systolic blood pressure.

Table 23-5 Classification of Blood Pressure (Canadian, British, and European Societies)

Classification	SBP, mm Hg		DBP, mm Hg
Optimal	<120	And	<80
Normal	<130	And	<85
High normal	130-139	Or	85-89
Grade 1 hypertension	140-159	Or	90-99
Grade 2 hypertension	160-179	Or	100-109
Grade 3 hypertension	≥180	Or	≥110

Abbreviations: DBP, diastolic blood pressure; SBP, systolic blood pressure.

lished. There is now an established role for using properly obtained self-measured BP values. As such, the patient's report can be considered part of the medical history when assessing for hypertension.

CHANGES IN THE REFERENCE STANDARD

Indirect BP measurement through auscultation of Korotkoff sounds is the pragmatic reference standard for clinical care and research. BP measurement with oscillometric techniques is acceptable for following treatment in patients with established hypertension, but the initial diagnosis should be confirmed with auscultatory methods. This may be especially important in elderly patients or those with arrhythmias. The seventh report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure (JNC-VII)³ recommends the classification of BP for adults as shown in Table 23-4.

The Canadian Hypertension Education Program,⁴ British Hypertension Society,⁵ and European Society of Hypertension⁶ avoid labeling patients as "prehypertensive." Instead, they describe "high-normal" BP and recommend that this group receive more frequent monitoring, given their higher risk of developing hypertension (Table 23-5).

RESULTS OF LITERATURE REVIEW

The principles of BP measurement remain the same:

1. Auscultatory methods with a properly calibrated device are the reference standard.
2. Patients should be seated quietly for 5 minutes in a chair, with their feet on the floor and the arm supported at heart level.
3. The cuff's bladder should encircle at least 80% of the arm.
4. The systolic pressure is the point at which the first of 2 or more consecutive heart sounds is heard (phase 1). The

diastolic pressure is the point at which the sounds disappear (phase 5). If sounds are heard all the way to 0 mm Hg, the point of diastolic muffling (phase 4) is used as the diastolic pressure.

Several points are worth observing from these recommendations. First, devices should be calibrated at least once a year, which is especially important because mercury sphygmomanometers are being replaced by aneroid or oscillometric devices. Aneroid devices use a spring mechanism that is subject to wear and can cause inaccurate readings. However, when properly maintained, aneroid sphygmomanometers are accurate and underestimate reference devices by a mean of only 0.5 mm Hg.⁷

Second, once the proper equipment is obtained (ie, a calibrated device and an appropriately sized cuff), the exact procedure must be followed. The 4 points listed above highlight important potential technical errors that are avoidable when the proper procedure is followed. The patient must be *seated* (rather than supine, which can increase the systolic pressure 3 mm Hg). The arm must be *supported* by an armrest or supported by the examiner so that the patient does not create effort in elevating the arm (a potential 2 mm Hg increase in pressure). Finally, the arm must be *at the level of the heart* rather than dangling (a potential 10 mm Hg increase in systolic pressure).

Third, diagnosis when the BP approaches the threshold values should never be based on a single measure. At a single visit, at least 2 measures should be taken. Because of biological variability and white coat hypertension, patients should return to the clinic for additional measures when the diagnosis is not obvious. At the initial diagnosis, experts suggest measuring the BP in both arms. Patients with large discrepancies (eg, >20 mm Hg systolic or >10 mm Hg diastolic) need further assessment.

Self-monitoring of BP can be used for both diagnosis and treatment monitoring. As such, the results should be considered as part of the patient's medical history. When used to diagnose hypertension, a mean self-recorded BP greater than 135 mm Hg systolic or 85 mm Hg diastolic should be considered hypertensive.⁸ The patient should use a fully automated monitor with an appropriate-sized arm cuff. The physician should discard the first day of patient reading and then use all other data to calculate the mean BP.⁹ Randomized trials of home monitoring for BP control used a frequency of twice-daily to twice-weekly recordings.¹⁰

EVIDENCE FROM GUIDELINES

All guidelines emphasize the need for correctly measured BPs. The US Preventive Services Task Force recommends that a diagnosis be established only after 2 or more elevated measures on at least 2 visits during at least 1 week.¹¹

Special Populations

Two populations of adults deserve special mention because BP measurement may be misleading. Elderly patients may have greater BP variability than younger patients, and they may develop decreased arterial compliance that creates the phenomenon of pseudohypertension. These patients may have an unidentifiable phase V. Because systolic hypertension is so important and the prevalence of hypertension is so high in the elderly, additional testing with self-monitored BP or ambulatory BP measures may be needed.

Patients with irregular arrhythmias can display large beat-to-beat BP liability. Automated devices may be particularly “confused” by the variability, so all patients with arrhythmia should have their BP confirmed with indirect auscultation.

CLINICAL SCENARIO—RESOLUTION

Using JNC-VII standards, it is highly likely that she at least has prehypertension and perhaps stage 1 hypertension. However, the initial diagnosis should be established with more certainty. Your nurse reported a single value obtained with your office automated cuff; you should confirm that proper techniques were used and whether the measure was repeated. Your nurse may have repeated the BP and recorded only the lower of 2 values; that would underestimate the BP because the mean value should be used. At the initial diagnosis, indirect auscultation is necessary (usually with a calibrated aneroid sphygmomanometer). If you are not sure whether your office cuffs have been calibrated, assigning one of your office staff responsibility for calibration at the manufacturer's recommended interval is an important quality measure. You should repeat the patient's BP measurement yourself, making sure that you follow the appropriate principles of measurement (5-minute rest, correct cuff size, arm supported, arm at the level of the heart). The patient should return in 1 to 2 weeks for a repeated measurement, or you might opt for self-measured home BPs to establish the diagnosis.

ADULT HYPERTENSION—MAKE THE DIAGNOSIS

PRIOR PROBABILITY

Approximately 25% of all US adults have hypertension. The prevalence is much greater at older ages. The JNC-VII review observed that a normotensive 55-year-old individual has a 90% lifetime risk of developing hypertension. More than half of patients aged 60 to 69 years have hypertension, with the estimate increasing to 75% by age 70 years.

POPULATION FOR WHOM HYPERTENSION SHOULD BE CONSIDERED

All persons older than 18 years.

DETECTING THE LIKELIHOOD OF HYPERTENSION

Indirect auscultation is the reference standard for detecting hypertension. Because prehypertension and grade 1 hypertension produce no symptoms, there are no screening tests and there is a universal recommendation to evaluate all adults for high BP at least every 2 years. Patients with prehypertensive values should be monitored more frequently. Self-monitored BPs may be used for diagnosis when the patient uses appropriate measurement techniques.

Despite proper measurement techniques, inaccurate results at a single visit may be attributed to biologic variability or white coat hypertension. Thus, patients should have values repeated at several visits to confirm stage 1 hypertension. Assessment of white coat hypertension may use self-monitored BP measurement. In addition, continuous ambulatory BP measurement may be obtained as an additional diagnostic test.

REFERENCE STANDARD TESTS

Indirect auscultation with a mercury or well-calibrated aneroid sphygmomanometer provides the pragmatic reference standard. Oscillometric measures for diagnosis should be confirmed with indirect auscultation. Self-monitored BPs may be used, although the threshold should be less than 135 mm Hg systolic and less than 85 mm Hg diastolic.

REFERENCES FOR THE UPDATE

1. McAlister FA, Straus SE. Measurement of blood pressure: an evidence based review. *BMJ*. 2001;322(7292):908-911.
2. Beevers G, Lip GYH, O'Brien E. Blood pressure measurement, part I: sphygmomanometry: factors common to all techniques. *BMJ*. 2001;322(7292):981-985.
3. Chobanian AV, Bakris GL, Black HR, et al. The seventh report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure. *JAMA*. 2003;289(19):2560-2572.^a
4. Hemmelgarn BR, Zarnke KB, Campbell NRC, et al; Canadian Hypertension Education Program, Evidence-Based Recommendations Task Force. The 2004 Canadian Hypertension Education Program recommendations for the management of hypertension: part I—blood pressure measurement, diagnosis and assessment of risk. *Can J Cardiol*. 2004;20(1):31-40.
5. Williams B, Poulter NR, Brown MJ, et al, for the British Hypertension Society. British Hypertension Society guidelines for hypertension management 2004 (BHS-IV): summary. *BMJ*. 2004;328(7440):634-640.
6. European Society of Hypertension—European Society of Cardiology Guidelines Committee. 2003 European Society of Hypertension—European Society of Cardiology guidelines for the management of arterial hypertension. *J Hypertens*. 2003;21(6):1011-1053.
7. Canzanello VJ, Jensen PL, Schwartz GL. Are aneroid sphygmomanometers accurate in hospital and clinic settings? *Arch Intern Med*. 2001;161(5):729-731.
8. Thijss L, Staessen JA, Celis H, et al. Reference values for self-recorded blood pressure. *Arch Intern Med*. 1998;158(5):481-488.
9. O'Brien E, Beevers G, Lip GYH. Blood pressure measurement, part IV: automated sphygmomanometry: self blood pressure measurement. *BMJ*. 2001;322(7295):531-536.
10. Cappuccio FP, Kerry SM, Forbes L, Donald A. Blood pressure control by home monitoring: a meta-analysis of randomized trials [published correction appears in *BMJ*. 2004;329(7464):499]. *BMJ*. 2004;329(7458):145-151.
11. US Preventive Services Task Force. Screening for high blood pressure: US Preventative Services Task Force reaffirmation recommendation statement. *Ann Intern Med*. 2007;147(11):783-786.

^aFor the Evidence to Support the Update for this topic,
see <http://www.JAMAEvidence.com>.

EVIDENCE TO SUPPORT THE UPDATE:

23

Hypertension

TITLE The Seventh Report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure.

AUTHORS Chobanian AV, Bakris GL, Black HR, et al.

CITATION JAMA. 2003;289(19):2560-2572.

(Note: A complete version of the report appears in *Hypertension*. 2003;42(6):1206-1252.)

QUESTION Is the case definition for blood pressure (BP) appropriate?

DESIGN Formal, systematic review.

DATA SOURCES MEDLINE.

STUDY SELECTION AND ASSESSMENT Study population consisted of adults. Articles reviewed were published between January 1997 and April 2003. Quality schemata used in previous iterations of the Joint National Committee on Prevention, Detection, Evaluation and Treatment of High Blood Pressure were used.

MAIN RESULTS

Patients with BP of 130/80 mm Hg to 139/89 mm Hg are at twice the risk of developing hypertension compared with those with lower BP. Thus, the patients in this 130 to 139 mm Hg systolic range were reclassified to a category called prehypertension.

Properly calibrated instruments must be used. The method can be simplified to a few key points: (1) patients should sit quietly for at least 5 minutes before the BP is measured, (2) the patients should be seated with their arm supported at heart level, (3) a cuff that encircles at least 80% of the arm should be used, and (4) 2 measures should be obtained. The systolic BP is the point at which the first of 2 or more Korotkoff sounds is heard. The diastolic BP is the point at which the Korotkoff sounds are last heard.

The committee supported the use of patient self-measured BP, observing that a mean value that is higher than 135/85 mm Hg should be accepted as defining hypertension. Similarly, patients with home measures consistently less than 130/80 mm Hg and who lack target organ disease, despite increased office measures, do not meet criteria for hypertension.

CONCLUSIONS

LEVEL OF EVIDENCE Systematic review.

STRENGTHS A large number of participations in an ongoing review of the evidence base for the treatment of hypertension.

There were no changes in the recommendations for BP technique performed in the examination room. The use of self-monitored BP as part of the patient's medical history requires that the patient's monitoring device be accurate and that the patient use proper technique.

Reviewed by David L. Simel, MD, MHS

This page intentionally left blank

CHAPTER 24

Is This Adult Patient Hypovolemic?

Steven McGee, MD

William B. Abernethy III, MD

David L. Simel, MD, MHS

CLINICAL SCENARIOS

In each of the following clinical scenarios, clinicians need to identify which physical signs reliably and accurately indicate volume depletion or dehydration.

CASE 1 A 54-year-old man, taking ibuprofen for knee arthritis, presents with a 1-day history of melena. Physical examination reveals a pulse of 80/min and blood pressure (BP) of 140/82 mm Hg when supine and 115 and 132/86 mm Hg when standing. There is mild epigastric tenderness and a positive result on a guaiac test for occult blood in the stool. The hematocrit level is 39%.

CASE 2 A 62-year-old woman had 6 months of episodic vertigo and unilateral hearing loss, attributed to Ménière disease. She began hydrochlorothiazide, but 3 weeks later her dizziness is slightly worse. Her heart rate is 80/min and BP is 160/84 mm Hg when supine, 88 and 134/72 mm Hg when standing. On standing, she experiences slight dizziness.

CASE 3 An 82-year-old nursing home resident presents to the emergency department with a 1-day history of nausea and vomiting. Her underlying medical problems include dementia, coronary artery disease, atrial fibrillation, emphysema, and hypertension. She has been treated with aspirin, isosorbide dinitrate, furosemide, β -agonist inhalers, and lisinopril. The clinician diagnoses viral gastroenteritis or food poisoning because other members of the nursing home have an identical illness. On examination, the patient is afebrile and alert and demonstrates normal speech and strength. Her mental status is no different from her baseline. The pulse is 75/min and the BP is 154/90 mm Hg supine and 90 and 130/76 mm Hg when upright. The tongue, mucous membranes, and axillae are moist. Results of an examination of the heart, lungs, abdomen, and an electrocardiogram are normal.

WHY IS CLINICAL EXAMINATION IMPORTANT?

The term *volume depletion* describes the loss of sodium from the extracellular space (intravascular and interstitial fluid) that occurs after gastrointestinal (GI) hemorrhage, vomiting, diarrhea, and diuresis. In contrast, the term *dehydration* refers to losses of intracellular water that ultimately cause cellular desiccation which elevates the plasma sodium concentration and osmolality.¹ This distinction is important to clinicians (patients with volume depletion exhibit prominent circulatory instability and should receive 0.9% saline rapidly, whereas those with pure dehydration may lack circulatory instability and should receive 5% dextrose, usually more slowly). Most patients presenting with dehydration, however,

also have volume depletion. Moreover, in most clinical studies of related physical findings, investigators lump the 2 disorders together by using as a combined criterion standard either the presence of an elevated serum urea nitrogen/creatinine ratio (a measure of volume depletion) or an elevated serum sodium level or osmolality (a measure of dehydration). We will use the term *hypovolemia* to collectively refer to both conditions.

GI tract hemorrhage, the prototype of volume depletion, is a common and important problem. Hospitalizations for upper GI tract hemorrhage occur in 150/100 000 population per year² and are associated with a case fatality rate of 3% to 10%.^{2,3} Hypernatremia, the hallmark of dehydration, affects primarily elderly patients with infections and poor access to water, accounting for less than 1% of hospital admissions but associated with a mortality rate exceeding 40%.^{4,5} Risk factors for hypovolemia in the elderly include female sex, age older than 85 years, having more than 4 chronic medical conditions, taking more than 4 medications, and being confined to bed.⁶

Clinical examination attempts to address (1) whether the patient's symptoms are related to hypovolemia and (2) the degree of hypovolemia. In case 1, symptoms and laboratory data do not gauge the severity of the GI tract hemorrhage. For example, the presence of melena has been associated with both insignificant (as little as 100 mL of blood loss)⁷ and massive hemorrhage.⁸ The admission hematocrit level also correlates poorly with the degree of blood loss and overall mortality,³⁻¹⁰ especially in cases of persistent or recurrent bleeding, because a decrease in hematocrit is often delayed 24 to 72 hours after hemorrhage.¹¹⁻¹³ In one large study,³ however, postural vital signs were a significant

univariate predictor of mortality and complications. How accurate are postural vital signs and which component of the postural change, pulse or BP, provides more meaningful information?

In case 2, the clinician recognizes that hydrochlorothiazide may benefit patients with Ménière disease¹⁴ but also wonders if the diuretic caused volume depletion and aggravated her dizziness. How significant is the postural decrease in systolic BP of 26 mm Hg and the mild postural dizziness?

Finally, case 3 differs from case 1 in that the fluid losses are not directly from the vascular space and that emesis typically has only one-third the sodium concentration of serum. How reliable are findings of postural vital signs, capillary refill, and moist axilla, tongue, and mucous membranes in this patient?

METHODS

Using the MEDLINE database for articles from January 1966 to November 1997, an author (S.M.) used 3 search strategies, all limited to the English language and to humans 16 years or older, to retrieve all relevant publications on the bedside diagnosis of hypovolemia. The first strategy used the search terms "dehydration/di" or "hypotension, orthostatic" or "tilt table test." The second strategy used "exp dehydration" or "exp hypotension, orthostatic" or "exp heart rate" and "exp physical examination" or "exp medical history taking" or "exp professional competence" or "exp 'sensitivity and specificity'" or "reproducibility of results" or "observer variation" or "diagnostic tests, routine" or "exp decision support techniques" or "Bayes theorem." Finally, textword searches were completed for "skin turgor" or "acute blood loss" or "orthostatic vital signs or (postural and pulse)." According to review of titles and abstracts, relevant publications were retrieved. To complete the search, this author reviewed the bibliographies of these articles and those of textbooks on physical diagnosis. Studies on the physical diagnosis of hypovolemia in infants and children were not included in this review.¹⁵⁻²³

Two types of studies are presented. The first group (**Table 24-1**) investigated the postural vital signs and capillary refill time in healthy volunteers, some of whom underwent phlebotomy of up to 1150 mL of blood. Despite their limitations, these studies are included because they are the only studies that compare physical signs with objective measurements of blood loss. A second set (**Table 24-2**) included patients presenting to emergency departments with suspected hypovolemia, usually caused by vomiting, diarrhea, or decreased oral intake. Two authors (S.M. and W.B.A.) independently graded these studies A, B, or C, according to the criteria that appear in the footnote of Table 24-2. There was complete agreement regarding classification.

A random-effects model was used to generate summary measures and confidence intervals (CIs).^{37,38} The model was appropriate because the studies of pulse and pressure change in normovolemic individuals were representative of all such investigations and included a broad mix of relevant subjects. For studies of diagnostic accuracy, the random-effects summary

Table 24-1 Phlebotomy Studies in Normovolemic Individuals^a

Source, y	No. of Subjects	Amount of Blood Removed, mL
Moderate ^b		
Knopp et al, ²⁴ 1980	44	500
Baraff and Schriger, ²⁵ 1992	100	450
Witting et al, ²⁶ 1994	292	450
	44	450
Ralston et al, ²⁷ 1961	16	530-590
Warren et al, ²⁸ 1945	8	4.1-8.5 per kg
Large ^b		
Knopp et al, ²⁴ 1980	44	1000
Shenkin et al, ²⁹ 1944	11	1029 ± 81
Wallace and Sharpey-Schafer, ³⁰ 1941	25	9-16 per kg
Skillman et al, ³¹ 1967	9	764 ± 93
Bergenwald et al, ³² 1977	16	900
Ralston et al, ²⁷ 1961	3	920

^aMean age range of participants in these studies was 25 to 44 years. The exceptions are Warren et al,²⁸ who did not provide age information, and Witting et al,²⁶ who had 292 subjects who were younger than 65 years and 44 subjects who were older than 65 years.

^bModerate was defined as 450 to 630 mL; large, 630 to 1150 mL.

Table 24-2 Clinical Studies of Hypovolemia^a

Source, y	Grade of Study ^a	No. of Subjects	Age, Mean (Range), y	Patient Population	Physical Finding	Criterion Standard for Hypovolemia	Reason Study Not Grade A
Eaton et al, ³³ 1994	A	86	80 (70-98)	Patients older than 70 y admitted with acute medical conditions	Dry axilla	Serum urea nitrogen/creatinine ratio > 25 or plasma osmolarity > 295 mmol/kg H ₂ O	
Gross et al, ³⁴ 1992	B	38	82 ^b (61-98)	Stable patients older than 60 y in the emergency department with suspected fluid and electrolyte problems	Dry mucous membranes, dry tongue, tongue furrows, confusion, weakness, nonfluent speech, sunken eyes	Elevated serum urea nitrogen/creatinine ratio, serum osmolality, or serum sodium n < 50	
Johnson et al, ³⁵ 1995	C	23	NA (18-31)	Pregnant women in the emergency department with hyperemesis gravidarum and normal serum electrolyte and creatinine levels	Postural vital signs	≥5% Weight gain after rehydration	Convenience sample
Schriger and Baraff, ³⁶ 1991	C	32	44 (17-90)	Adults in the emergency department with decreased oral intake, vomiting, diarrhea, or blood loss, ^c and suspected hypovolemia	Capillary refill time	Hypotension or postural pulse increment > 20 beats/min or diastolic blood pressure decrement > 15 mm Hg	Criterion standard was postural vital signs or hypotension; question blinding from criterion standard

Abbreviation: NA, not available.

^aGrading was determined by these traits: A, an independent, blind comparison of a defined physical sign with an acceptable criterion standard of hypovolemia in more than 50 consecutive patients suspected of having hypovolemia; B, same traits as A but there were fewer than 50 consecutive patients suspected of having volume depletion; C, all other studies, including those using a criterion standard of uncertain validity, a physical finding not clearly defined, a comparison that was not blinded, or a selection of patients dependent on either the physical finding or criterion standard. An acceptable criterion standard was a chemical measure (either elevated serum sodium, osmolality, or blood urea nitrogen ratio, or blood urea nitrogen/creatinine ratio) or percentage of weight gain after the patient had received parenteral fluids. (See Table 1-7 for a summary of Evidence Grades and Levels.)

^bMedian age instead of mean age.

^cTotal number of subjects with blood loss equaled 6.

measures provided suitable benchmarks for clinicians' use in actual practice and avoided errors when testing for homogeneity among a number of investigations. Calculations of sensitivity and specificity were derived from graphs or tabulated data that appeared in the original articles or were available from the authors of the studies.^{26,34,35} Those phlebotomy studies that described their results only as mean and standard deviations of the postural change in heart rate and BP, before and after phlebotomy, were reviewed but excluded from the calculations of sensitivity and specificity.³⁹⁻⁴² We used the method of Simel et al⁴³ to calculate CIs for the likelihood ratios (LRs).

RESULTS

Postural Vital Signs

When obtaining postural vital signs, clinicians should wait 2 minutes before measuring the supine vital signs and 1 minute after patients stand before measuring the upright vital signs, according to investigations of healthy individuals discussed below. Having the patient sitting instead of standing markedly reduces the clinician's ability to detect the postural changes induced by blood loss.²⁴ Clinicians who count the pulse for 30 seconds and double the result are more accurate than those using only 15 seconds.⁴⁴

Within 1 to 2 minutes after the patient stands up from the supine position, about 7 to 8 mL/kg of blood shifts to the lower body, causing the thoracic blood volume, stroke volume, and cardiac output to decrease and circulating norepinephrine levels and systemic vascular resistance to increase.^{40,41,45-48} Table 24-3 presents data from 25 studies that investigated the postural vital signs of more than 3500 normovolemic individuals during tilt tests (moving from supine to upright positions by active standing was used in 97% of patients, tilt table testing in 3% of patients). The most prominent finding is an increment in heart rate of 11/min (95% CI, 8.9-13/min). This increase usually stabilizes after 45 to 60 seconds with the patient in the upright position.^{24,45,52,54} The systolic BP decreases slightly by 3.5 mm Hg (95% CI, -1.5 to -5.5 mm Hg), stabilizing 1 to 2 minutes after standing,^{45,54} whereas the diastolic BP increases by 5.2 mm Hg (95% CI, 2.8-7.6 mm Hg).

The variability of the postural pulse increment observed in these studies is in part attributable to the patients' ages and perhaps to the physical examination method. In Table 24-3, the mean age from each study correlates inversely with the observed mean pulse increment ($r, -0.50; P = .02$) (Table 24-3). Other studies^{47,56,57,61,62} also confirm that as patients age, the pulse increment becomes smaller, although no obvious cut point exists that allows the clinician to stratify patients. The duration of supine rest before the patient stands might also

Table 24-3 Postural Change in Vital Signs of Normovolemic Adults^a

Source, y	No. of Subjects	Age, Mean (Range), y	Pulse Change, beats/min (SD) ^a	Systolic Blood Pressure Change, mm Hg (SD)	Diastolic Blood Pressure Change, mm Hg (SD)
Tell et al, ⁴⁹ 1988	916	... ^b (14-16)	+14.0 (14.8)	-5.2 (8.6)	+10.2 (12.4)
Honda et al, ⁵⁰ 1977	496	16.5 (...)	+8.3 (8.8)	-9.5 (9.4)	+3.7 (11.9)
Horam and Roscelli, ⁵¹ 1992	34	... (16-19)	+18.4 (8.4)	+0.8 (7.4)	+8.3 (6.7)
Borst et al, ⁵² 1982	10	21 (...)	+15 (13)	+2 (5)	+19 (8)
Kaijser and Sachs, ⁵³ 1985	14	... (20-26)	+16 (8)	-1 (6)	+4 (5)
Moore and Newton, ⁵⁴ 1986	50	... (25-35)	+12.6 (11.7)	-12.1 (7.4)	-3.5 (6.1)
Baraff and Schriger, ²⁵ 1992	104	32 (...)	+2 (7)	-2 (6)	+4 (7)
Green and Metheny, ³⁹ 1947	25	32 (18-46)	+9 (7)	-5 (8.1)	+12 (9.3)
Currens, ⁵⁵ 1948	1000	33.2 (...)	+13.2 (...)
Koziol-McLain et al, ⁵⁶ 1991	132	34.1 (...)	+17.2 (11.1)	+2.8 (11.4)	+9.2 (7.8)
Knopp et al, ²⁴ 1980	79	36 (17-55)	+18.4 (...)	-2.8 (...)	+16.4 (...)
Tuckman and Shillingford, ⁴⁸ 1966	9	37 (...)	+13 (12)	+1 (8)	+7 (7)
Streeten et al, ⁴⁶ 1988	92	... (18-64)	+12.3 (4.8)	-6.5 (4.8)	+5.6 (3.8)
Wong et al, ⁴⁰ 1989	27	41.4 (...)	+14.6 (5.7)
Kaijser and Sachs, ⁵³ 1985	18	42.5 (38-47)	+13 (8)	-2 (8)	+6 (8)
Dambrink and Wieling, ⁵⁷ 1987	10	... (60-69)	+10 (9.5)	-2 (12.6)	+9 (9.5)
Kaijser and Sachs, ⁵³ 1985	15	67 (...)	+11 (8)	+9 (20)	+3 (13)
Dambrink and Wieling, ⁵⁷ 1987	10	... (70-79)	+11 (6.3)	-9 (12.6)	+2 (3.2)
Baraff and Schriger, ²⁵ 1992	96	76.5 (...)	+1 (7)	-5 (12)	-2 (7)
Green and Metheny, ³⁹ 1947	13	80 (...)	+2 (5.5)	-9 (12)	-5 (7.7)
Dambrink and Wieling, ⁵⁷ 1987	10	... (80-89)	+8 (3.2)	-5 (9.5)	+4 (6.3)
Lipsitz et al, ⁵⁸ 1985	15	87 (...)	+12 (7.5)	-3 (16)	...
Levitt et al, ⁵⁹ 1992	21	...	+6.8 (7.8)	-2.5 (8.0)	+5.3 (9.9)
Schneider and Truesdell, ⁶⁰ 1922	144	...	+13.8 (7.1)
Schneider and Truesdell, ⁶⁰ 1922	204	...	+12.5 (8.5)	+5.3 (...)	...
Summary measure ^c	NA	NA	+11 (8.9-12.8)	-3.5 (-1.5, -5.5)	+5.2 (2.8-7.6)

Abbreviations: CI, confidence interval; NA, not available.

^aValues are expressed as upright minus resting supine value.

^bEllipses indicate data not available.

^cExpressed as random effects summary measure (95% CI).

affect the variability of the pulse change, according to the one outlier study in Table 24-3,²⁵ which demonstrated a mean postural pulse increment of only 2/min and used the shortest time of supine rest before having the patient stand (only 1 minute; all other studies waited at least 2 minutes). Longer periods of supine rest before standing may produce a greater immediate pulse increment, perhaps by causing a greater transfer of blood to the legs and decrement in cardiac output.^{48,63} Aside from the patient's age and period of supine rest, however, no other trend was evident. There was no clear relationship between the postural pulse increment and period of supine rest beyond 2 minutes, resting supine pulse rate, time upright before vital signs measurement (all > 45 seconds), technique of pulse measurement (palpation vs automated), setting of the study (emergency department, prephlebotomy vs other), or method of assuming the upright posture (active stand vs tilt table).

According to the studies in Table 24-3 that enrolled more than 25 individuals and presented tabulated data (n = 774), the specificity of a postural pulse increment of 30/min or

more (ie, the most common threshold used in clinical studies) was 96% (95% CI, 92%-98%).

Postural hypotension, defined as a decrement in systolic BP of more than 20 mm Hg after standing from the supine position, occurs in up to 10% of normovolemic individuals younger than 65 years²⁶ and in 11% to 30% of those who are older than 65 years.⁶⁴⁻⁷¹ Postural hypotension is more likely if the patient has supine systolic hypertension^{58,67,68,71-73} but is not more likely, surprisingly, if the patient takes cardiovascular or psychotropic medications.^{47,65,68,71,74} Finally, the symptom of mild or moderate postural dizziness is a poor predictor of postural hypotension in most studies.^{56,67,70}

Pathogenesis and Definition of Other Physical Findings

The capillary refill time is determined by compressing the distal phalanx of the patient's middle finger, positioned level with the heart, for 5 seconds and then timing the return of normal color to the finger. With an ambient temperature of 21°C, the upper limits

of normal for the refill time are 2 seconds for children and adult men, 3 seconds for adult women, and 4 seconds for the elderly.⁷⁵

Poor skin turgor refers to the slow return of skin to its normal position after being pinched between the examiner's thumb and forefinger.⁷⁶ The protein elastin, which is responsible for the recoil of skin, is markedly affected by moisture content. As little as 3.4% loss in wet weight may prolong the recoil time 40-fold.⁷⁶ Elastin deteriorates with age, suggesting that the recoil of skin normally decreases with age, although this has never been formally studied to the authors' knowledge. No studies on the normal recoil time or precise definitions of technique could be found.

Cellular dehydration, interstitial space dehydration, and poor perfusion are presumably responsible for many of the other classic signs of hypovolemia, such as longitudinal tongue furrows, dry mucous membranes, dry axillae, and sunken eyes. No studies on the pathogenesis of these findings, however, could be found.

Precision of Physical Signs

Reproducible measurements of BP depend on many variables, including the examiner's technique, the patient examined, and various observer biases and errors, all of which are thoroughly reviewed in another article.⁷⁷

Outside of an extensive literature devoted to patients with syncope that uses different methods and end points than those

discussed in this article, the few studies of tilt test reproducibility focus more on biological variation (ie, reproducibility of the test when repeated days later) than on immediate interobserver reproducibility. When postural vital signs of 911 elderly nursing home residents were measured 4 times daily, postural hypotension was present only 1 of the 4 times in 18% of the residents, 2 or 3 times in 20%, and all 4 times in only 13%.⁶⁸ Postural hypotension is more reproducible in the morning than in the afternoon^{68,78} or if cardiovascular medications are withheld. Cardiovascular medications can unmask supine systolic hypertension, a known risk factor for postural hypotension.^{58,73}

In acutely ill elderly patients, interobserver agreement for axillary sweating (dry vs moist) was moderate (κ , 0.50; 80% simple agreement).³³ In addition, the clinician's assessment of axillary moisture correlated well with the weight gain of a piece of preweighed tissue paper applied to the patient's axilla for 15 minutes.³³ With stopwatches, the measurements of capillary refill time by 2 observers were within 0 to 0.3 seconds of each other.⁷⁵

Accuracy of Physical Signs for Acute Blood Loss

Table 24-4 reveals that the 2 most valuable observations from the tilt test are either a postural pulse increment of 30/min or more or the inability of the patient to stand for

Table 24-4 Diagnostic Accuracy of Vital Signs for Acute Blood Loss

Finding	Source, y	Moderate Blood Loss, Sensitivity (95% CI), % ^a	Large Blood Loss, Sensitivity (95% CI), % ^b	Before Blood Loss, Specificity (95% CI), %
Postural pulse increment ≥ 30/min or severe postural dizziness ^c	Knopp et al, ²⁴ 1980	57	98	98
	Shenkin et al, ²⁹ 1944	... ^d	100	...
	Baraff and Schriger, ²⁵ 1992	8	...	100
	Witting et al, ²⁶ 1994	14	...	99
	Summary measure ^e	22 (6-48)	97 (91-100)	98 (97-99)
Postural hypotension (>20 mm Hg decrease in SBP) ^{c,f}	Baraff and Schriger, ²⁵ 1992	7	...	98
	Witting et al, ²⁶ 1994	9	...	90
	Summary measure ^e	9 (6-12)	...	94 (84-99)
	Age ≥ 65 y	Witting et al, ²⁶ 1994	27 (14-40)	...
Supine tachycardia (pulse > 100/min)	Ralston et al, ²⁷ 1961	0	0	100
	Shenkin et al, ²⁹ 1944	...	9	91
	Wallace and Sharpey-Schafer, ³⁰ 1941	...	16	100
	Skillman et al, ³¹ 1967	...	0	100
	Summary measure ^e	0 (0-42)	12 (5-24)	96 (88-99)
Supine hypotension (SBP < 95 mm Hg)	Warren et al, ²⁸ 1945	13	...	100
	Shenkin et al, ²⁹ 1944	...	36	100
	Wallace and Sharpey-Schafer, ³⁰ 1941	...	32	96
	Skillman et al, ³¹ 1967	...	56	100
	Bergenwald et al, ³² 1977	...	13	...
	Summary measure ^e	13 (0-50)	33 (21-47)	97 (90-100)

Abbreviations: CI, confidence interval; SBP, systolic blood pressure.

^aModerate blood loss, 450 to 630 mL.

^bLarge blood loss, 630 to 1150 mL.

^c"Postural" indicates change from supine to standing position.

^dEllipses indicate data not available.

^eSummary measures calculated with random-effects model.

^fExcludes those patients unable to stand because of severe dizziness.

vital signs because of severe dizziness. After blood loss of 450 to 630 mL, only 1 in 5 patients demonstrates these findings. The sensitivity increases to 97% (95% CI, 91%-100%) after 630- to 1150-mL blood loss. The specificity is 98% (95% CI, 97%-99%), a value similar to that generated from the studies in Table 24-3. Either of these findings is durable after hemorrhage, lasting at least 12 to 72 hours if intravenous fluids are withheld.^{11,30,39} If the patient sits instead of stands from the supine position, the sensitivity decreases, being 39%²⁴ and 78%³⁰ in 2 studies after 1000 mL of hemorrhage. Because the studies on large blood loss (630-1150 mL) enrolled younger healthy individuals, the sensitivity may also be lower in elderly patients or those taking medications such as β -blockers. A patient's complaint of postural dizziness, not severe enough to prevent standing and accompanied by a pulse increment lower than 30/min, has little predictive value.^{26,56}

After excluding those unable to stand to have vital signs measured, postural hypotension (a more than 20 mm Hg decrease in systolic BP) has little additional predictive value. Its sensitivity for 450 to 630 mL of blood loss is only 9% in those younger than 65 years and 27% in those older than 65 years. These numbers are similar to the false-positive rates in some studies of the same age groups, 10% (<65 years)²⁶ and 28% (>65 years),⁷¹ resulting in positive LRs (LR+) close to unity. There are insufficient data to address the value of isolated postural hypotension after 630 to 1150 mL of blood loss.

Supine tachycardia (pulse > 100/min) is a specific but insensitive indicator of blood loss (specificity, 96%). Thus, patients without supine tachycardia can still have significant

blood loss. In contrast, bradycardia occurs frequently after significant blood loss, often immediately preceding the decrease in systemic vascular resistance and the fainting that may occur.^{11,27-32,39,79-81} One study⁸⁰ showed a strong correlation between the decrease in heart rate after blood loss and the maximal decrease in BP ($r = 0.79$), and, in hypotensive patients receiving fluid resuscitation, the pulse may paradoxically increase initially.⁸¹

In patients with suspected blood loss, supine hypotension (systolic BP < 95 mm Hg) is a specific finding of hypovolemia (specificity, 97%), although it is insensitive to both moderate blood loss of 450 to 630 mL (sensitivity, 13%) and more significant loss of 630 to 1150 mL (sensitivity, 33%).

Using the age- and sex-specific upper limits of normal for capillary refill time defined earlier, a prolonged refill time does not accurately predict 450 mL of blood loss (sensitivity 6%; specificity, 93%) and yields an LR+ of 1.0.³⁶ If the clinician instead uses an arbitrary upper limit of 2 seconds, diagnostic performance is no better (sensitivity, 11%; specificity, 89%; LR+, 1.0).³⁶

Accuracy of Physical Signs for Other Causes of Hypovolemia

Table 24-5 reviews the sensitivity and specificity of various physical signs for the diagnosis of hypovolemia derived from studies of individuals usually presenting to emergency departments with vomiting, decreased oral intake, or diarrhea. Except for 1 study,³⁵ which enrolled young women with hyperemesis gravidarum, these studies generally recruited older adults.

Table 24-5 Diagnostic Accuracy of Physical Signs for Hypovolemia Not Due to Blood Loss

Physical Finding	Source, y	Grade of Study ^a	Definition of Abnormal Finding	Sensitivity, %	Specificity, %	LR+ (95% CI)	LR- (95% CI)
Postural vital signs	Johnson et al, ³⁵ 1995	C	Pulse increment > 30 beats/min	43	75	1.7 (0.7-4.0)	0.8 (0.5-1.3)
	Johnson et al, ³⁵ 1995	C	Postural hypotension (SBP decline > 20 mm Hg)	29	81	1.5 (0.5-4.6)	0.9 (0.6-1.3)
Skin, eyes, and mucous membranes	Eaton et al, ³³ 1994	A	Dry axilla	50	82	2.8 (1.4-5.4)	0.6 (0.4-1.0)
	Gross et al, ³⁴ 1992	B	Mucous membranes of mouth and nose dry	85	58	2.0 (1.0-4.0)	0.3 (0.1-0.6)
	Gross et al, ³⁴ 1992	B	Tongue dry	59	73	2.1 (0.8-5.8)	0.6 (0.3-1.0)
	Gross et al, ³⁴ 1992	B	Longitudinal furrows on tongue	85	58	2.0 (1.0-4.0)	0.3 (0.1-0.6)
	Gross et al, ³⁴ 1992	B	Sunken eyes	62	82	3.4 (1.0-12)	0.5 (0.3-0.7)
Neurologic findings	Gross et al, ³⁴ 1992	B	Confusion present	57	73	2.1 (0.8-5.7)	0.6 (0.4-1.0)
	Gross et al, ³⁴ 1992	B	Upper or lower extremity weakness present	43	82	2.3 (0.6-8.6)	0.7 (0.5-1.0)
	Gross et al, ³⁴ 1992	B	Speech not clear or expressive	56	82	3.1 (0.9-11)	0.5 (0.4-0.8)
Capillary refill time	Schriger and Baraff, ³⁶ 1991	C	Capillary refill time greater than age- and sex-specific upper normal limit (see "Results")	34	95	6.9 (3.2-15)	0.7 (0.5-0.9)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio; SBP, systolic blood pressure.

^aSee Table 24-2 footnotes for grading determinations. See also Table 1-7 for a summary of Evidence Grades and their relationship to Evidence Levels.

A dry axilla increases the probability of hypovolemia (LR+, 2.8; 95% CI, 1.4-5.4), although it is an insensitive physical sign (sensitivity, 50%).³³ A moist axilla decreases the probability of volume depletion only slightly (negative [LR-], 0.6; 95% CI, 0.4-1.0).

In the study by Johnson et al³⁵ of 23 women with hyperemesis gravidarum, neither postural hypotension nor a postural pulse increment of more than 30/min was particularly helpful (Table 24-5). In this study, however, the specificity of a pulse increment of more than 30/min was unusually low (75%). One reason for this could be the authors' definition of dehydration ($\geq 5\%$ weight gain after 12 hours of rehydration), which led them to classify as nondiseased the dehydrated women with less than 5% weight gain, thus devaluing the specificity calculation. Alternatively, the postural pulse increment may be less specific because of pregnancy.

In another study of 202 individuals with acute illnesses, investigators used multiple analysis of variance to identify which clinical findings best explained the variation in total body water deficit, as calculated from the patient's serum osmolality.⁵⁹ The finding of a dry axilla was significantly associated with the severity of dehydration ($P = .03$). The postural pulse increment was also significantly associated but only weakly so ($r = 0.22$; $P = .02$).⁵⁹ The mean water deficit in this study was only 3.9%, correlating with a 140-mL deficit from the vascular space (or about 250 mL of blood), a level below that in the phlebotomy studies discussed earlier. This study found no association between dehydration and postural changes of systolic BP.

In Table 24-5, the capillary refill time seems to perform impressively, especially when the capillary refill time is prolonged (LR+, 6.9).³⁶ However, the criterion standard in this study was the supine and postural vital signs, raising the question whether capillary refill time has any incremental diagnostic value. Another study found no correlation between capillary refill time, tested over the patella, and objective measures of hypovolemia.³⁴

In a study of 55 elderly patients presenting with suspected hypovolemia, the 7 physical signs of confusion, extremity weakness, nonfluent speech, dry mucous membranes, dry tongue, furrowed tongue, and sunken eyes correlated best with measurement of the serum sodium and serum urea nitrogen/creatinine ratio (Table 24-5).³⁴ According to the CIs of the LRs, however, none of these findings is particularly helpful when present in isolation. Combinations of findings may be more helpful—on average, patients with severe and moderate hypovolemia had 5.7 and 3.9, respectively, of these 7 signs, whereas those without dehydration had only 1.3—but this requires validation.³⁴ The most helpful negative findings, arguing against hypovolemia, are moist mucous membranes, absence of sunken eyes, and absence of furrows on the tongue.

Another study found no correlation between degree of hypovolemia and dryness of mucous membranes.⁵⁹ In adults, 2 studies have found poor skin turgor to have no diagnostic value.^{34,59}

THE BOTTOM LINE

When obtaining postural vital signs, clinicians should wait at least 2 minutes before measuring the supine vital signs and 1 minute after the patient stands before measuring the upright vital signs. Counting the pulse for 30 seconds and doubling the result is more accurate than 15 seconds of observation.⁴⁴ In normovolemic individuals, a postural pulse increment of more than 30/min is uncommon, affecting only about 2% to 4% of individuals.

When patients with suspected blood loss are evaluated, the most helpful physical findings are severe postural dizziness (preventing measurement of upright vital signs) or a postural pulse increment of 30/min or more. Having the patient sit instead of stand reduces the sensitivity of the tilt test. After excluding those unable to stand, postural hypotension has no incremental diagnostic value.

Supine hypotension and tachycardia are frequently absent, even with more than 1000 mL of blood loss, and the symptom of mild postural dizziness has no proven diagnostic value. Bradycardia is common after significant blood loss.

Rigorous conclusions about the role of physical examination for assessing the volume and hydration status of patients with vomiting, diarrhea, or decreased oral intake are difficult to make because there are few relevant studies. Severe postural dizziness or a postural pulse increment of 30/min or more should be just as accurate as after blood loss, although one study³⁵ of the pulse increment in patients with hyperemesis gravidarum failed to confirm this. A dry axilla supports the diagnosis of hypovolemia in the elderly, and moist mucous membranes and a tongue without furrows argue against it. However, clinicians should recall that the criterion standard of hypovolemia in these studies—simple serum chemistry measurements—is easily accessible to clinicians.

Case 1 demonstrates a postural pulse increment of more than 30/min, suggesting significant blood loss. The clinician should start fluid resuscitation. In case 2, postural hypotension and mild postural dizziness lack the specificity necessary to condemn the diuretic at this time. The clinician could continue the diuretic treatment if the physician believes the patient's dizziness comes from inner-ear vertigo. Finally, despite the negative physical examination findings in case 3, this patient has many risk factors for significant hypovolemia, and the clinician should measure the serum blood urea nitrogen, creatinine, and electrolyte levels before making the decision to discharge the patient.

Author Affiliations at the Time of the Original Publication

Department of Medicine, University of Washington, Seattle-Puget Sound VA Health Care System, Seattle (Dr McGee); Department of Cardiovascular Medicine, Massachusetts General Hospital, Boston, Massachusetts (Dr Abernethy); and Department of Medicine, Ambulatory Care, Duke University and Durham Veterans Affairs Medical Center, Durham, North Carolina (Dr Simel).

Acknowledgments

We thank Kenneth Goldberg, MD, and Bill Yarger, MD, for their comments.

REFERENCES

1. Mange K, Matsuura D, Cizman B, et al. Language guiding therapy: the case of dehydration versus volume depletion. *Ann Intern Med.* 1997;127(9):848-853.
2. Cutler JA, Mendeloff AI. Upper gastrointestinal bleeding: nature and magnitude of the problem. *Dig Dis Sci.* 1981;26(7 suppl):90S-96S.
3. Silverstein FE, Gilbert DA, Tedesco FJ, et al. The national ASGE survey on upper gastrointestinal bleeding, II. *Gastrointest Endosc.* 1981;27(2):80-93.
4. Palevsky PM, Bhagrath R, Greenberg A. Hypernatremia in hospitalized patients. *Ann Intern Med.* 1996;124(2):197-203.
5. Snyder NA, Feigel DW, Arief AI. Hypernatremia in elderly patients. *Ann Intern Med.* 1987;107(3):309-319.
6. Lavizzo-Mourey R, Johnson J, Stolley P. Risk factors for dehydration among elderly nursing home residents. *J Am Geriatr Soc.* 1988;36(3):213-218.
7. Schiff L, Stevens RJ, Shapiro N, Goodman S. Observations on the oral administration of citrated blood in man, II. *Am J Med Sci.* 1942;203:409-412.
8. Wara P, Stokilde H. Bleeding pattern before admission as guideline for emergency endoscopy. *Scand J Gastroenterol.* 1985;20(1):72-78.
9. Tibbs DJ, Leslie WG. Blood volumes in gastroduodenal haemorrhage. *Lancet.* 1956;2:266-274.
10. Schiller KFR, Truelove SC, Williams DG. Haematemesis and melaena, with special reference to factors influencing the outcome. *BMJ.* 1970;2(5700):7-14.
11. Ebert RV, Stead EA, Gibson JG. Response of normal subjects to acute blood loss. *Arch Intern Med.* 1941;68:578-590.
12. Taylor PH, Heydinger DK, Bowers JO, Callendine GW. Evaluation of small acute blood loss in man. *Am J Surg.* 1967;114(6):913-916.
13. Bennett TI, Dow J, Lander L, Wright S. Severe haemorrhage from the stomach and duodenum. *Lancet.* 1938;2:651-655.
14. Klockhoff I, Lindblom U. Ménière's disease and hydrochlorothiazide. *Acta Otolaryngol (Stockh).* 1967;63(4):347-365.
15. Bergman GE, Reisner FF, Anwar RAH. Orthostatic changes in normovolemic children: an analysis of the "tilt test." *J Emerg Med.* 1983;1(2):137-141.
16. Castro W, Skarin R, Roscelli JD. Orthostatic heart rate and arterial blood pressure changes in normovolemic children. *Pediatr Emerg Care.* 1985;1(3):123-127.
17. Duggan C, Refat M, Hashem M, et al. How valid are clinical signs of dehydration in infants? *J Pediatr Gastroenterol Nutr.* 1996;22(1):56-61.
18. Fuchs SM, Jaffe DM. Evaluation of the "tilt test" in children. *Ann Emerg Med.* 1987;16(4):386-390.
19. Laron Z. Skin turgor as a quantitative index of dehydration in children. *Pediatrics.* 1957;19:816-822.
20. Mackenzie A, Barnes G, Shann F. Clinical signs of dehydration in children. *Lancet.* 1989;2(8663):605-607.
21. Paxton LA, Redd SC, Steketee RW, et al. An evaluation of clinical indicators for severe paediatric illness. *Bull World Health Organ.* 1996;74(6):613-618.
22. Saavedra JM, Harris GD, Li S, Finberg L. Capillary refilling skin turgor in the assessment of dehydration. *Am J Dis Child.* 1991;145(3):296-298.
23. Yurdakok K. Dry cry: a new definition for old criteria in dehydrated children. *Arch Pediatr Adolesc Med.* 1997;151(2):208-209.
24. Knopp R, Claypool R, Leonardi D. Use of the tilt test in measuring acute blood loss. *Ann Emerg Med.* 1980;9(2):72-75.
25. Baraff LJ, Schriger DL. Orthostatic vital signs: variation with age, specificity, and sensitivity in detecting a 450-mL blood loss. *Am J Emerg Med.* 1992;10(2):99-103.
26. Witting MD, Wears RL, Li S. Defining the positive tilt test. *Ann Emerg Med.* 1994;23(6):1320-1323.
27. Ralston LA, Cobb LA, Bruce RA. Acute circulatory effects of arterial bleeding as determined by indicator-dilution curves in normal human subjects. *Am Heart J.* 1961;61:770-776.
28. Warren JV, Brannon ES, Stead EA, Merrill AJ. The effect of venesection and the pooling of blood in the extremities on the atrial pressure and cardiac output in normal subjects with observations on acute circulatory collapse in three instances. *J Clin Invest.* 1945;24(3):337-344.
29. Shenkin HA, Cheney RH, Govons SR, Hardy JD, Fletcher AG, Starr I. On the diagnosis of hemorrhage in man. *Am J Med Sci.* 1944;208:421-436.
30. Wallace J, Sharpey-Schafer EP. Blood changes following controlled haemorrhage in man. *Lancet.* 1941;241:393-395.
31. Skillman JJ, Olson JE, Lyons JH, Moore FD. The hemodynamic effect of acute blood loss in normal man, with observations on the effect of the Valsalva maneuver and breath holding. *Ann Surg.* 1967;166(5):713-738.
32. Bergenwald L, Freyschuss U, Sjostrand T. The mechanism of orthostatic and haemorrhage fainting. *Scand J Clin Lab Invest.* 1977;37(3):209-216.
33. Eaton D, Bannister P, Mulley GP, Conolly MJ. Axillary sweating in clinical assessment of dehydration in ill elderly patients. *BMJ.* 1994;308(6939):1271.
34. Gross CR, Lindquist RD, Woolley AC, et al. Clinical indicators of dehydration severity in elderly patients. *J Emerg Med.* 1992;10(3):267-274.
35. Johnson DR, Douglas D, Hauswald M, Tandberg D. Dehydration and orthostatic vital signs in women with hyperemesis gravidarum. *Acad Emerg Med.* 1995;2(8):692-697.
36. Schriger DL, Baraff LJ. Capillary refill: is it a useful predictor of hypovolemic states? *Ann Emerg Med.* 1991;20(6):601-605.
37. Eddy DM, Hasselblad V, Schacter RD. *Meta-Analysis by the Confidence Profile Method: The Statistical Synthesis of Evidence.* San Diego, CA: Academic Press Inc; 1992:93-98.
38. Eddy DM, Hasselblad V. *Fast*Pro v1.8: Software for Meta-Analysis by the Confidence Profile Method.* San Diego, CA: Academic Press Inc; 1992.
39. Green DM, Metheny D. The estimation of acute blood loss by the tilt test. *Surg Gynecol Obstet.* 1947;84:1045-1050.
40. Wong DH, O'Connor D, Tremper KK, et al. Changes in cardiac output after acute blood loss and position change in man. *Crit Care Med.* 1989;17(10):979-983.
41. Velasquez MT, Menitove JE, Skelton MM, Cowley AW. Hormonal responses and blood pressure maintenance in normal and hypertensive subjects during acute blood loss. *Hypertension.* 1987;9(5):423-428.
42. Triedman JK, Cohen RJ, Saul JP. Mild hypovolemic stress alters autonomic modulation of heart rate. *Hypertension.* 1993;21(2):236-247.
43. Simel DL, Samsa GP, Matchar DB. Likelihood ratios with confidence. *J Clin Epidemiol.* 1991;44(8):763-770.
44. Hollerbach AD, Sneed NV. Accuracy of radial pulse assessment by length of counting interval. *Heart Lung.* 1990;19(3):258-264.
45. Smith JJ, Porth CM, Erickson M. Hemodynamic response to the upright posture. *J Clin Pharmacol.* 1994;34(5):375-386.
46. Streeten DHP, Anderson GH, Richardson R, Thomas FD. Abnormal orthostatic changes in blood pressure and heart rate in subjects with intact sympathetic nervous function: evidence for excessive venous pooling. *J Lab Clin Med.* 1988;111(3):326-335.
47. Jansen RW, Lenders JW, Thien T, Hoefnagels WH. The influence of age and blood pressure on the hemodynamic and humoral response to head-up tilt. *J Am Geriatr Soc.* 1989;37(6):528-532.
48. Tuckman J, Shillingford J. Effect of different degrees of tilt on cardiac output, heart rate, and blood pressure in normal man. *Br Heart J.* 1966;28(1):32-39.
49. Tell GS, Prineas RJ, Gomez-Marin O. Postural changes in blood pressure and pulse rate among black adolescents and white adolescents. *Am J Epidemiol.* 1988;128(2):360-369.
50. Honda K, Nose T, Yoshida N, Tanimura M, Tanaka K. Responses to the postural change and orthostatic dysregulation symptoms. *Jpn Circ J.* 1977;41(6):629-641.
51. Horam WJ, Roscelli JD. Establishing standards of orthostatic measurements in normovolemic adolescents. *Am J Dis Child.* 1992;146(7):848-851.
52. Borst C, Wieling W, van Brederode JFM, et al. Mechanisms of initial heart rate response to postural change. *Am J Physiol.* 1982;243(5):H676-H681.
53. Kajser L, Sachs C. Autonomic cardiovascular responses in old age. *Clin Physiol.* 1985;5(4):347-357.
54. Moore KI, Newton K. Orthostatic heart rates and blood pressures in healthy young women and men. *Heart Lung.* 1986;15(6):611-617.
55. Currens JH. A comparison of the blood pressure in the lying and standing positions. *Am Heart J.* 1948;35:646-654.
56. Koziol-McLain J, Lowenstein SR, Fuller B. Orthostatic vital signs in emergency department patients. *Ann Emerg Med.* 1991;20(6):606-610.
57. Dambrink JHA, Wieling W. Circulatory response to postural change in healthy male subjects in relation to age. *Clin Sci (Colch).* 1987;72(3):335-341.
58. Lipsitz LA, Storch HA, Minaker KL, Rowe JW. Intra-individual variability in postural blood pressure in the elderly. *Clin Sci (Colch).* 1985;69(3):337-341.
59. Levitt MA, Lopez B, Lieberman ME, Sutton M. Evaluation of the tilt test in an adult emergency medicine population. *Ann Emerg Med.* 1992;21(6):713-718.
60. Schneider EC, Truesdell D. A statistical study of the pulse rate and the arterial blood pressures in recumbency, standing, and after a standard exercise. *Am J Physiol.* 1922;61:429-474.

61. Smith JJ, Hughes CV, Ptacin MJ, et al. The effect of age on hemodynamic response to graded postural stress in normal men. *J Gerontol.* 1987;42(4):406-411.
62. Patel A, Maloney A, Damato AN. On the frequency and reproducibility of orthostatic blood pressure changes in healthy community-dwelling elderly during 60-degree head-up tilt. *Am Heart J.* 1993;126(1):184-188.
63. Ten Harkel AD, Van Lieshout JJ, Van Lieshout EJ, Wieling W. Assessment of cardiovascular reflexes. *J Appl Physiol.* 1990;68(1):147-153.
64. Aronow WS, Lee NH, Sales FF, Etienne F. Prevalence of postural hypotension in elderly patients in a long-term health care facility. *Am J Cardiol.* 1988;62(4):336.
65. Caird FI, Andrews GR, Kennedy RD. Effect of posture on blood pressure in the elderly. *Br Heart J.* 1973;35(5):527-530.
66. Johnson RH, Smith AC, Spalding JMK, Wollner L. Effect of posture on blood-pressure in elderly patients. *Lancet.* 1965;1(7388):731-733.
67. Mader SL, Josephson KR, Rubenstein LZ. Low prevalence of postural hypotension among community-dwelling elderly. *JAMA.* 1987;258(11):1511-1514.
68. Ooi WL, Barrett S, Hossain M, Kelley-Gagnon M, Lipsitz LA. Patterns of orthostatic blood pressure change and their clinical correlates in a frail, elderly population. *JAMA.* 1997;277(16):1299-1304.
69. Palmer KT. Studies into postural hypotension in elderly patients. *N Z Med J.* 1983;96(724):43-45.
70. Rodstein M, Zeman FD. Postural blood pressure changes in the elderly. *J Chronic Dis.* 1957;6(6):581-588.
71. Raiha I, Luutonen S, Piha J, et al. Prevalence, predisposing factors, and prognostic importance of postural hypotension. *Arch Intern Med.* 1995;155(9):930-935.
72. MacLennan WJ, Hall MRP, Timothy JI. Postural hypotension in old age. *Age Ageing.* 1980;9(1):25-32.
73. Jansen RW, Kelley-Gagnon MM, Lipsitz LA. Intraindividual reproducibility of postprandial and orthostatic blood pressure changes in older nursing-home patients. *J Am Geriatr Soc.* 1996;44(4):383-389.
74. Myers MG, Kearns PM, Kennedy DS, Fisher RH. Postural hypotension and diuretic therapy in the elderly. *CMAJ.* 1978;119(6):581-585.
75. Schriger DL, Baraff L. Defining normal capillary refill. *Ann Emerg Med.* 1988;17(9):932-935.
76. Dorrington KL. Skin turgor: do we understand the clinical sign? *Lancet.* 1981;1(8214):264-265.
77. Reeves RA. Does this patient have hypertension? *JAMA.* 1995;273(15):1211-1218.
78. Ward C, Kenny RA. Reproducibility of orthostatic hypotension in symptomatic elderly. *Am J Med.* 1996;100(4):418-422.
79. Barcroft H, Edholm OG, McMichael J, Sharpey-Schafer EP. Posthaemorrhagic fainting. *Lancet.* 1944;1:489-491.
80. Ruetz PP, Johnson SA, Callahan R, et al. Fainting: a review of its mechanisms and a study in blood donors. *Medicine (Baltimore).* 1967;46(5):363-384.
81. Sander-Jensen K, Secher NH, Bie P, Warberg J, Schwartz TW. Vagal slowing of the heart during haemorrhage. *BMJ.* 1986;292(6517):364-366.

This page intentionally left blank

Prepared by David L. Simel, MD, MHS
Reviewed by Kenneth Goldberg, MD, and Ali Raja, MD

CLINICAL SCENARIO

A 75-year-old man fell at home and was on the floor for 8 hours, unable to ambulate. When his family checked on him, they brought him immediately to the emergency department. You suspect a hip fracture, but you are also concerned about intravascular volume depletion and rhabdomyolysis. Although he cannot stand up, he can change position from lying down to about a 45-degree angle before the hip begins to hurt. His pulse increases 22/min when he sits at the angle.

UPDATED SUMMARY ON HYPOVOLEMIA

Original Review

McGee S, Abernethy WB 3rd, Simel DL. Is this patient hypovolemic? *JAMA*. 1999;281(11):1022-1029.

UPDATED LITERATURE SEARCH

Our literature search replicated that done in the original publication. We used the parent search strategy for The Rational Clinical Examination series and combined it with “dehydration/di,” “exp hypotension,” “tilt-table test.mp,” and “exp hypovolemia.” We also searched on the text words “orthostatic vital,” “orthostatic pulse,” “postural pulse,” and “postural vital.” This strategy yielded 258 English-language articles published between 1998 and September 2004. We excluded case reports and then reviewed the title to identify potentially eligible articles. The focus was on adults with acute hypovolemia, rather than chronic orthostatic hypotension, using the clinical evaluation or commonly available bedside tests. We identified 23 articles for review, but only 3 contained prospectively collected data applicable to the clinical scenario of acute volume depletion in adults. The reference list for each article was reviewed but yielded no additional studies. To validate the literature search, we also used the SUMSearch strategy (<http://sumsearch.uthscsa.edu>; accessed May 31, 2008) in PubMed for the same search, limited to physical examination since 1997; we found no additional articles for review.

NEW FINDINGS

- A pulse change of 30/min on going from supine to standing remains the most helpful physical finding. A change of only 20/min should be used for the change from sitting to standing.
- Individual clinical findings are not useful in the intensive care unit (ICU), but combinations of findings may be helpful.
- In healthy young patients, one study suggests that the bedside specific gravity cutoff of 1.020 identifies patients at higher and lower risk of dehydration.

Details of the Update

Although there have been many studies on orthostatic hypotension (especially in the elderly), the focus of this review was orthostatic hypotension secondary to hypovolemia. Thus, the results apply only to patients for whom there is a suspicion of intravascular volume depletion. Examples of clinical conditions would be acute blood loss, gastrointestinal illness with fluid loss, decreased oral intake, or “unmeasured” losses as might occur with heat-induced illness.

IMPROVEMENTS IN THE DATA PRESENTED IN THE ORIGINAL PUBLICATION

No additional data were found that modify the original results.

CHANGES IN THE REFERENCE STANDARD

There have been no changes in the reference standard. One new study used radiolabeling to measure circulating blood volume. A second study from a metabolic laboratory used radiolabeling to quantify changes in total body water and extracellular water. Although hypovolemia from blood loss can be established clinically and with laboratory tests, a pragmatic clinical reference standard continues to be a problem for both clinical work and research studies of other types of hypovolemia. Most clinicians would accept a combination of laboratory findings and the response to rehydration as the reference standard in typical clinical settings.

RESULTS OF LITERATURE REVIEW

One study assessed a variety of clinical variables for detecting hypovolemia in the ICU patient.¹ The variables are all readily obtainable. However, these findings taken individually were essentially useless, as exhibited by likelihood confidence intervals that included 1. When used in combination, the most important variables were the assessments of third spacing (ie, ascites or pleural effusion), a clinical diagnosis of heart failure, and pulmonary edema (Tables 24-6 and 24-7).

For patients who are not so acutely ill that they must be supine, physicians (or nurses) might often obtain sitting and standing vital signs rather than supine and standing. A study of sitting-to-standing orthostatic changes was done on patients in the emergency department who did not have an acute illness that would have affected orthostatic vital signs.² This study found that a change in pulse of greater than or equal to 20/min should be the cut point for sitting to standing changes. This recommendation has face validity but should preferably be validated in patients with a suspicion for hypovolemia.

A study of controlled dehydration in collegiate wrestlers assessed the role of the urine specific gravity level determined with a urine dipstick.³ The measurement of specific gravity in the correct setting on the appropriate patient may have merit. In young, healthy subjects for whom there is a suspi-

Table 24-6 Increases the Likelihood of Hypovolemia in Intensive Care Unit Patients

1. Presence of obvious fluid losses as occurs through drainage tubes
2. Fluid balance from input and output sheets

Table 24-7 Decreases the Likelihood of Hypovolemia in Intensive Care Unit Patients

1. Peripheral edema
2. Pulmonary edema
3. Third spacing
4. Skin mottling
5. Clinically evident heart failure

cion of hypovolemia not caused by blood loss, a specific gravity threshold of 1.020 might be useful for both ruling in and ruling out intravascular volume depletion.

Multivariate Findings for Hypovolemia

Although a quantitative predictive model has been developed that uses clinical features, it was developed and validated only for ICU patients for whom the diagnosis of hypovolemia was uncertain. Because of that, we cannot assess the generalizability of these features, especially because most of the features apply only to patients who have been in the ICU for several days. Until the results are confirmed, clinicians might want to collect these variables and assess their importance more qualitatively.

EVIDENCE FROM GUIDELINES

No guidelines apply to the assessment of intravascular volume depletion in adults.

CLINICAL SCENARIO—RESOLUTION

From the clinical history, the likelihood of intravascular volume depletion seems high. The patient has had no oral intake for 8 hours. In addition, he may have hemorrhage from his hip fracture and resulting intravascular blood loss. The increase in pulse of more than 20/min from lying down to sitting supports the diagnosis, but it could also be from pain on movement of the hip. Although a change in postural tachycardia would be helpful to assess intravascular volume depletion, it is not necessary to measure this because you have enough evidence to obtain laboratory tests for assessing the effect of intravascular volume depletion. Furthermore, this patient could be considered a “trauma” patient and the presence of tachycardia in blood loss is not universal. A urinalysis would likely be obtained (for assessing rhabdomyolysis), but the urine specific gravity in this older patient may be a marker of his renal function rather than his intravascular volume.

ADULT HYPOVOLEMIA—MAKE THE DIAGNOSIS

PRIOR PROBABILITY

Hypovolemia occurs for a variety of causes. There are no reasonable estimates for the prior probability that would be uniformly helpful. Clinicians should use their best judgment in assessing the probability of intravascular volume depletion according to the patient's medical history and findings that suggest the possibility of fluid losses.

POPULATION FOR WHOM HYPOVOLEMIA DISEASE SHOULD BE CONSIDERED

- Acute blood loss
- Illness with fluid loss
- Decreased oral intake
- “Unmeasured” losses as might occur with heat-induced illness

See Tables 24-8 and 24-9 for the likelihood of hypovolemia caused by blood loss.

Table 24-8 Detecting the Likelihood of Hypovolemia Not Caused by Blood Loss

Finding, Patient Population	LR+ (95% CI)	LR- (95% CI)
Urine specific gravity > 1.020	11 (3-43)	0.09 (0.03-0.36)
• Young, healthy college wrestlers		
• Dehydration secondary to sweating		
Dry axilla	2.8 (1.4-5.4)	0.6 (0.4-1.0)
Patients > 70 y with acute illness		
Pulse increment of > 30/min (supine to standing) ^a	1.7 (0.7-4.0)	0.8 (0.5-1.3)
• Pregnant women in emergency department (1 study)		
• Normal electrolyte and creatinine levels		

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

^aIn other populations, data suggest lowering the threshold to ≥ 20/min when the patient moves from sitting to standing.

Table 24-9 Detecting the Likelihood of Hypovolemia Caused by Blood Loss

Pulse Increment 30/min or Postural Dizziness ^a	Sensitivity, % (95% CI)	Specificity, % (95% CI)
Moderate blood loss (450-630 mL)	22 (6-48)	98 (97-99)
Larger blood loss (630-1150 mL)	97 (91-100)	

Abbreviation: CI, confidence interval.

^aBased on phlebotomy studies in normovolemic individuals. Specificity is based on results for these normovolemic adults before phlebotomy.

REFERENCE STANDARD TESTS

Intravascular volume depletion typically relies on a clinical diagnosis, with appropriate laboratory measures that correct with rehydration. In controlled settings, blood volume and total body water can be measured indirectly with radiolabeled agents.

REFERENCES FOR THE UPDATE

1. Stéphan F, Flahault A, Dieudonné N, Hollande J, Paillard F, Bonnet F. Clinical evaluation of circulating blood volume in critically ill patients—contribution of a clinical scoring system. *Br J Anaesth.* 2001;86(6):754-762.^a
2. Witting MD, Gallagher K. Unique cutpoints for sitting-to-standing orthostatic vital signs. *Am J Emerg Med.* 2003;21(1):45-47.^a

3. Bartok C, Schoeller DA, Sullivan JC, Clark RR, Landry GL. Hydration testing in collegiate wrestlers undergoing hypertonic dehydration. *Med Sci Sports Exerc.* 2004;36(3):510-517.^a

^aFor the Evidence to Support the Update for this topic,
see <http://www.JAMAEvidence.com>.

This page intentionally left blank

EVIDENCE TO SUPPORT THE UPDATE:

24

Hypovolemia, Adult

TITLE Hydration Testing in Collegiate Wrestlers Undergoing Hypertonic Dehydration.

AUTHORS Bartok C, Schoeeler DA, Sullivan JC, Clark RR, Landry GL.

CITATION *Med Sci Sport Exerc.* 2004;36(3):510-517.

QUESTION In a controlled situation of iatrogenically induced dehydration, what are the thresholds for commonly measured laboratory tests?

DESIGN Prospective.

SETTING Metabolic laboratory.

PATIENTS Twenty-five healthy collegiate wrestlers.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

The students were evaluated in a euvolemic state on day 1. On day 2, they were randomized to dehydration levels of 2%, 3%, 4%, or 5%. Careful measurements of diet, fluid intake, weight, and both total body and extracellular water (radiodilution techniques) confirmed that they reached the level of prespecified dehydration. The urine specific gravity and protein levels were determined by 2 independent observers. The data were compared with laboratory measures.

MAIN OUTCOME MEASURES

The screening test was urine specific gravity and urine protein levels measured by a bedside test (Multistix; Miles Diagnostics, Elkhart, Indiana).

MAIN RESULTS

Only 1 subject had dipstick proteinuria when euvolemic; all had proteinuria during dehydration.

A receiver operating characteristic curve selected a specific gravity of 1.020 as the appropriate cut point for the dipstick (Table 24-10).

Table 24-10 Likelihood Ratio of Urine Specific Gravity for Dehydration

Test	Cut Point	LR+ (95% CI)	LR- (95% CI)
Urine specific gravity, dipstick (g/mL)	>1.020	11 (3-43)	0.09 (0.03-0.36)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

CONCLUSIONS

LEVEL OF EVIDENCE Level 3.

STRENGTHS Carefully controlled study.

LIMITATIONS Small sample size and unique population of patients limit generalizability. No physical examination findings were included.

We include this study in our review despite the small sample size and lack of clinical examination findings because they evaluated a bedside paraclinical test (urine dipstick) in a highly controlled situation. Although the authors observed a lack of correlation between the absolute specific gravity level measured in the laboratory and the percentage of dehydration, that finding belied the excellent discriminative properties of the specific gravity. The National Collegiate Athletic Association does use a specific gravity of 1.020 as the threshold for further testing to make sure that collegiate wrestlers have not dehydrated themselves to gain eligibility in a lower weight class.¹

The question for clinicians is whether these data apply to patients treated in an uncontrolled situation in an emergency department or outpatient clinic. The subjects in this study on day 1 were used for determining the specificity. On day 2, they underwent controlled dehydration and were used for establishing the sensitivity. Thus, there were 2 populations of patients: one in which hypovolemia was expected and one in which it was not. This sort of enrollment is different from what would happen in clinical practice, in which all the patients are enrolled because of a suspicion of hypovolemia. On the other hand, a prospective study with an enriched population of patients most likely to have hypovolemia would almost certainly yield results with some verification bias (underestimated specificity). Thus, the specificity found in this study is plausible and should be validated.

To apply these results, at the very least the patients must be young and healthy, without chronic illness, and there must be a reasonable basis (ie, acute illness) for suspecting intravascular volume depletion not caused by blood loss. As with collegiate wrestlers, additional laboratory evaluation makes sense when the patient's specific gravity exceeds 1.020 and the subject meets these criteria. Larger studies in a clinical population would be necessary to determine whether lower specific gravity values really do rule out dehydration. Whatever the case, the clinical evaluation must first identify the patients for whom the measure would apply.

REFERENCE FOR THE EVIDENCE

- Bubb RG. 2004 *Wrestling Rules and Interpretation: Appendix H*. National Indianapolis, IN: Collegiate Athletic Association; 2003: WA-27. http://www.ncaa.org/library/rules/2004/2004_wrestling_rules.pdf. Accessed May 31, 2008.

Reviewed by David L. Simel, MD, MHS

either rales or crackles on physical examination or from a chest radiograph that showed alveolar edema and vascular redistribution), (5) clinically diagnosed congestive heart failure, (6) peripheral edema, or (7) evidence of third-spacing of fluid (defined by ascites or pleural effusion). Central venous pressure was measured with a pressure transducer, zeroed to the midchest level. The reference standard for volume lost was assessment of circulating blood volume using radiolabelled albumin. Hypovolemia was defined as a circulating blood volume at least 10% lower than the predicted mean for healthy subjects of the same sex, height, weight, and age. The authors reported that their circulating blood volume was precise to $\pm 5\%$.

In addition to the clinical findings, vital signs (blood pressure and pulse) and laboratory measures were obtained.

MAIN OUTCOME MEASURES

Interobserver variability for clinical findings, and the sensitivity, specificity, and likelihood ratios (LRs) (compared with circulating blood volume). A clinical prediction model was developed from the LRs and tested prospectively.

MAIN RESULTS

Thirty-six (53%) of the prospectively enrolled patients were hypovolemic. For patients who were hypovolemic, the mean blood volume deficit was 514 mL (SD = 194).

The clinical examination components that the clinicians elicited showed excellent observer agreement: The heart rate, systolic and diastolic pressures, and urinary sodium levels were not statistically significant between the hypovolemic and nonhypovolemic groups.

The clinical findings with the highest diagnostic odds ratios (**Table 24-11**) were also the findings that carried the most weight in a predictive score when the variables are considered together.

Table 24-11 Observer Variability (κ), Likelihood Ratios, and Diagnostic Odds Ratios for Clinical Findings of Hypovolemia

Test		LR+ (95% CI)	LR- (95% CI)	DOR
Peripheral edema ^a	0.82	1.5 (0.94-2.4)	0.64 (0.38-1.1)	2.3
Fluid balance		1.5 (0.76-2.9)	0.79 (0.54-1.1)	1.9
Pulmonary congestion ^a	0.78	1.3 (1.0-1.7)	0.27 (0.08-0.90)	5.0
Skin mottling	1.0	1.3 (0.56-3.0)	0.92 (0.70-1.2)	1.4
Clinical diagnosis of heart failure ^a	0.84	1.1 (0.93-1.3)	0.36 (0.07-1.7)	3.1
Third spacing	0.86	1.1 (0.91-1.3)	0.44 (0.12-1.6)	2.5

Abbreviations: CI, confidence interval; DOR, diagnostic odds ratio; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

^aFor these items, the absence of the finding would be considered a "positive" result for hypovolemia. As an example, the lack of a clinical diagnosis of heart failure confers an LR+ for hypovolemia of 1.1. The presence of heart failure makes hypovolemia less likely and therefore has an LR- of 0.36.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

Clinical assessment was done by 2 attending clinicians, independently of each other. Disagreements were resolved by a third clinician. The clinical findings were all readily available at the bedside and included (1) fluid losses (defined by body drainage tubes or aspiration of gastric contents), (2) fluid balance (from the intake and output records), (3) skin mottling, (4) pulmonary congestion (defined by the presence of

The hypovolemia score =
 -5 (from the pretest probability) +
 Fluid loss (14 if present, -4 if absent) +
 Fluid balance (41 if there are more fluids "out" than "in,"
 -24 if balance is equal or positive) +
 Skin mottling (29 if present, -10 if absent) +
 Pulmonary congestion (20 if absent,
 -90 if congestion is present) +
 Heart failure (11 if no heart failure,
 -105 if heart failure is present) +
 Peripheral edema (25 if no edema, -90 if edema is present) +
 Third spacing (27 if no ascites or pleural effusion,
 -184 if ascites or pleural effusion is present)
 Central venous pressure (117 if < 2 mm Hg,
 -42 if > 2 mm Hg)

The results of the score are placed in the equation below to estimate the probability of hypovolemia. Note that if the central venous pressure is not measured, a value of 0 is assigned for the component. The probability can be calculated directly from the equation:

$$\text{Probability (\%)} = \{1 / [(\exp^{(-\text{score}/100)} + 1)]\} \times 100$$

CONCLUSIONS

LEVEL OF EVIDENCE Level 3.

STRENGTHS Objective reference standard for circulating blood volume in a population of patients for whom the clinicians were uncertain about hypovolemia. The clinical findings were assessed independently, and the observer agreement was determined. Definitions for the clinical findings are provided.

LIMITATIONS The authors observe that the reference standard might slightly overestimate circulating blood volume. The comparison of circulating blood volume results with a normal population may not be correct for intensive care unit (ICU) patients (although it does seem reliable). The overall clinical assessment was used to identify patients eligible for this study, resulting in a high prevalence of hypovolemia

compared with that in all other ICUs. The results should not be generalized to settings other than the one in which it was studied. The figures presented in the article do suggest a good correlation at prevalence values of increased circulating blood volume exceeding 50%. The details for data reduction to create a parsimonious model are not given, so it is difficult to know whether all the variables in the model are necessary.

This is a clever study and the investigators use a criterion standard that was applied close to the time of the clinical assessment, independent of the clinical findings, and that was reproducible. They observe, correctly, that the clinical findings all assess extravascular volume excess, which allows clinicians to make inferences about intravascular volume.

The issue of verification bias is difficult to sort out. The patients were selected specifically because the clinicians could not "rule in" or "rule out" hypovolemia—a common problem in ICUs. A bias toward consistently better specificity does not seem to exist. Furthermore, given the number of findings assessed, it seems unlikely that the presence or absence of any one finding consistently identified patients for the study (ie, selection bias). If that is the case, then the findings should have been distributed randomly among those with and without hypovolemia. Given the poor performance characteristics of the individual findings, it seems unlikely that verification bias had a major effect on the final results.

Although the individual findings function poorly, the combination of findings may work well in this setting and with patients for whom the presence of hypovolemia is uncertain. The results suggest the potential importance of evaluating combinations of findings even when the individual clinical examination results lack discriminating power. The model needs validation in the emergency department, but the components suggest it would be less useful in patients who are not acutely ill. First, these patients had been ICU patients for at least a day so that the variables could be assessed. Second, some of the variables would not apply to the acute emergency department or clinic patient (eg, fluid loss through drainage tubes). Third, the model needs assessment at different prevalences of hypovolemia because the starting score of -5 and the scores for the component measures could change as the prior probability deviates from 50%.

Reviewed by David L. Simel, MD, MHS

TITLE Unique Cut Points for Sitting-to-Standing Orthostatic Vital Signs.

AUTHORS Witting MD, Gallagher K.

CITATION *Am J Emerg Med.* 2003;21(1):45-47.

QUESTION Are the thresholds for detecting abnormal vital sign changes when going from supine to standing the same as when going from sitting to standing?

DESIGN Prospective convenience sample compared to prospectively collected data on blood donors.

SETTING Emergency department.

PATIENTS A total of 176 patients in the emergency department, with no cardiovascular symptoms, hypertension, anemia, diabetes, substance abuse, orthostatic hypotension history, or cancer. All patients were presumed to be euvolemic, and none had chest discomfort, dyspnea, palpitations, lightheadedness, vomiting, diarrhea, decreased appetite, abdominal pain, pharyngitis, melena, laceration, or diffuse trauma. The data for supine to standing were from 292 healthy blood donors, obtained before they donated blood.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

Blood pressure (BP) and pulse were measured after the subject had been sitting 5 minutes. The patient then stood up and, after 1 minute, the vital signs were retaken. An automated BP device was used, prohibiting observer variability.

MAIN OUTCOME MEASURES

The specificity of vital sign changes for sitting to standing compared with the vital sign changes from supine to standing.

MAIN RESULTS

The mean change in pulse from sitting to standing was 5.3/min (95% confidence interval [CI], 4.3–6.3/min), whereas the mean change in systolic pressure from sitting to standing was –1.2 mm Hg (95% CI, –0.3 to 2.6 mm Hg). The specificity for both findings was high (Table 24-12).

Table 24-12 Specificity of Findings for Change in Pulse Rate or Pressure as a Function of Posture Change

Test		Cut Point	Specificity (95% CI)
Pulse change	Sitting to standing	+20/min	0.98 (0.94-0.99)
	Supine to standing	+30	0.98 (0.96-0.99)
Systolic blood pressure, mm Hg	Sitting to standing	–20	0.97 (0.92-0.99)
	Supine to standing	–25	0.98 (0.95-0.99)

Abbreviation: CI, confidence interval.

CONCLUSIONS

LEVEL OF EVIDENCE Level 1.

STRENGTHS Simple study that used an automated device to prevent observer variability and bias.

LIMITATIONS Historical control. This study evaluates the magnitude of differences in pulse and systolic BP when the clinician chooses to have the patient go from sitting to standing rather than supine to standing. In normovolemic subjects, a pulse change of 30/min occurs in only 2% to 4% (ie, a specificity of 96%–98%). Patients who go from sitting to standing will not have as great a pulse change. These data show that a threshold of 20/min should be used for these patients. The CI around the beats per minute is slightly narrower than the CI for systolic BP. Thus, the change in pulse should be preferred as the screening test. The authors also evaluated a combination of the 2 findings, pulse/systolic BP (called the shock index); the specificity for changes at adjusted cut points had a wider CI than the pulse.

Reviewed by David L. Simel, MD, MHS

CHAPTER 25

Is This Child Dehydrated?

Michael J. Steiner, MD

Darren A. DeWalt, MD, MPH

Julie S. Byerley, MD, MPH

CLINICAL SCENARIOS

CASE 1 A 20-month-old girl is brought to the emergency department (ED) after 2 days of vomiting and diarrhea. Her father reports that she has not eaten normally since the illness began and now will not drink. She has had 8 stools so far today, but he does not think there were any diapers with urine in them. The child appears mildly ill but does make tears while crying. Her respiratory rate and quality are normal, along with her other vital signs. Her mouth is somewhat dry, capillary refill time is 1.5 seconds, and skin turgor is normal. Her serum (blood) urea nitrogen concentration (BUN) is 12 mg/dL, and bicarbonate concentration is 19 mEq/L.

CASE 2 A 5-month-old boy presents to a health care clinic in a developing country. The child lives in a rural area, and there is no running water in the family home. The child began having nonbloody, profuse, watery stools approximately 7 days ago. The family has World Health Organization (WHO) oral rehydration packets at home that the child has eagerly consumed. He seemed less interested in drinking this morning so his parents began the trip to the clinic. The child is now quiet and hyperpneic. He has sunken eyes and a dry mouth. His capillary refill time is 3 seconds, and his skin turgor is prolonged.

WHY IS THE CLINICAL EXAMINATION IMPORTANT?

Dehydration is one of the leading causes of morbidity and mortality in children throughout the world.^{1,2} Diarrheal disease and dehydration account for as much as 30% of worldwide deaths among infants and toddlers; 8000 children younger than 5 years die each day because of gastroenteritis and dehydration.²⁻⁴ In the United States, children younger than 5 years have an average of 2 episodes of gastroenteritis per year, leading to 2 to 3 million office visits and 10% of all pediatric hospital admissions.^{1,5,6} The direct costs of outpatient and hospital visits are more than \$2 billion per year, not including indirect costs to families and society.⁴ Despite aggressive medical care, as many as 300 US children still die each year as a result of gastroenteritis and associated dehydration.^{1,6}

Many other childhood illnesses in addition to gastroenteritis are associated with dehydration. Gingivostomatitis, bronchiolitis, pyloric stenosis, and focal bacterial infections such as pneumonia, meningitis, and urinary tract infections can all lead to dehydration. For this reason, the morbidity and mortality related to dehydration are actually much higher than that associated solely with gastroenteritis. Dehydration is such a common concern in pediatrics that clinicians in primary care offices, EDs, and hospital settings all assess volume status as part of their evaluation. This assessment helps guide decisions about therapy and patient disposition.

The American Academy of Pediatrics (AAP), Centers for Disease Control and Prevention (CDC), and WHO have all developed treatment guidelines for gastroenteritis according to the clinical assessment of dehydration. The AAP guideline states that “the treatment of a child with diarrhea is directed primarily by the degree of dehydration present.”⁴ They recommend clinically deciding whether a patient is mildly (3%-5%), moderately (6%-9%), or severely ($\geq 10\%$) dehydrated and then treating according to that classification. The CDC uses a similar assessment and scale in its recommendations on the initial management of diarrhea.¹³ WHO has also incorporated signs of dehydration into the Integrated Management of Childhood Illness Scale, which assists practitioners in developing countries in making treatment and referral decisions.⁷

Inaccurate assessment of dehydration can have important consequences. Unrecognized and untreated fluid deficits can create electrolyte disturbances, acidosis, and end-organ damage, including cardiovascular instability, renal insufficiency, and lethargy. These complications can produce devastating results, including permanent injury or death. Conversely, unnecessary interventions can occur after erroneous assessment that a child has moderate or severe dehydration when he or she is actually euvolemic or only mildly dehydrated.⁵ Despite recommendations for oral rehydration in mild or moderate dehydration, this therapy is used in less than 30% of the cases of diarrhea in the United States for which it is indicated.⁸ Clinicians may rely on the more invasive intravenous rehydration in part because they overestimate the degree of dehydration. Both overestimating and underestimating the degree of dehydration can increase health care costs and cause unnecessary morbidity.

Pediatricians generally use the terms *dehydration*, *volume depletion*, and *hypovolemia* interchangeably to represent fluid loss in outpatient settings. Literature that focuses on physiologic changes caused by different types of fluid loss differentiates among these terms.⁹ Because this discrimination can have unclear clinical implications and to simplify discussion, much of the clinical literature combines terminology.¹⁰ Herein, we follow this convention and use the term *dehydration* to represent all fluid deficits except in circumstances such as whole blood loss or significant sodium alteration, in which important clinical implications are evident.

The quantification of dehydration is an important and commonly used skill for assessment of pediatric patients. Despite this importance, the utility of the clinical history, physical examination, and laboratory tests to assess dehydration in children has not been systematically reviewed. Most teaching regarding the assessment of dehydration is based on clinical experience and medical tradition. We conducted a systematic review of the literature on the precision and accuracy of medical history, physical examination, and laboratory tests in identifying dehydration in children between 1 month and 5 years old.

Anatomic/Physiologic Origins of Dehydration Signs

Many signs in pediatric assessment are attributed to the fluid and electrolyte shifts caused by dehydration. Early work to

understand dehydration in children focused on intracellular and extracellular physiologic changes associated with fluid loss. Researchers have fastidiously documented fluid and electrolyte losses in dehydration and have even performed biopsies of the muscle of children with severe diarrhea to understand intracellular fluid and electrolyte shifts.¹¹ Particularly instructive experiments used radiolabeled albumin to demonstrate that the percentage of body weight lost was directly proportional to the percentage of plasma volume lost.¹² For example, children who had lost 5% of their body weight lost approximately 5% of their plasma volume. Because plasma volume is only a small percentage of total body water, this experiment indirectly demonstrated that the majority of fluid lost in childhood dehydration actually comes from either interstitial or intracellular sources.

The correlation of losses from specific fluid compartments to corresponding physical signs has not been clearly documented. The signs of dehydration appear to represent an actual desiccation of tissue (eg, dry mucous membranes), a compensatory reaction of the body to maintain vital perfusion (eg, tachycardia), or some combination of both (eg, capillary refill time). Although some authors offer more specific explanations of theoretic fluid compartments and their examination correlates, these 3 principles should be sufficient for clinical assessment of patients.

How to Elicit Symptoms and Signs

Pediatrics practitioners often elicit historical points from adult caregivers instead of directly from the patient. When assessing volume status in infants, physicians may ask about number of wet diapers (surrogate for urine output), presence or absence of vomiting and diarrhea, and amount and type of oral intake. Caregivers also frequently report their interpretation of examination signs by clarifying whether the child is active, whether the eyes appear sunken, and whether the child drinks vigorously. Clinicians should ask parents whether they have given a successful trial of clear fluids at home, whether the child has been treated by another medical practitioner during the illness, and the date and value of the child's most recent weight measurement.^{1,13}

The ability to elicit some examination signs is impaired when pediatric patients are crying and uncooperative. Therefore, assessment of hydration status should progress from the least to the most invasive maneuvers. The examination should begin with the child across the room in a position of comfort (eg, in the parent's lap). Overall appearance, activity, and response of the child to stimulation should be observed. Evaluating the respiratory pattern is important for assessment of dehydration and all other acute illnesses. Respiratory rate should be measured for 60 seconds by observing chest wall movements. The precise measurement requires a quiet and comfortable child. The rate should be compared with age-based norms.¹⁴ In a potentially dehydrated child, the examiner should specifically look for hyperpnea (deep, rapid breathing without other signs of respiratory distress), suggestive of an acidosis.¹ Other vital signs, including temperature, pulse, and blood pressure, should also be evaluated while the child is comfortable.¹

Next, the clinician should assess skin turgor and capillary refill time. Skin turgor has been used to diagnose dehydration for more than 50 years and, when abnormal, is also called “tenting” or “inelastic skin.”^{15,16} To elicit the sign, the examiner should use the thumb and index finger to pinch a small skin fold on the lateral abdominal wall at the level of the umbilicus.¹⁵ The fold should be promptly released, and then the time is measured for the skin fold’s return to normal form.¹⁵ Clear norms for this time have not been published, and most clinicians simply qualify skin turgor as immediate, slightly delayed, or prolonged.

Excess subcutaneous fat and hypernatremia may falsely normalize the turgor in dehydrated children, whereas malnutrition may falsely prolong the recoil time.^{15,17-21} Primary skin disorders complicate the interpretation of skin turgor.¹⁹

To assess capillary refill time, the examiner compresses a superficial capillary bed and estimates the time it takes for normal color to return after the pressure is released. Capillary refill time varies as a function of ambient temperature, site of application, lighting, medications, and primary (eg, reflex sympathetic dystrophy) or secondary (eg, cardiogenic shock) autonomic changes.^{16,18,22-24} Extremes in patient temperature may also affect the capillary refill time; for example, capillary refill times are markedly prolonged after cold immersion.²⁵ However, Gorelick et al²² found that fever did not affect the test characteristics in children with vomiting, diarrhea, or poor oral intake. According to the available studies, and to standardize examination techniques, we recommend assessing capillary refill time on a finger with the arm at the level of the heart in a warm ambient temperature. Pressure should be gradually increased on the palmar surface of the distal fingertip and then released immediately after the capillary bed blanches. The time elapsed until restoration of normal color should be estimated. Although many practitioners use other sites to measure capillary refill time, most

studies of this sign use the palmar surface of the distal fingertip.²²⁻²⁶ Using this approach, values for nondehydrated children are less than 1.5 to 2 seconds.²⁵

METHODS

Search Strategy and Quality Review

We identified articles by direct searches of the MEDLINE database via the PubMed search engine. The first and broadest search strategy used “dehydration” and “diagnosis,” “hypovolemia” and “diagnosis,” or “intravascular volume depletion” and “diagnosis.” All searches were limited by age (all children: 0-18 years) and publication date (January 1966 to April 2003). These searches produced 1537 articles. We supplemented this preliminary search with the standardized search technique used in The Rational Clinical Examination series (available from the authors). This second search produced 24 additional articles.

Each of the authors reviewed the titles and available abstracts from the 1561 articles, selecting for further review those that appeared to address the evaluation of dehydration in children aged 1 month to 5 years. We did not exclude articles if the study enrolled some children outside that age range. Through consensus, we identified 68 articles as potential sources of primary data or reviews with potential background information and thorough reference lists.

To ensure a comprehensive literature review, we used additional techniques to identify articles (Figure 25-1). One author (M.J.S.) searched for individual symptoms and signs associated with the diagnosis of dehydration in children. These terms included “capillary refill,” “skin turgor,” “dry cry,” “tears,” “mucous membrane,” “sunken eyes,” “fontanelle” and “dehydration,” “urine specific gravity,” “urine” and “dehydration,” “hemoconcentration,” “BUN,” “urine,” “blood

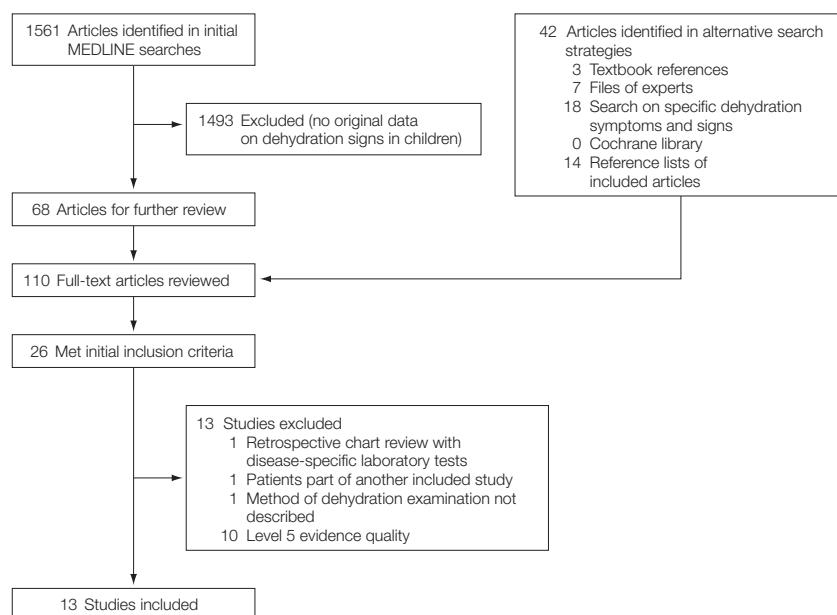


Figure 25-1 Selection Process for Studies Included in Review

pressure," "bioimpedance," "orthostasis," "respiration," "parent" and "dehydration," "pulse," and "heart rate" (all limit: aged 0-18 years, human, NOT "dehydration" and "diagnosis"). The Cochrane Library, reference lists of pediatric and physical examination textbooks,²⁷⁻³² reference lists of all included articles, and articles from the collections of experts in the field were reviewed. Forty-two potential articles were identified from the supplemental searches.

We performed a full review of the 110 retained articles to identify those with primary data comparing dehydration with a symptom, sign, or laboratory value in pediatric patients. Twenty-six articles met these criteria and underwent a full quality assessment with an established methodologic filter that has been consistently used and described in The Rational Clinical Examination series (see Table 1-7).³³ A second author then checked the initial quality review. The group always arrived at a consensus on the final evidence quality level assigned.

Nine of the 110 articles that underwent a full-text review were written in languages other than English. Medical school faculty, residents, or students at our institution who were primary speakers of the written language read each of these articles. Six of these 9 articles did not meet inclusion criteria and were excluded, whereas 3 were assigned an evidence quality level according to a translation of the article.

No studies on physical examination signs, symptoms, or laboratory results in childhood dehydration demonstrated

evidence quality criteria for level 1 or 2. Four studies were assigned to level 3, but one of these was eventually excluded because the study population overlapped with that in another included study.²² Twelve studies were initially assigned to level 4, although one was excluded because of methodologic flaws¹² and another was excluded because of its retrospective design and restriction to children with pyloric stenosis.³⁴

We chose the difference between the rehydration weight and the acute weight divided by the rehydration weight as the best available gold standard of percentage of volume lost.³⁵ Ten articles used gold standards based solely on examination signs or a general dehydration assessment. These were assigned an evidence quality level of 5 and were subsequently excluded. Figure 25-1 shows a schematic representation of the methods, and Table 25-1 summarizes the 13 included studies.

Statistical Analyses

We report precision data as a range of κ values obtained directly from the published results. Two-by-two tables were created from the published information regarding accuracy and were used to calculate point estimates and 95% confidence intervals (CIs) for the sensitivity, specificity, and likelihood ratios (LRs) for each test.⁴⁴ One author provided original data to calculate these values because they were not calculable from the original publication.¹⁸ We created these 2×2 tables for detecting both 5% and 10% dehydration

Table 25-1 Summary of Included Studies

Source, y	Evidence Quality Level	Country	Setting	No. of Participants	Age Range	Inclusion Criteria
Porter et al, ¹³ 2003	3	United States	Emergency department	71	1 mo-5 y	Chief complaint of vomiting, diarrhea, or poor oral intake
Laron, ¹⁵ 1957	4	United States	Hospital	21	1 mo-3.5 y	Admitted with diarrhea
Saavedra et al, ¹⁶ 1991	4	United States	Hospital	32	2-24 mo	Admitted with diarrhea
Duggan et al, ¹⁸ 1996	4	Egypt	Gastroenteritis clinic	135	3-18 mo	Acute diarrhea and dehydrated
Gorelick et al, ³⁵ 1997	3	United States	Emergency department	225	1 mo-5 y	Chief complaint of vomiting, diarrhea, or poor oral intake
Duggan et al, ³⁶ 1997 (precision only)	3	Egypt	Gastroenteritis clinic	100	2 mo-4 y	>5 Stools in last 24 h
MacKenzie et al, ³⁷ 1989	4	Australia	Hospital	102	<4 y	Admitted with gastroenteritis and dehydration
English et al, ³⁸ 1997	3	Kenya	Hospital	119	>1 mo	Admitted with malaria and coma, respiratory distress, or prostration
Plata Rueda and Diaz Cruz, ³⁹ 1974	4	Columbia	Hospital	100	<73 mo	Admitted with diarrhea and dehydration
Vega and Avner, ⁴⁰ 1997	4	United States	Emergency department	97	2 wk-15 y	Dehydrated and needed intravenous fluids
Amin et al, ⁴¹ 1980	4	Indonesia	Hospital	36	<24 mo	Admitted with diarrhea and dehydration
Teach et al, ⁴² 1997	4	United States	Emergency department	40	2 wk-12 y	Dehydrated and needed intravenous fluids
Yilmaz et al, ⁴³ 2002	4	Turkey	Emergency department	168	1-21 mo	Received intravenous fluids and hospitalized for gastroenteritis and dehydration

when data were available. A range of values was provided when only 2 studies evaluated an individual diagnostic test. If more than 2 studies evaluated a test, then we combined the results with a random-effects model. Data for meta-analysis were not weighted according to the quality of included studies. Statistical tests were performed with STATA software, version 7.0 (StataCorp, College Station, Texas).

We performed tests of heterogeneity for data used in all meta-analyses and found significant heterogeneity for most signs. Analysis of data with a random-effects model is complicated by the presence of heterogeneity. However, combining data in this manner allows clinicians to make general summary “best estimates” of utility according to all of the included studies. Furthermore, the degree of uncertainty between LRs of summary estimates was more obvious with the broad range of 95% CIs as opposed to the narrower range for the individual point estimates. Thus, the summary LRs lower the risk of clinicians being overly confident about the utility of clinical findings.

RESULTS

Precision of Symptoms and Signs

Porter et al¹³ evaluated the agreement between parental observation of examination signs and the signs elicited by trained ED nurses. The κ value demonstrated substantial agreement beyond chance when assessing for a sunken anterior fontanelle ($\kappa = 0.73$) and presence of cool extremities ($\kappa = 0.70$). There was moderate agreement on general appearance ($\kappa = 0.46$), presence of sunken eyes ($\kappa = 0.49$), absence of tears ($\kappa = 0.57$), and presence of dry mouth ($\kappa = 0.52$).

Three included studies reported interrater agreement among clinicians, ranging from chance to good agreement (Table 25-2).^{16,35,36} Agreement on respiratory rate and pattern may be no better than that which occurs by chance. The other signs had higher levels of agreement, although the range of κ levels for these findings was broad.

Accuracy of Symptoms, Signs, and Laboratory Studies

Symptoms

Three studies evaluated the accuracy of history-taking in assessing dehydration.^{13,35,37} All 3 of these studies evaluated history of low urine output as a test for dehydration. In the pooled analysis, low urine output did not increase the likelihood of 5% dehydration (LR, 1.3; 95% CI, 0.9-1.9). Porter et al¹³ showed that a history of vomiting, diarrhea, decreased oral intake, reported low urine output, a previous trial of clear liquids, and having seen another clinician during the illness before presenting to the ED yielded LRs that lacked utility in the assessment of dehydration. However, their data did suggest that children who had not been previously evaluated by a physician during the illness might be less likely to be dehydrated on presentation (LR, 0.09; 95% CI, 0.01-1.4). Similarly, parental report of a normal urine output decreases the likelihood of dehydration (Gorelick et al³⁵ reported an LR of 0.27 [95% CI, 0.14-0.51] and Porter et al¹³ reported an LR of 0.16 [95% CI, 0.01-2.5]).

Table 25-2 Precision of Examination Signs for Dehydration

Finding	Reference	Total No. of Participants	Range of κ Values
Prolonged capillary refill	16, 35, 36	216	0.01 to 0.65
Abnormal skin turgor	35, 36	184	0.36 to 0.55
Abnormal respiratory pattern	35, 36	184	-0.04 to 0.40
Extremity perfusion	35	100	0.23 to 0.66
Absent tears	35, 36	184	0.12 to 0.75
Sunken fontanelle	36	100	0.10 to 0.27
Sunken eyes	35, 36	184	0.06 to 0.59
Dry mucous membranes	35, 36	184	0.28 to 0.59
Weak pulse	35, 36	184	0.15 to 0.50
Poor overall appearance	35, 36	184	0.18 to 0.61

Examination Signs

Table 25-3 is a comprehensive list of individual physical examination signs and their test characteristics in evaluating children for 5% dehydration. Signs were included when they were evaluated in 2 or more studies, and calculations based on pooled results were performed when evaluated in 3 or more studies.

Three signs were evaluated in multiple studies, had a clinically helpful pooled LR in detecting 5% dehydration, and had 95% CIs wholly above 1.0. Capillary refill time was evaluated in 4 studies, and the pooled sensitivity of prolonged capillary refill time was 0.60 (95% CI, 0.29-0.91), with a specificity of 0.85 (95% CI, 0.72-0.98), for detecting 5% dehydration.^{16,35,37,38} The LR for abnormal capillary refill time was 4.1 (95% CI, 1.7-9.8). This was the highest value among examination signs with pooled results. Abnormal skin turgor had a pooled LR of 2.5 (95% CI, 1.5-4.2)^{15,18,35,37,38} and abnormal respiratory pattern had a pooled LR of 2.0 (95% CI, 1.5-2.7).^{18,35,37,38}

Presence of cool extremities or a weak pulse or absence of tears also may be helpful tests for dehydration. Absence of tears had a pooled LR of 2.3 (95% CI, 0.9-5.8), but the potential utility is limited by a wide 95% CI that crosses 1.0.^{13,35,37} Two studies examined a weak pulse quality as a test for dehydration. One study found a reasonably precise LR for weak pulse of 3.1 (95% CI, 1.8-5.4),³⁵ but in the other study, the 95% CI was too wide to make a reasonable estimate (LR, 7.2; 95% CI, 0.4-150).¹⁸ The 2 studies that evaluated cool extremities as a test of dehydration found imprecise point estimates for the positive likelihood ratio (LR+) in detecting 5% dehydration (LR, 19; 95% CI,¹⁸ 1.1-330 and LR, 1.5; 95% CI,¹³ 0.2-12).

Sunken eyes and dry mucous membranes offer little help clinically; both had narrow 95% CIs but pooled LRs of 1.7. An increased heart rate, a sunken fontanelle in young infants, and an overall poor appearance are frequently taught as good tests for dehydration. However, the objective evidence reveals that all have summary LRs of less than 2.0 and 95% CIs that cross 1.0.

Some tests may be clinically useful in decreasing the likelihood of dehydration. Absence of dry mucous membranes (LR, 0.41; 95% CI, 0.21-0.79), a normal overall appearance

Table 25-3 Summary Test Characteristics for Clinical Findings to Detect 5% Dehydration

Finding	Reference	Total No. of Participants	LR Summary, Value (95% CI) or Range		Sensitivity (95% CI)	Specificity (95% CI)
			Present	Absent		
Prolonged capillary refill	16, 35, 37, 38	478	4.1 (1.7-9.8)	0.57 (0.39-0.82)	0.60 (0.29-0.91)	0.85 (0.72-0.98)
Abnormal skin turgor	15, 18, 35, 37, 38	602	2.5 (1.5-4.2)	0.66 (0.57-0.75)	0.58 (0.40-0.75)	0.76 (0.59-0.93)
Abnormal respiratory pattern	18, 35, 37, 38	581	2.0 (1.5-2.7)	0.76 (0.62-0.88)	0.43 (0.31-0.55)	0.79 (0.72-0.86)
Sunken eyes	13, 18, 35, 37	533	1.7 (1.1-2.5)	0.49 (0.38-0.63)	0.75 (0.62-0.88)	0.52 (0.22-0.81)
Dry mucous membranes	13, 18, 35, 37	533	1.7 (1.1-2.6)	0.41 (0.21-0.79)	0.86 (0.80-0.92)	0.44 (0.13-0.74)
Cool extremity	13, 18	206	1.5-19	0.89-0.97	0.10-0.11	0.93-1.0
Weak pulse	18, 35	360	3.1-7.2	0.66-0.96	0.04-0.25	0.86-1.0
Absent tears	13, 35, 37	398	2.3 (0.9-5.8)	0.54 (0.26-1.1)	0.63 (0.42-0.84)	0.68 (0.43-0.94)
Increased heart rate	18, 35, 37	462	1.3 (0.8-2.0)	0.82 (0.64-1.0)	0.52 (0.44-0.60)	0.58 (0.33-0.82)
Sunken fontanelle	13, 18, 37	308	0.9 (0.6-1.3)	1.12 (0.82-1.5)	0.49 (0.37-0.60)	0.54 (0.22-0.87)
Poor overall appearance	13, 35, 37	398	1.9 (0.97-3.8)	0.46 (0.34-0.61)	0.80 (0.57-1.0)	0.45 (-0.1 to 1.0)

Abbreviations: CI, confidence interval; LR, likelihood ratio.

(LR, 0.46; 95% CI, 0.34-0.61), and absence of sunken eyes (LR, 0.49; 95% CI, 0.38-0.63) had pooled LRs of less than 0.5. Most clinical scenarios require lower LRs than these to rule out dehydration effectively.

Four studies evaluated clinical prediction models or groups of signs.^{18,35,39,40} Vega and Avner⁴⁰ evaluated a table similar to that used in many pediatric textbooks and also commonly taught to medical students as the best evaluation tool for dehydration.³⁰ This scale, displayed in Table 25-4, is similar to the one used by the AAP and CDC in their recommendations for the management of acute gastroenteritis.^{1,3,4} The tool uses the assessment of 9 physical examination findings to classify chil-

dren as mildly (4%-5%), moderately (6%-9%), or severely ($\geq 10\%$) dehydrated. In 97 children presenting to the ED with dehydration requiring intravenous fluids, a classification of severe on the scale had an LR of 3.4 (95% CI, 1.5-7.7) for the presence of at least 5% dehydration. Classification of severe dehydration also yielded an LR of 4.3 (95% CI, 2.4-7.8) for at least 10% dehydration. A moderate classification by examination was less useful to diagnose 5% dehydration (LR, 2.1; 95% CI, 0.9-4.8).⁴⁰

Duggan et al¹⁸ evaluated 2 dehydration assessment scales that classified children as mild, moderate, or severe according to the number of dehydration examination signs present. The authors reported the final mean percentage of dehydration within each group, and these averages increased significantly as the severity assessment increased,¹⁸ which suggests that as more signs of dehydration appear, children tend to be more dehydrated. Plata Rueda and Diaz Cruz³⁹ also presented groupings of signs and symptoms that attempted to stratify children into different degrees of dehydration. Minor physical examination changes did not significantly change the likelihood of dehydration; however, the presence of abnormal skin turgor on the abdomen, thorax, extremities, and face, combined with sunken eyes, dry mucous membranes, and a sunken fontanelle, did increase the likelihood of 10% dehydration (LR, 3.7; 95% CI, 1.6-8.1).³⁹

Gorelick et al³⁵ created a scale giving equal weight to 10 commonly elicited signs: decreased skin elasticity, capillary refill time greater than 2 seconds, general appearance, absence of tears, abnormal respirations, dry mucous membranes, sunken eyes, abnormal radial pulse, tachycardia (heart rate $> 150/\text{min}$), and decreased urine output. The presence of at least 3 of the 10 signs had a sensitivity of 0.87 and a specificity of 0.82 in detecting 5% dehydration (LR+, 4.9; 95% CI, 3.3-7.2, and negative LR, 0.15; 95% CI, 0.08-0.30). Similarly, 7 of 10 signs had an LR+ of 8.4 (95% CI, 5.0-14) in diagnosing 10% dehydration. A logistic regression analysis performed by Gorelick et al³⁵ showed that capillary refill time, dry mucous membranes, absence of tears, and abnormal overall appearance contained most of the predictive power. A simplified assessment tool

Table 25-4 Example of a Commonly Taught Dehydration Assessment Scale^a

Variable/Sign	Dehydration		
	Mild (4%-5%)	Moderate (6%-9%)	Severe ($\geq 10\%$)
General appearance	Thirsty, restless, alert	Thirsty, drowsy, postural hypotension	Drowsy, limp, cold, sweaty, cyanotic extremities
Radial pulse	Normal rate and strength	Rapid and weak	Rapid, thready, sometimes impalpable
Respirations	Normal	Deep, may be rapid	Deep and rapid
Anterior fontanelle	Normal	Sunken	Very sunken
Systolic blood pressure	Normal	Normal or low	Low
Skin elasticity	Pinch retracts immediately	Pinch retracts slowly	Pinch retracts very slowly
Eyes	Normal	Sunken	Grossly sunken
Tears	Present	Absent	Absent
Mucous membranes	Moist	Dry	Very dry

^aAdapted from Vega and Avner,⁴⁰ with permission.

Table 25-5 Summary Test Characteristics for Laboratory Tests Assessing Dehydration

Laboratory Value	Reference	Total No. of Participants	Sensitivity, Value (95% CI) or Range	Specificity, Value (95% CI) or Range	LR Summary, Value (95% CI) or Range	
					Present	Absent
Blood urea nitrogen, mg/dL						
>8	37, 38		0.38-0.71	0.71-0.82	2.1-2.4	0.41-0.76
>18	41, 43		0.63-0.90	0.55-0.57	1.4-2.1	0.17-0.68
>27	41	36	0.44 (0.19-0.68)	0.85 (0.69-1.0)	2.9 (0.9-9.5)	0.66 (0.41-1.1)
>45	43	168	0.43 (0.34-0.52)	0.99 (0.96-1.0)	46.1 (2.9-733)	0.58 (0.49-0.68)
Blood urea nitrogen/creatinine ratio > 40						
	42	40	0.23 (0.01-0.46)	0.89 (0.77-1.0)	2.1 (0.5-8.9)	0.87 (0.62-1.2)
Bicarbonate, mEq/L						
<17	40	97	0.83 (0.72-0.94)	0.76 (0.64-0.88)	3.5 (2.1-5.8)	0.22 (0.12-0.43)
<15	43	168	0.93 (0.88-0.98)	0.40 (0.26-0.53)	1.5 (1.2-1.9)	0.18 (0.08-0.37)
Base deficit > 7 mEq/L						
	37, 38		0.67-0.75	0.52-0.59	1.4-1.8	0.42-0.68
pH < 7.35	37	102	0.43 (0.28-0.58)	0.80 (0.70-0.91)	2.2 (1.2-4.1)	0.71 (0.53-0.95)
Anion gap > 20 mEq/L	42	40	0.46 (0.19-0.73)	0.74 (0.58-0.91)	1.8 (0.8-4.2)	0.73 (0.42-1.3)
Uric acid > 10 mg/dL	42	40	0.23 (0.01-0.46)	0.78 (0.62-0.93)	1.0 (0.3-3.5)	0.99 (0.69-1.4)

Abbreviations: CI, confidence interval; LR, likelihood ratio.

using the presence of 2 of these 4 signs yielded an LR+ of 6.1 (95% CI, 3.8-9.8) for diagnosing 5% dehydration.³⁵

Laboratory Tests

Six studies evaluated the utility of laboratory tests in assessing dehydration (Table 25-5).^{37,38,40-43} Five studies evaluated BUN concentration or BUN/serum creatinine ratio as a test for dehydration.^{37,38,41-43} BUN cutoffs of 8, 18, and 27 mg/dL produced LRs ranging from 1.4 to 2.9. Yilmaz et al⁴³ found that in a group of hospitalized children with gastroenteritis, BUN greater than 45 mg/dL was specific for at least 5% dehydration (specificity of 1.0). However, this was a small study and the estimated 95% CI for an LR+ was 3 to 730.

Four studies evaluated acidosis as a test for dehydration.^{37,38,40,43} Most patients enrolled in these studies had acute diarrhea, a potential cause of acidosis. Mackenzie et al³⁷ and English et al³⁸ used a base deficit of greater than 7 as the measure of acidosis. (Base deficit estimates the severity of metabolic acidosis by comparing the patient's bicarbonate concentration to historical norms for a given pH and PCO₂.) In both studies, the LR+ was less than 2.0. Although Yilmaz et al⁴³ found that an absolute serum bicarbonate concentration of less than 15 mEq/L was not helpful (LR for low serum bicarbonate, 1.5; 95% CI, 1.2-1.9), Vega and Avner⁴⁰ found that an absolute bicarbonate concentration of less than 17 mEq/L offered some help in diagnosing children with 5% dehydration (LR, 3.5; 95% CI, 2.1-5.8). Teach et al⁴² evaluated serum uric acid and an increased anion gap as tests for dehydration but found that abnormal results were not helpful. Urine specific gravity was evaluated by English et al³⁸ but was not found to be significantly correlated with dehydration. The only laboratory measurement that appears to be valuable in decreasing the likelihood of 5% dehydration is serum bicarbonate. A serum bicarbonate concentration of more than 15 or 17 mEq/L has

an LR range of 0.18 to 0.22, reducing the likelihood of dehydration if the child has gastroenteritis.^{40,43}

Limitations

The published literature on assessment of dehydration has significant limitations affecting both internal and external validity. As discussed in the “Methods” section, none of the identified studies met the criteria for high-quality (level 1 or level 2) evidence according to the established methodologic filter. The best available studies had modest sample sizes, used nonconsecutive patients, and did not compare the included children with those excluded from the study populations. The most common bias in level 4 evidence studies was that they enrolled children already thought to be dehydrated and to need intravenous fluids or who were admitted to the hospital. The diagnostic tests may perform better in children who are thought to be dehydrated compared with children solely at risk of dehydration. Thus, there may be limitations to the generalizability of these results when applied to an unselected group of children simply at risk of dehydration.

The results of the study by Gorelick et al³⁵ differed from those of the other included studies. Gorelick et al³⁵ evaluated the interrater reliability for 10 physical examination signs. The κ values ranged from 0.40 to 0.75, which were clearly better than those found in the other studies on precision by Saavedra et al¹⁶ and Duggan et al.³⁶ The accuracy of signs was also generally better in the study by Gorelick et al³⁵ than in other included studies. The LRs of positive tests were all statistically significant and ranged from 1.8 to 12. All 10 of the signs evaluated by Gorelick et al³⁵ were assessed in other studies. For 9 of the 10 signs, the results by Gorelick et al³⁵ produced the highest LRs of any included study, which is difficult to explain. The study by Gorelick et al³⁵ is of high methodologic quality in comparison with the other included studies. It achieved an

evidence quality level 3 according to nonconsecutive patient selection that did not introduce a clear systematic bias. They enrolled a relatively large group of patients and followed them meticulously. The sensitivity values of the tests were generally similar to those found in other studies, but the specificity was often much higher. The high percentage of true-negative test results may have been affected by a patient population with a relatively low incidence of disease in comparison with patients enrolled in the other studies.³⁵

Ten of the 26 articles that met initial inclusion criteria were later found to have a methodologic flaw with the diagnostic standard and were excluded from the final analysis. These studies used a gold standard for dehydration according to examination signs or clinical assessment, which represents a circular flaw in assessing the utility of the history taking or examination in establishing dehydration. Conversely, the difference between an ill weight and a rehydrated weight (after illness) appears to be the best pragmatic diagnostic standard for dehydration that has been validated in the literature.³⁵ However, problems can be introduced by the timing of the rehydration weight. For example, if it is obtained too early, children may still be dehydrated or may actually be overhydrated because of aggressive intravenous fluid administration. The timing of the rehydration weight varied among the included studies, and most studies used additional assessments to validate their perception of a true rehydration weight. For example, Teach et al⁴² used the weight when the physical examination findings had normalized and the urine-specific gravity level was low. Incorporating other assessments that were not based on weight into the gold standard could bias the results. Some studies avoided this problem by documenting the rehydration weight when measured weight remained unchanged over time.³⁵ Another criticism of a weight-based gold standard is that infants may “gain” a significant percentage of their body weight if they have a full bladder and colon, which they may then “lose” when they void.²⁰ In studies of large sample size, the weight contribution of a full bladder would be unlikely to have a major effect on the LRs for clinical findings. Additionally, the number of children with weight “gained” or “lost” because of impending or recent voids should balance.

Pediatricians are taught that hypernatremia may alter the test characteristics of signs in dehydration.³⁰ For example, prolonged skin turgor is less sensitive in detecting significant dehydration in children with diabetes insipidus and pure water loss than in children with diarrhea.¹⁵ Because of this clinical experience, some studies excluded children with significant hypernatremia.^{35,38} Other studies used subgroup analysis to demonstrate that assessment had not been affected by hypernatremia.^{37,43} Because tests of dehydration are usually applied without any knowledge of the serum sodium level in the patient, it seems appropriate to structure studies without excluding hypernatremic children.

THE BOTTOM LINE

Dehydration is an important cause of morbidity and mortality as a complication of pediatric illness. However, the

literature evaluating the symptoms, signs, and laboratory values for assessing dehydration is limited. We found few high-quality studies with accurate gold standards and minimal systematic bias.

The evidence shows that tests of dehydration are imprecise, generally showing only fair to moderate agreement among examiners. Historical points have moderate sensitivity as a screening test for dehydration. However, parental reports of dehydration symptoms are so nonspecific that they may not be clinically useful. The best 3 individual examination signs for assessing dehydration are prolonged capillary refill time, abnormal skin turgor, and abnormal respiratory pattern. Groups of signs or use of clinical scales improves diagnostic characteristics. Commonly obtained laboratory tests such as BUN and bicarbonate concentrations generally are only helpful when results are markedly abnormal. A normal bicarbonate concentration helps somewhat to reduce the likelihood of dehydration. These laboratory tests should not be considered definitive for dehydration.

The literature reports more than 30 potential tests for detecting dehydration. This large number should not distract clinicians from focusing on signs and symptoms with proven diagnostic utility. Unfortunately, the data also suggest that signs of dehydration can be imprecise and inaccurate, making clinicians unable to predict the exact degree of dehydration. For this reason, we agree with the WHO and other groups that recommend using the physical examination to classify dehydration as “none,” “some,” or “severe.”^{1,45} This general assessment can then be used to guide clinical management.

CLINICAL SCENARIOS—RESOLUTIONS

CASE 1 The historical clues provided by the father are minimally helpful in assessing the child’s dehydration. There are no signs present that increase the likelihood of dehydration. The negative LRs associated with the absence of multiple examination signs and the serum bicarbonate concentration of 19 mEq/L make significant dehydration much less likely. This child probably has “no” dehydration instead of “some” or “severe” dehydration.

CASE 2 The hyperpnea, prolonged capillary refill time, and delayed skin turgor all increase the likelihood of dehydration. Because there are multiple signs of dehydration, the possibility of severe dehydration should be considered and treated appropriately.

Author Affiliations at the Time of the Original Publication

Departments of Pediatrics (Drs Steiner, DeWalt, and Byerley) and Internal Medicine (Drs Steiner and DeWalt), University of North Carolina School of Medicine, and Robert Wood Johnson Clinical Scholars Program, University of North Carolina (Dr DeWalt), Chapel Hill. Dr Steiner is now at the Clinica Sierra Vista, Bakersfield, California.

Acknowledgments

Dr DeWalt received salary support from the Robert Wood Johnson Clinical Scholars Program.

We thank Kenneth B. Roberts, MD, Lee S. Benjamin, MD, Peter Margolis, MD, Anita K. Ying, MD, and William Miller, MD, PhD, for their thoughtful comments on drafts of the manuscript. We also thank Marcos Mestre, MD, Myriam F. Bauer, MD, Sandro Pinheiro, PhD, Kaori Longphre, MD, and Pia J. Hauk, MD, for their assistance in translating articles written in languages other than English.

REFERENCES

- King CK, Glass R, Bresce JS, Duggan C. Managing acute gastroenteritis among children. *MMWR Recomm Rep*. 2003;52(RR-16):1-16.
- Black RE, Morris SS, Bryce J. Where and why are 10 million children dying every year? *Lancet*. 2003;361(9376):2226-2234.
- Duggan C, Santoshamb M, Glass RI. The management of acute diarrhea in children. *MMWR Recomm Rep*. 1992;41(RR-16):1-20.
- American Academy of Pediatrics, Provisional Committee on Quality Improvement, Subcommittee on Acute Gastroenteritis. Practice parameter: the management of acute gastroenteritis in young children. *Pediatrics*. 1996;97:424-435.
- McConnochie KM, Conners GP, Lu E, Wilson C. How commonly are children hospitalized for dehydration eligible for care in alternative settings? *Arch Pediatr Adolesc Med*. 1999;153(12):1233-1241.
- Glass RI, Lew JF, Gangarosa RE, LeBaron CW, Ho MS. Estimates of morbidity and mortality rates for diarrheal diseases in American children. *J Pediatr*. 1991;118(4 pt 2):S27-S33.
- Gove S. Integrated management of childhood illness by outpatient health workers: technical basis and overview. *Bull World Health Organ*. 2003;75:7-24.
- Santoshamb M, Keenan EM, Tulloch J, et al. Oral rehydration therapy for diarrhea: an example of reverse transfer of technology. *Pediatrics*. 1997;100(5):E10.
- Mange K, Matsuura D, Cizman B, et al. Language guiding therapy: the case of dehydration versus volume depletion. *Ann Intern Med*. 1997;127(9):848-853.
- McGee S, Abernathy WB, Simel DL. Is this patient hypovolemic? *JAMA*. 1999;281(11):1022-1029.
- Darrow DC. The retention of electrolyte during recovery from severe dehydration due to diarrhea. *Pediatrics*. 1949;3:129-156.
- Czaczkes JW. Plasma volume as an index of total fluid loss. *Am J Dis Child*. 1961;102:190-193.
- Porter S, Fleisher G, Kohane I, Mandl K. The value of parental report for diagnosis and management of dehydration in the emergency department. *Ann Emerg Med*. 2003;41(2):196-205.
- Margolis P, Gadomski A. Does this infant have pneumonia? *JAMA*. 1998;279(4):308-313.
- Laron Z. Skin turgor as a quantitative index of dehydration in children. *Pediatrics*. 1957;19(5):816-822.
- Saavedra JM, Harris GD, Li S, Finberg L. Capillary refilling (skin turgor) in the assessment of dehydration. *Am J Dis Child*. 1991;145(3):296-298.
- Bruck E, Abal G, Aceto Jr T. Therapy of infants with hypertonic dehydration due to diarrhea. *Am J Dis Child*. 1968;115(3):281-301.
- Duggan C, Refat M, Hashem M, et al. How valid are clinical signs of dehydration in infants? *J Pediatr Gastroenterol Nutr*. 1996;22(1):56-61.
- Dorrington KL. Skin turgor: do we understand the clinical sign? *Lancet*. 1981;1(8214):264-266.
- Leung A, Robson W. Clinical signs of dehydration in children. *Lancet*. 1989;2(8670):1038.
- Dehydration and fat babies. *BMJ*. 1971;1(5741):125.
- Gorelick MH, Shaw KN, Murphy KO, Baker MD. Effect of fever on capillary refill time. *Pediatr Emerg Care*. 1997;13(5):305-307.
- Gorelick MH, Shaw K, Baker M. Effect of ambient temperature on capillary refill in healthy children. *Pediatrics*. 1993;92(5):699-702.
- Brown L, Prasad N, Whitley T. Adverse lighting condition effects on the assessment of capillary refill. *Am J Emerg Med*. 1994;12(1):46-47.
- Schriger DL, Baraff L. Defining normal capillary refill: variation with age, sex, and temperature. *Ann Emerg Med*. 1988;17(9):932-935.
- Tibby S, Hatherill M, Murdoch I. Capillary refill and core-peripheral temperature gap as indicators of haemodynamic status in paediatric intensive care patients. *Arch Dis Child*. 1999;80(2):163-166.
- DeGowin R. *DeGowin & DeGowin's Diagnostic Examination*. 6th ed. New York, NY: McGraw-Hill Co; 1994.
- Nechyba C. *Harriet Lane Pediatric Handbook*. 16th ed. St Louis, MO: Elsevier Science; 2002.
- Sapira J. *The Art and Science of Bedside Diagnosis*. Baltimore, MD: Urban & Schwarzenberg; 1990.
- Behrman R, Kliegman R, Jenson H. *Nelson Textbook of Pediatrics*. 16th ed. Philadelphia, PA: WB Saunders Co; 2000.
- Bates B, Bickley L, Hoekelman R. *A Guide to Physical Examination and History Taking*. 6th ed. Philadelphia, PA: JB Lippincott Co; 1995.
- Zitelli B, Davis H. *Atlas of Pediatric Physical Diagnosis*. 4th ed. St Louis, MO: Mosby; 2002.
- Holleman DR, Simel DL. Does the clinical examination predict airflow limitation? *JAMA*. 1995;273(4):313-319.
- Dawson KP, Graham D. The assessment of dehydration in congenital pyloric stenosis. *N Z Med J*. 1991;104(910):162-163.
- Gorelick MH, Shaw KN, Murphy KO. Validity and reliability of clinical signs in the diagnosis of dehydration in children. *Pediatrics*. 1997;99(5):E6.
- Duggan C, Refat M, Hashem M, et al. Interrater agreement in the assessment of dehydration in infants. *J Trop Pediatr*. 1997;43(2):119-121.
- Mackenzie A, Barnes G, Shann F. Clinical signs of dehydration in children. *Lancet*. 1989;2(8663):605-607.
- English M, Waruiru C, Mwakesi R, Marsh K. Signs of dehydration in severe childhood malaria. *Trop Doct*. 1997;27(4):235-236.
- Plata Rueda E, Diaz Cruz G. Clinical and biochemical evaluation of the degree of dehydration in children with acute diarrhea [in Spanish]. *Bol Med Hosp Infant Mex*. 1974;31(3):561-576.
- Vega RM, Avner JR. A prospective study of the usefulness of clinical and laboratory parameters for predicting percentage of dehydration in children. *Pediatr Emerg Care*. 1997;13(3):179-182.
- Amin SS, Jusniar B, Suharjono. Blood urea nitrogen (B.U. N.) in gastroenteritis with dehydration. *Paediatr Indones*. 1980;20(3-4):77-82.
- Teach SJ, Yates EW, Feld LG. Laboratory predictors of fluid deficit in acutely dehydrated children. *Clin Pediatr (Phila)*. 1997;36(7):395-400.
- Yilmaz K, Karabocuoglu M, Citak A, Uzel N. Evaluation of laboratory tests in dehydrated children with acute gastroenteritis. *J Paediatr Child Health*. 2002;38(3):226-228.
- Simel DL, Samsa GP, Matchar DB. Likelihood ratios with confidence. *J Clin Epidemiol*. 1991;44(8):763-770.
- World Health Organization. *A Manual for the Treatment of Diarrhoea*. Geneva, Switzerland: World Health Organization Program for Control of Diarrhoeal Diseases; 1990.

This page intentionally left blank

Prepared by Michael J. Steiner, MD
Reviewed by Darren A. DeWalt, MD, MPH

CLINICAL SCENARIO

Worried parents bring a 3-year-old boy who is refusing to eat or drink to your office. His illness started 4 days ago, with temperatures as high as 39°C, increased sleepiness, and decreased oral intake. On examination, his temperature is 38.7°C and he is alert but mildly tachycardic and tachypneic. He has normal skin turgor, although his mucous membranes are dry and his capillary refill is 3 seconds. Also observed on examination are small vesicular and ulcerated lesions on the posterior pharynx and red macules on the hands and feet.

UPDATED SUMMARY ON DEHYDRATION IN CHILDREN

Original Review

Steiner MJ, DeWalt DA, Byerley JS. Is this child dehydrated? *JAMA*. 2004;291(22):2746-2754.

UPDATED LITERATURE SEARCH

We repeated the literature search from April 2004 to March 2006 and found 258 new abstracts, but there were no additional studies of the diagnostic accuracy (both sensitivity and specificity) of the physical examination components or an explicit grouping of findings for predicting the presence of dehydration in young children. One potential article that assessed the validity and reliability of dehydration assessment in children with diabetic ketoacidosis was excluded because it enrolled only 5 children who met the age range criteria of our original search (1 month to 5 years).¹ We identified one article on the precision of individual findings and a second article that contained information on the accuracy and precision of the findings included in a new childhood dehydration scale.^{2,3}

NEW FINDINGS

Details of the update

Friedman et al³ evaluated the measurement properties of 12 findings for dehydration, each measured on a 3-point ordinal

Table 25-6 Rating Scale Based on Severity of 4 Clinical Signs³

Characteristic (Intraclass Correlation Coefficient)	Score		
	0	1	2
General appearance (0.55)	Normal	Thirsty, restless, or lethargic but irritable when touched	Drowsy, limp, cold, sweaty, or comatose
Eyes (0.61)	Normal	Slightly sunken	Very sunken
Mucous membranes, moistness on the tongue (0.71)	Moist	"Sticky"	Dry
Tears (0.66)	Tears	Decreased tears	Absent tears

scale. Nine items occurred frequently enough to merit closer evaluation of differing combinations. The patients were aged 1 to 36 months, with gastroenteritis and clinically diagnosed dehydration. A 4-item scale (Table 25-6) had the best measurement characteristics, as assessed by correlation with change in weight, interobserver variability, discrimination between levels of dehydration, and change after treatment.

The intraclass correlation coefficients (a measure of interobserver variability for items on an ordinal scale) were comparable to the range of κ values reported in the original study. Results of the 4 findings are summed, and if one or more are abnormal, the authors report a sensitivity of 0.85 (95% confidence interval [CI], 0.73-0.97) and a specificity of 0.32 (95% CI, 0.20-0.44) for dehydration at a cut point of greater than or equal to 3% (according to data from the original research; written communication,³ Patricia Parkin, MD, University of Toronto, Canada, April 2006). The sensitivity of this model was similar to that reported by Gorelick et al⁴ for detecting 5% dehydration (sensitivity, 0.79; specificity, 0.87), with much lower specificity, but it is difficult to compare the scales directly because of differing dehydration cutoff levels. The model presented by Gorelick et al⁴ was considered positive for 5% dehydration if any 2 of the following were present: capillary refill greater than 2 seconds, dry mucous membranes, absent tears, or change in general appearance.

The interobserver variability for signs of shock was evaluated in Kenyan children admitted to a pediatric ward.² The

diagnoses were unknown to the 4 independent examining clinicians. Capillary refill time, dry mucous membranes, decreased skin turgor, and sunken eyes each had κ values well within the ranges reported in the original Rational Clinical Examination article.

IMPROVEMENTS IN THE DATA PRESENTED IN THE ORIGINAL PUBLICATION

We rereviewed all of the original studies to establish the pretest probability of dehydration in children. Only 2 studies assessed the prevalence of dehydration (5%) among children presenting for emergency care with diarrhea, vomiting, or poor oral intake. There was heterogeneity in the prevalence (11/71 vs 63/186),^{4,5} but the random-effects summary prevalence provides a useful anchor for infants and children presenting with these symptoms (summary prevalence, 25%; 95% CI, 14%-39%).

We rereviewed the data from Gorelick et al⁴ on the performance of a combination of findings for dehydration. This study had the largest number of children of any high-quality study in our original review. We reported a likelihood ratio (LR) of 6.1 (95% CI, 3.8-9.7) to predict at least 5% dehydration when 2 of 4 signs of dehydration were present. However, we did not provide the LR associated with fewer findings. The presence of 0 to 1 finding has an LR of 0.24 (95% CI, 0.14-0.39), making dehydration less likely. By changing the threshold to greater than or equal to 3 findings, the model predicts more severe dehydration ($\geq 10\%$), with an LR of 4.7 (95% CI, 3.1-7.3).

DIFFERENCES IN THE REFERENCE STANDARD

There have been no changes in the reference standards for dehydration.

RESULTS OF LITERATURE REVIEW

Univariate Findings

There were no new data on the accuracy of individual symptoms and signs of dehydration at a threshold of 5%. When measured on an ordinal scale, the intraclass correlations as measures of reliability are good for general appearance (0.55), presence of dry mucous membranes (0.71), sunken eyes (0.61), tears (0.66), and capillary refill time (0.65).⁶

EVIDENCE FROM GUIDELINES

There have been no updates to the 2003 guideline published by the Centers for Disease Control and Prevention (CDC).⁶ Since the publication of the original article, the American Academy of Pediatrics has retired their clinical practice parameters for the management of acute gastroenteritis⁷ and endorsed the CDC guideline.

CLINICAL SCENARIO—RESOLUTION

The clinical history of this 3-year-old boy puts him at risk for dehydration. It is difficult to establish a pretest probability of dehydration for this child presenting to a clinician's office. However, according to published studies from emergency departments where children were enrolled solely because of potentially dehydrating symptoms, his pretest probability of dehydration can be estimated at 25%.^{4,5} According to our reviews, his prolonged capillary refill and tachypnea independently make dehydration more likely. Although some other clinical signs are normal, his dry mucous membranes, with a prolonged capillary refill time, give him a positive result on the Gorelick clinical scale (LR 6.1; 95% CI, 3.8-9.7). According to these values, the posttest probability of dehydration is 70%, so appropriate treatment should be initiated.

CHILDHOOD DEHYDRATION—MAKE THE DIAGNOSIS

PRIOR PROBABILITY

Dehydration develops progressively, depending on the underlying condition, and therefore, a consistent prior probability of dehydration cannot be established for most general conditions. For infants and children whose parents bring them for emergency care for diarrhea, vomiting, or poor oral intake, the prevalence of at least 5% dehydration is approximately 25% (95% CI, 14%-39%).

POPULATION FOR WHOM CHILDHOOD DEHYDRATION SHOULD BE CONSIDERED

In our initial article, we were unable to identify published parental historical elements that made dehydration more likely. However, vomiting, diarrhea, change in oral intake, decreased urine output, fever, change in mental status, or the presence of potentially dehydrating underlying conditions (eg, diabetes insipidus) prompts an evaluation for dehydration.⁶

DETECTING THE LIKELIHOOD OF CHILDHOOD DEHYDRATION

Accurately identifying the presence of dehydration requires the use of combinations of signs. Combinations of findings can include results being either present or absent or graded on an ordinal scale (eg, 0, 1, 2) and then summed across findings. Each scale must be assessed in comparison with the reference standard. See Table 25-7.

Table 25-7 Likelihood Ratio of Combinations of Findings for Greater Than or Equal to 5% Dehydration

Findings ^a	Positive	$\geq 5\%$ Dehydration	
		LR+ (95% CI)	LR- (95% CI)
Capillary refill time > 2 s, dry mucous membranes, absent tears, altered general appearance	≥ 2 Findings ^a	6.1 (3.8-9.7)	0.24 (0.14-0.39)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

^aWhen 3 to 4 findings are present, the likelihood of severe dehydration ($\geq 10\%$) is 4.7 (95% CI, 3.1-7.3).

REFERENCE STANDARD TESTS

The difference between the “well” weight and the acute weight divided by the well weight represents the standard for the percentage of volume lost because of dehydration.

REFERENCES FOR THE UPDATE

- Koves IH, Neutze J, Donath S, et al. The accuracy of clinical assessment of dehydration during diabetic ketoacidosis in childhood. *Diabetes Care*. 2004;27(10):2485-2487.
- Otieno H, Were E, Ahmed I, Charo E, Brent A, Maitland K. Are bedside features of shock reproducible between different observers? *Arch Dis Child*. 2004;89(10):977-979.
- Friedman JN, Goldman RD, Srivastava R, Parkin PC. Development of a clinical dehydration scale for use in children between 1 and 36 months of age. *J Pediatr*. 2004;145(2):201-207.^a
- Gorelick MH, Shaw KN, Murphy KO. Validity and reliability of clinical signs in the diagnosis of dehydration in children. *Pediatrics*. 1997;99(5):E6.
- Porter SC, Fleisher GR, Kohane IS, Mandl KD. The value of parental report for diagnosis and management of dehydration in the emergency department. *Ann Emerg Med*. 2003;41(2):196-205.
- King CK, Glass R, Bresee JS, Duggan C. Managing acute gastroenteritis among children: oral rehydration, maintenance, and nutritional therapy. *MMWR Recomm Rep*. 2003;52(RR-16):1-16.
- American Academy of Pediatrics. Managing acute gastroenteritis among children: oral rehydration, maintenance, and nutritional therapy. *Pediatrics*. 2004;114(2):507.

^aFor the Evidence to Support the Update for this topic, see <http://www.JAMAevidence.com>.

This page intentionally left blank

EVIDENCE TO SUPPORT THE UPDATE:

25

Hypovolemia, Child

TITLE Development of a Clinical Dehydration Scale for Use in Children Aged Between 1 and 36 Months.

AUTHORS Friedman JN, Goldman RD, Srivastava R, Parkin PC.

CITATION *J Pediatr.* 2004;145(2):201-207.

QUESTION Can a clinical dehydration scale accurately and reliably distinguish between degrees of dehydration and help to assess the response to therapy?

DESIGN A prospective study enrolled a convenience sample of children and assessed dehydration signs before and after rehydration.

SETTING Participants were enrolled through the emergency department of a tertiary-care pediatric hospital.

PATIENTS Children aged 1 to 36 months and presenting to the hospital for treatment of presumed viral gastroenteritis were enrolled. All participants were judged by the attending physician to be dehydrated and to need rehydration therapy (either oral or intravenous). Exclusion criteria were another cause of dehydration, the presence of any chronic disease, recent intravenous fluid therapy, or important serum sodium alterations (<130 or >150 mmol/L).

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

Twelve clinical tests for dehydration were initially identified by a review of the published literature and survey of experts in the field. Reported urine output, general appearance, capillary refill, skin turgor, sunkeness of eyes, mucous membranes, tears, respiratory rate, and heart rate were endorsed frequently enough to be included for further analysis of test characteristics. Signs with the strongest measurement properties on univariate analyses were then combined to form the clinical scale.

The diagnostic standard for initial percentage of dehydration was calculated with the following equation: (rehydration weight – the dehydrated weight) $\times 100$ /rehydration weight.

MAIN OUTCOME MEASURES

The rating scale possessed the strongest measurement properties (Table 25-8 and Table 25-9).

In addition, the reliability of this scale was assessed between examiners. The intraclass correlation coefficient was 0.77, demonstrating a high level of agreement.

Table 25-8 Rating Scale Based on Severity of 4 Clinical Signs

Clinical Signs	0	1	2
General appearance	Normal	Thirsty, restless, lethargic but irritable when touched	Drowsy, limp, cold, comatose ^a
Eye appearance	Normal	Slightly sunken	Very sunken
Mucous membranes	Moist	"Sticky"	Dry
Tear presence	Tears	Decreased	Absent tears

^aChildren who are comatose automatically fall into this category.

Table 25-9 Likelihood Ratio for Rating Scale at a Threshold of ≥ 1 for Dehydration of at Least 3%

Findings	Result	$\geq 3\%$ Dehydration			
		Sensitivity	Specificity ^a	LR+ (95% CI)	LR- (95% CI)
General appearance, sunken eyes, dry tongue, decreased tears	≥ 1 Abnormal	0.85	0.32	1.3 (1.0-1.6)	0.46 (0.19-1.1)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

^aPatricia Parkin, MD, provided the specificity data from the results of her original research (written communication, April 2006).

CONCLUSIONS

LEVEL OF EVIDENCE Level 3.

STRENGTHS This study used established methodology for the development of outcome measures and applied them to a

common clinical concern in the care of children. The scale is easy to use and evaluate in clinical settings.

LIMITATIONS The use of an immediate rehydration weight instead of a true well weight to determine the exact degree of dehydration is the most important limitation of this study. Determining the percentage of dehydration in this manner has not been previously validated.

This meticulously conducted study illustrates how ineffective clinicians are at accurately identifying dehydration. All subjects were thought to be dehydrated by pediatric emergency department specialists, yet 16% of the subjects had no dehydration and 49% had clinically insignificant (<3%) degrees of dehydration when it was retrospectively measured with the diagnostic standard. Unfortunately, the test characteristics demonstrated by their clinical scale do not further assist clinicians with the accurate identification of dehydration.

The authors hoped to establish a clinical dehydration scale whose purpose was to discriminate between degrees of dehy-

dration, though recent practice guidelines recommend grouping children into “none,” “some,” or “severe” dehydration and basing treatment accordingly instead of on estimates of percentile-based groupings.^{1,2} The responsiveness to change of their clinical scale suggests a potentially important clinical use; normalization of the scale may signal that a child is rehydrated and can safely stop therapy. However, this potential needs to be confirmed in future clinical trials.

Reviewed by Michael J. Steiner, MD

REFERENCES FOR THE EVIDENCE

1. King CK, Glass R, Bresee JS, Duggan C. Managing acute gastroenteritis among children: oral rehydration, maintenance, and nutritional therapy. *MMWR Recomm Rep.* 2003;52(RR-16):1-16.
2. World Health Organization. *The Treatment of Diarrhoea: A Manual for Physicians and Other Senior Health Workers*. Geneva, Switzerland: WHO; 2003. http://libdoc.who.int/hq/2003/WHO_FCH_CAH_03.7.pdf. Accessed June 1, 2008.

CHAPTER 26

Does This Patient Have Influenza?

Stephanie A. Call, MD, MSPH

Mark A. Vollenweider, MD, MPH

Carlton A. Hornung, PhD, MPH

David L. Simel, MD, MHS

W. Paul McKinney, MD

CLINICAL SCENARIO

A 45-year-old eighth-grade math teacher visited your office in mid-December 2003, complaining of temperature to 38.6°C (101.5°F), dry cough, sore throat, myalgias, and malaise. Her symptoms began approximately 24 hours earlier, but she continued to teach through the end of the school day. A number of children in her classes were absent because of similar complaints during the past 2 weeks. Her physical examination results revealed readily apparent malaise, temperature of 38.5°C (101°F), mild pharyngeal erythema with no exudates, no adenopathy, and clear lung fields. She took acetaminophen and ibuprofen for fever and muscle aches, with modest relief. Her medical history was notable for hypertension and gastroesophageal reflux disease, for which she took hydrochlorothiazide and lansoprazole, respectively. Aside from 2 normal deliveries more than 10 years previously and an appendectomy during childhood, she had never been hospitalized. As in previous years, she chose not to receive influenza vaccine. She came to you suspecting that she might have the flu and asking whether any medication would help her return to the classroom more quickly.

WHY IS THIS AN IMPORTANT CLINICAL ISSUE?

Ten percent to 20% of US residents contract influenza annually, accounting for an average of 36000 deaths throughout the past decade¹ and 133900 pneumonia and influenza hospitalizations per year from 1979 to 2001.² Given its propensity for antigenic drifts and shifts, influenza has the capability to cause periodic epidemics and global pandemics. A shortfall in production of vaccine because of problems at one manufacturer's facilities (<http://www.hhs.gov/news/press/2004/pres/20041005.html>; accessed March 28, 2008) created the potential for increased morbidity and mortality in the 2004-2005 influenza season. The effect on society during major outbreaks is substantial in terms of both direct medical costs and indirect costs associated with illness, including missed workdays and reduced productivity.³ In 2003, there were concerns about early season reports of influenza-related severe illnesses and deaths in the United States.⁴ The fixed number of doses of vaccine (approximately 83 million) and the increased demand for its use in 2003 led to a redistribution of vaccine to clinicians caring for individuals with the greatest immediate need.⁴ This situation was compounded by a vaccine that may have had reduced effectiveness because of a suboptimal antigenic match. Early in the 2004-2005 season, one of the manufacturers of the trivalent inactivated vaccine did not provide vaccine to the United States; consequently, the available vaccine for the nation was only

about half that projected for the year.⁵ Under these circumstances, early diagnosis and intervention were even more critical.

Two agents, zanamivir and oseltamivir (for either type A or type B strains), are currently recommended and reduce the duration of clinical illness,⁶ but they are expensive and must be instituted within 48 hours of symptom onset for maximal benefit. Consequently, they should be used only when the probability of infection with influenza and the expected benefit are both high.

Influenza-like illness, defined by the Centers for Disease Control and Prevention (CDC) US Influenza Sentinel Providers Surveillance Network as temperature higher than 37.8°C (100°F) plus either cough or sore throat (<http://www.cdc.gov/flu/weekly/>; accessed June 1, 2008) but sometimes defined differently by others, is a syndrome characterized by other nonspecific symptoms that may be observed with a variety of upper respiratory tract infections. The frequency of infections attributable to the various viral agents that cause influenza-like illness varies geographically and from week to week throughout the influenza season. Fortunately, excellent weekly reports are available that help clinicians understand both the incidence of influenza-like illness and the current influenza activity rates applicable to their geographic locations. The CDC produces weekly influenza reports that are available online (<http://www.cdc.gov/flu/weekly/fluactivity.htm>; accessed June 1, 2008). These reports provide a synopsis of epidemiologic information, including laboratory surveillance data, influenza-like illness frequency as reported by US sentinel providers, and regional variability of outbreaks (Figure 26-1). Similar reports are available from individual state health departments, Canada (through Health Canada), the World Health Organization (WHO) International Influenza Program, the WHO Flunet, and the European Influenza Surveillance Scheme (hyperlinks available at <http://www.cdc.gov/flu/weekly/intsurv.htm>; accessed June 1, 2008).

In the 2003-2004 influenza season, the weekly percentage of patient visits for influenza-like illness exceeded the national baseline of 2.5% for 9 consecutive weeks, with a peak of 7.6% in the week ending December 27, 2003.⁴ Thus, during the peak week of the 2003-2004 outbreak, about 1 of every 13 primary care visits in the United States was for an influenza-like illness.

Laboratory surveillance monitoring in the United States showed that most samples in the 2003-2004 influenza season tested negative for influenza. Although not specifically reported, the implication is that these patients often had other viruses such as rhinoviruses, adenoviruses, and parainfluenza. Although many of these are relatively benign and self-limited, others may be serious; for example, early infection during an epidemic of the coronavirus causing severe acute respiratory syndrome (SARS) produced influenza-like illness.⁷ Bacterial agents, including *Legionella* species, *Chlamydia pneumoniae*, *Mycoplasma pneumoniae*, and *Streptococcus pneumoniae*, may also be responsible for influenza-like illnesses.

When faced with a patient with influenza-like illness, a physician must be able to accurately estimate the probability of influenza as opposed to other infections. This probability estimate guides the clinician in further diagnostic testing and treatment. Appropriate and prompt diagnosis and therapy affect not only the individual patient but society as well, in that local outbreaks may be detected and control measures initiated. Influenza is difficult to diagnose because of non-specific symptoms and the host of other diseases that cause similar symptoms. Our objective in this review was to identify clinical factors that may be valuable in distinguishing which patients with influenza-like illness have a higher probability of truly having influenza.

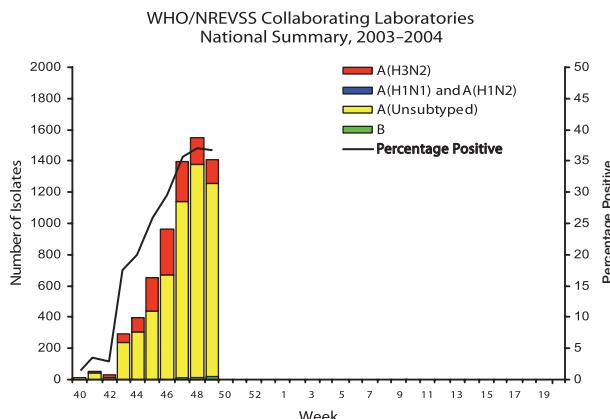
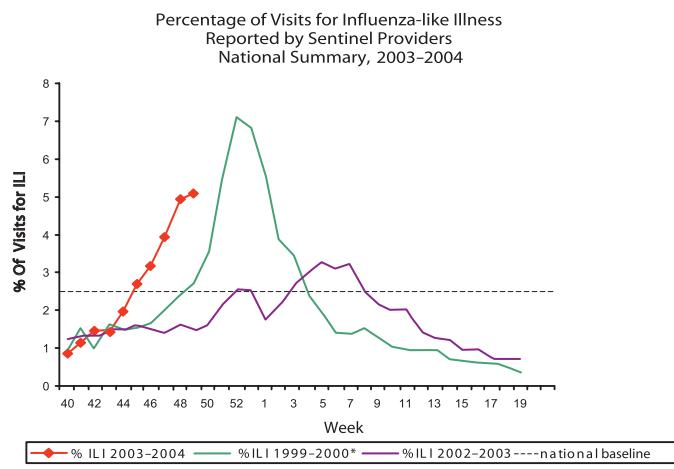
METHODS

Search Strategy and Quality Review

We searched MEDLINE (January 1966 to September 2004) to identify articles pertaining to the diagnosis of influenza according to individual clinical signs and symptoms. We intentionally limited the search to the period before the SARS epidemic to avoid implying that the same operating characteristics could be applied during an outbreak with a highly virulent agent causing similar symptoms. The search strategy used the following Medical Subject Headings: “EXP influenza” or “EXP influenza A virus” or “EXP influenza A virus human” or “EXP influenza B virus.” These terms were then combined with the Medical Subject Headings and text words “EXP sensitivity and specificity” or “EXP medical history taking” or “EXP physical examination” or “EXP reproducibility of results” or “EXP observer variation” or “symptoms.mp” or “clinical signs.mp” or “sensitivity.mp” or “specificity.mp.” We also searched for academic reviews on influenza (“EXP influenza” or “EXP influenza A virus” or “EXP influenza B virus,” limited to human, English-language academic reviews). From this search, we retained only systematic reviews. We reviewed the references and citations to identify other relevant articles. We also reviewed the references in a recent systematic review by Ebell et al.⁸ Unpublished primary data were not sought.

Abstracts of the identified articles were reviewed for relevance. Only articles describing primary studies dealing with the diagnosis of influenza according to clinical signs and symptoms were selected for complete review.

Two of the authors (S.A.C., W.P.M.) independently reviewed the final set of 17 articles for quality.⁹⁻²⁵ Differences in assessment were discussed and resolved by consensus. Studies in the final set were excluded from analysis if they did not meet the following criteria: (1) study design qualifying as prospective cohort, randomized controlled trial, or meta-analysis; (2) inclusion of primary assessment of individual clinical signs and symptoms as predictors of diagnosis; (3) definition of at least 1 of the outcomes as influenza type A or B infection that was proven by (a) culture, (b) 4-fold increase in diagnostic antibody titer, eg, hemagglutination inhibition, complement fixation, or enzyme immunoassay from acute to convalescent serum,

A Laboratory Surveillance**B** Influenza-Like Illness (ILI) Surveillance

* The 1999-2000 season was selected for comparison because it was the most recent A(H3N2) season of moderate severity.

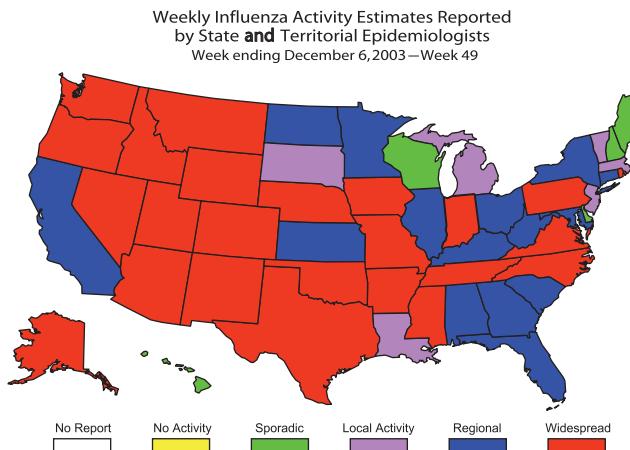
C Weekly Influenza Activity Estimates Reported by State and Territorial Epidemiologists

Figure 26-1 Centers for Disease Control and Prevention Weekly Report: Influenza Summary Update, Week Ending December 6, 2003—Week 49

Source: <http://www.cdc.gov/flu/weekly/weeklyarchives2003-2004/weekly49.htm>; accessed June 1, 2008.

(c) polymerase chain reaction, or (d) immunofluorescent antibody; and (4) study quality graded A or B using the scheme appearing previously in The Rational Clinical Examination series, adapted from Holleman and Simel²⁶ as shown (see Table 1-7 for a summary of Evidence Grades and Levels).

Grade A: Independent blinded comparison of signs or symptoms with criterion standard among a large number of consecutive patients (≥ 300) who might have influenza.
Grade B: Independent blinded comparison of signs or symptoms with criterion standard among a small number of consecutive patients (< 300) who might have influenza.

Grade C1: Independent blinded comparison of signs or symptoms with criterion standard in nonconsecutive patients or nonindependent comparison in patients known to have influenza.

Grade C2: Comparison of signs or symptoms with standard of uncertain validity.

Ten articles met all of the inclusion criteria.^{9,12,14,17-20,23-25}

Because the interpretation of rapid influenza test results is tightly coupled to the interpretation of the clinical examination, we added information to the article about the usefulness of diagnostic testing. This information was obtained through an additional MEDLINE database search (January 1996 to October 2004) for English-language articles pertaining to rapid diagnostic kits for human influenza. This strategy was devised to focus on articles describing the most current and relevant tests available to clinicians and to find citations in which direct comparisons of the most recent tests might be available. The search strategy used the following medical subject headings: "EXP influenza" and "EXP sensitivity and specificity" and "EXP reagent kits, diagnostic." Data from manufacturers were also sought to establish the products' range of sensitivity and specificity. Unpublished primary data were not sought. Abstracts of identified articles were reviewed for relevance.

Statistical Methods

We used data from the identified articles to calculate the sensitivity, specificity, positive likelihood ratio (LR), and negative LR, as well as a summary LR and the diagnostic odds ratio (OR) for individual medical history and physical examination findings. The positive LR is a measure of how strongly a positive result increases the odds of disease; the negative LR is a measure of how well a negative result decreases the odds of disease. An LR greater than 1.0 increases the likelihood of disease; an LR less than 1.0 decreases the likelihood; an LR close to 1.0 does not change the likelihood. FastPro (Academic Press, Boston, Massachusetts) was used for all analyses; $P < .05$ was used to determine statistical significance.

The Diagnostic OR

The diagnostic OR is a single indicator of diagnostic test performance, reflecting its accuracy.²⁷ The diagnostic OR can also be viewed as presenting the odds (likelihood) of the symptom or finding among individuals with disease (ie, the positive LR) compared with the odds of the symptom or finding among those not having the disease (ie, the negative LR). The diagnostic OR should always be assessed in comparison with the paired sensitivity and specificity because the same diagnostic OR can be associated with different pairs. The value of the diagnostic OR ranges from 0 to infinity, with higher values indicating better test performance. Values less than 1 indicate more negative test results among individuals with disease. The diagnostic OR can also be used to develop summary estimates in meta-analyses.

We tested the LRs for heterogeneity between studies using the Mantel-Haenszel Q-statistic.²⁸ We used conservative random-effects models to describe the summary estimates and confidence intervals (CIs), making it easier to discern the relative usefulness of symptoms and signs.^{29,30}

RESULTS

The search strategy identified 915 articles (bibliography available on request). We found only 10 studies that met all the inclusion criteria.^{9,12,14,17-20,23-25} Most of the excluded articles were not primary studies. We were unable to obtain primary data for 3 of the 10 studies,^{9,18,19} and data from 1 study were included in 2 articles; thus, the final data (Table 26-1) are based on 6 studies and included 7105 patients.^{12,14,20,23,25,26} We identified a recent systematic review that included several studies for which we were unable to obtain the primary data.⁸ Thus, not all the references in this systematic review met our inclusion criteria. One additional study included in this review, but not identified in our literature search, did meet our inclusion criteria.²⁵

The second search strategy identified 13 articles dealing with rapid diagnostic tests for influenza (bibliography available on request). Only 6 original articles³¹⁻³⁶ describing the comparison of a commercially available rapid diagnostic test for influenza vs viral culture as the criterion standard were selected for complete review. Of these, only 1 article³⁵ presented direct comparison of results among 4 test kits studied; the data from this article were evaluated in detail.

Precision of Signs and Symptoms

None of the studies assessed the precision of signs or symptoms of influenza. Measurements of objective clinical signs such as temperature are assumed to have high precision.

Accuracy of Signs and Symptoms

The studies presented used varying definitions for fever, ranging from 37.8°C to 38.5°C (Table 26-1). We defined fever as *present* or *absent* according to the individual article's definition. *Feverishness* was reported by the patient and could have been based on either a temperature taken at home or a subjective sense of having an elevated temperature. The sensitivity, specificity, positive LR, negative LR, and diagnostic OR for clinical variables evaluated in at least 2 of the 5 studies are reported in Tables 26-2 and 26-3. Summary estimates are also presented. Eleven of the 13 clinical factors had heterogeneous diagnostic ORs (all with $P < .05$). The patient's sense of feverishness and vaccination history provided homogeneous results across studies. Despite the heterogeneity, the studies we reviewed seem representative of the universe of patients with influenza, and most of the differences in estimates created by the statistical heterogeneity were small. The heterogeneity, expressed in the CIs, never moved a finding from useless (LR approaching 1) to obviously useful (LR so different from 1 that it would make influenza extremely likely or extremely unlikely). Therefore, we present the summary LRs and diagnostic ORs as an efficient way of conveying the relative diagnostic effect of the symptoms and signs.

Table 26-1 Studies of the Diagnostic Performance of Clinical Findings in Diagnosing Influenza

Source, y	Study Period	Location	No. of Patients	Age Range, y	Design	Selection Criteria	Grade of Evidence ^a	Diagnostic Test	Prevalence of Influenza, %
Nicholson et al, ²⁵ 1997	Winters of 1992-1993 and 1993-1994	Leicester-shire, England	533	60-90	Prospective cohort	Weekly telephone surveillance for symptoms of upper respiratory tract infection; home visit as soon as possible thereafter if symptoms noted	A	4-Fold increase in hemagglutination inhibition titer	8
Govaert et al, ¹⁴ 1998	Influenza season, 1991-1992	The Netherlands	1838	≥60	Randomized controlled trial (of influenza vaccine)	Tested all persons in the study. Persons were originally selected from general practice offices, not "high-risk" groups	B	4-Fold increase in titer (influenza A)	7
Carrat et al, ¹² 1999	Influenza epidemic, 1995-1996	France	610	Included all ages ≥ 1 y	Prospective cohort	Sudden onset of ≥ 1 of the following: influenza-like illness, upper or lower respiratory tract infectious syndrome, or temperature of > 38°C (100.4°F) without any symptoms or signs of other infectious syndromes	A	ELISA, immunofluorescence (influenza A)	28
Monto et al, ²⁰ 2000	Fall and winters, 1994-1998	231 Study centers in North America, Europe, southern hemisphere	3744	≥12	Retrospective, pooled analysis of clinical trials	Fever or ≥ 2 symptoms (headache, myalgias, cough, sore throat)	B	Positive culture result for influenza A or B or 4-fold increase in titer or PCR or immunofluorescence	66
Hulson et al, ¹⁷ 2001	3 Consecutive influenza outbreaks, 1999-2000	Oklahoma	358	10 mo-73 y	Prospective cohort	Any of: fever (temperature > 38°C [100.4°F]), cough, sore throat, headache, myalgia	A	Positive culture result for influenza A or B	67
van Elden et al, ²³ 2001	Influenza season, 1997-1998	The Netherlands	81	Included all ages	Prospective cohort	Fever (temperature > 38°C [100.4°F]) plus ≥ 1 constitutional symptom (malaise, headache, myalgia, chills) plus ≥ 1 respiratory symptom (coryza, sneezing, cough, sore throat, hoarseness)	B	PCR (influenza A)	5

Abbreviations: ELISA, enzyme-linked immunosorbent assay; PCR, polymerase chain reaction.

^aSee Table 1-7 for a summary of Evidence Grades and Levels.

No single clinical finding consistently had a positive LR high enough to clinically rule in influenza nor did any single finding have a negative LR low enough to clinically rule out influenza (Tables 26-2 and 26-3). However, several patterns do emerge when the data are evaluated from the multiple studies. Among studies that enrolled patients without regard to age, no single finding had a summary LR greater than 2. For decreasing the likelihood of influenza, the absence of fever (LR, 0.40; 95% CI, 0.25-0.66), cough (LR, 0.42; 95% CI, 0.31-0.57), or nasal congestion (LR, 0.49; 95% CI, 0.42-0.59) was the only finding with an LR less than 0.5. Feverishness, myalgia, malaise, sore throat, and sneezing each had a positive and negative LR that was indistinguishable from 1.0 and therefore of no diagnostic value for the patients in studies that evaluated the entire age spectrum. Among the studies of patients limited to those aged 60 years or older, the strongest univariate indicators of influenza were fever (LR, 3.8; 95% CI, 2.8-5.0), malaise (LR, 2.6; 95% CI, 2.2-3.1), and chills (LR, 2.6; 95% CI, 2.0-3.2). Among

older patients exclusively, the presence of sneezing reduced the likelihood of influenza (LR, 0.47; 95% CI, 0.24-0.92).

Two studies, by Govaert et al¹⁴ and Monto et al,²⁰ assessed the diagnostic usefulness of fever with cough in persons aged 60 years or older and in the unrestricted age group (Table 26-3). The LRs when both fever and cough were present were 5.0 and 1.9, respectively. The addition of a third variable, acute onset of symptoms, added minimally to the discriminatory accuracy in either study.

The calculation of diagnostic ORs for the individual variables in each study allows us to compare the diagnostic performance of the different variables and combinations of variables using a single measure (Tables 26-2 and 26-3). The 3 studies with the lowest frequency of influenza tended to have the best overall accuracy as expressed by the diagnostic OR. In comparison with the calculated diagnostic ORs for other symptoms, fever (summary diagnostic OR, 4.5; 95% CI, 1.8-11) and cough (summary diagnostic OR, 2.8; 95% CI, 2.1-3.7) are the most useful single

Table 26-2 Test Characteristics of Clinical Findings, by Study

Source, y	Sensitivity	Specificity	LR+ (95% CI) ^a	LR- (95% CI) ^a	DOR (95% CI) ^a
Fever					
No age restriction					
Carrat et al, ¹² 1999	0.84	0.73	3.1 (2.6-3.7)	0.21 (0.15-0.31)	14 (8.8-23)
Monto et al, ²⁰ 2000	0.68	0.60	1.7 (1.6-1.8)	0.53 (0.49-0.57)	3.2 (2.8-3.7)
Hulson et al, ¹⁷ 2001	0.86	0.25	1.1 (1.0-1.3)	0.59 (0.35-0.87)	1.9 (1.0-3.4)
Summary			1.8 (1.1-2.9)	0.40 (0.25-0.66)	4.5 (1.8-11)
Only patients ≥ 60 y					
Govaert et al, ¹⁴ 1998	0.34	0.91	3.8 (2.8-5.0)	0.72 (0.64-0.82)	5.2 (3.4-7.9)
Feverishness					
No age restriction					
Monto et al, ²⁰ 2000					1.1 (0.89-1.4)
van Elden et al, ²³ 2001	0.88	0.15	1.0 (0.86-1.2)	0.70 (0.27-2.5)	1.3 (0.35-4.6)
Summary					1.1 (0.88-1.4)
Only patients ≥ 60 y					
Nicholson et al, ²⁵ 1997	0.47	0.78	2.1 (1.2-3.7)	0.68 (0.45-1.0)	3.1 (1.2-8.1)
Cough					
No age restriction					
Carrat et al, ¹² 1999	0.84	0.29	1.2 (1.1-1.3)	0.58 (0.39-0.85)	2.0 (1.3-3.2)
Monto et al, ²⁰ 2000	0.93	0.20	1.2 (1.1-1.2)	0.35 (0.29-0.42)	3.3 (2.7-4.1)
Hulson et al, ¹⁷ 2001	0.96	0.07	1.0 (0.95-1.1)	0.61 (0.25-1.5)	1.9 (0.71-5.0)
van Elden et al, ²³ 2001	0.98	0.23	1.3 (1.1-1.5)	0.11 (0.01-0.82)	12 (1.4-97)
Summary			1.1 (1.1-1.2)	0.42 (0.31-0.57)	2.8 (2.1-3.7)^b
Only patients ≥ 60 y					
Nicholson et al, ²⁵ 1997	0.53	0.56	1.2 (0.75-1.9)	0.85 (0.52-1.4)	1.4 (0.5-3.7)
Govaert et al, ¹⁴ 1998	0.66	0.77	2.9 (2.5-3.4)	0.44 (0.34-0.56)	6.7 (4.5-10)
Summary			2.0 (1.1-3.5)	0.57 (0.37-0.87)	3.4 (1.2-9.7)
Myalgia					
No age restriction					
Monto et al, ²⁰ 2000	0.94	0.06	1.0 (0.98-1.0)	1.0 (0.76-1.3)	0.99 (0.75-1.3)
Hulson et al, ¹⁷ 2001	0.64	0.21	0.81 (0.71-0.93)	1.7 (1.2-2.5)	0.50 (0.29-0.83)
van Elden et al, ²³ 2001	0.60	0.38	0.97 (0.68-1.4)	1.0 (0.60-1.8)	0.94 (0.38-2.3)
Summary			0.93 (0.83-1.0)	1.2 (0.90-1.6)	0.79 (0.54-1.1)
Only patients ≥ 60 y					
Nicholson et al, ²⁵ 1997	0.47	0.83	2.7 (1.5-5.0)	0.64 (0.41-0.98)	4.3 (1.6-12)
Govaert et al, ¹⁴ 1998	0.45	0.81	2.4 (1.9-3.0)	0.68 (0.58-0.80)	3.4 (2.3-5.0)
Summary			2.4 (1.9-2.9)	0.68 (0.58-0.79)	3.5 (2.4-5.0)
Malaise					
No age restriction					
van Elden et al, ²³ 2001	0.73	0.26	0.98 (0.75-1.3)	1.1 (0.51-2.2)	0.91 (0.34-2.5)
Only patients ≥ 60 y					
Govaert et al, ¹⁴ 1998	0.57	0.78	2.6 (2.2-3.1)	0.55 (0.44-0.67)	4.9 (3.3-7.1)
Headache					
No age restriction					
Carrat et al, ¹² 1999	0.84	0.26	1.1 (1.0-1.2)	0.62 (0.42-0.91)	1.9 (1.2-3.0)
Monto et al, ²⁰ 2000	0.91	0.11	1.0 (0.99-1.0)	0.81 (0.66-0.99)	1.3 (1.0-1.6)
Hulson et al, ¹⁷ 2001	0.88	0.16	1.1 (0.95-1.1)	0.75 (0.43-1.3)	1.4 (0.76-2.7)
van Elden et al, ²³ 2001	0.70	0.43	1.2 (0.87-1.7)	0.70 (0.38-1.3)	1.8 (0.76-4.5)
Summary			1.0 (1.0-1.1)	0.75 (0.63-0.89)	1.4 (1.2-1.8)^b
Only patients ≥ 60 y					
Nicholson et al, ²⁵ 1997	0.68	0.57	1.6 (1.1-2.3)	0.56 (0.28-1.1)	2.8 (1.0-7.8)
Govaert et al, ¹⁴ 1998	0.44	0.79	2.1 (1.7-2.6)	0.71 (0.60-0.83)	3.0 (2.0-4.4)
Summary			1.9 (1.6-2.3)	0.70 (0.60-0.82)	3.0 (2.1-4.3)^b

Abbreviations: CI, confidence interval; DOR, diagnostic odds ratio; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

^aLR+ is the likelihood ratio when the finding is present; LR- is the likelihood ratio when the finding is absent; DOR is an indicator of the test's overall accuracy.^bHomogeneous DOR ($P > .05$). When the DOR was heterogeneous, we assessed for homogeneity separately for the positive and negative LRs.

Table 26-3 Test Characteristics of Clinical Findings, by Study

Source, y	Sensitivity	Specificity	LR+ (95% CI) ^a	LR- (95% CI) ^a	DOR (95% CI) ^a
Sore Throat					
No age restriction					
Monto et al, ²⁰ 2000	0.84	0.16	1.0 (0.97-1.0)	1.0 (0.85-1.2)	1.0 (0.8-1.2)
Hulson et al, ¹⁷ 2001	0.75	0.28	1.0 (0.91-1.2)	0.89 (0.62-1.3)	1.2 (0.72-2.0)
van Elden et al, ²³ 2001	0.80	0.33	1.2 (0.91-1.6)	0.61 (0.28-1.3)	1.9 (0.69-5.3)
Summary			1.0 (0.98-1.0)	0.96 (0.83-1.1)	1.1 (0.87-1.3) ^b
Only patients ≥ 60 y					
Nicholson et al, ²⁵ 1997	0.58	0.36	0.91 (0.61-1.4)	1.2 (0.66-2.1)	0.8 (0.3-2.1)
Govaert et al, ¹⁴ 1998	0.40	0.81	2.1 (1.7-2.7)	0.74 (0.64-0.85)	2.9 (2.0-4.3)
Summary			1.4 (0.81-2.5)	0.77 (0.66-0.89)	1.8 (0.81-4.0)
Sneezing					
No age restriction					
Carrat et al, ¹² 1999	0.50	0.59	1.2 (1.0-1.5)	0.85 (0.71-1.0)	1.4 (1.0-2.1)
van Elden et al, ²³ 2001	0.33	0.69	1.1 (0.55-2.0)	0.97 (0.71-1.3)	1.1 (0.42-2.8)
Summary			1.2 (1.0-1.5)	0.87 (0.75-1.0)	1.3 (0.95-1.9) ^b
Only patients ≥ 60 y					
Nicholson et al, ²⁵ 1997	0.32	0.33	0.47 (0.24-0.92)	2.1 (1.4-3.1)	0.2 (0.1-0.6)
Nasal Congestion					
No age restriction					
Monto et al, ²⁰ 2000	0.91	0.19	1.1 (1.1-1.2)	0.47 (0.40-0.56)	2.4 (2.0-2.9)
van Elden et al, ²³ 2001	0.68	0.41	1.1 (0.81-1.6)	0.79 (0.44-1.4)	1.4 (0.58-3.6)
Summary			1.1 (1.1-1.2)	0.49 (0.42-0.59)	2.3 (1.9-2.8) ^b
Only patients ≥ 60 y					
Nicholson et al, ²⁵ 1997	0.47	0.50	0.95 (0.57-1.6)	1.0 (0.67-1.7)	0.9 (0.3-2.4)
Chills					
No age restriction					
Carrat et al, ¹² 1999	0.83	0.25	1.1 (1.0-1.2)	0.68 (0.46-0.99)	1.6 (1.0-3.0)
Only patients ≥ 60 y					
Govaert et al, ¹⁴ 1998	0.46	0.82	2.6 (2.0-3.2)	0.66 (0.55-0.77)	3.9 (2.7-5.7)
Vaccine History					
No age restriction					
Hulson et al, ¹⁷ 2001	0.12	0.83	0.71 (0.41-1.2)	1.1 (0.96-1.2)	0.69 (0.37-1.3)
van Elden et al, ²³ 2001	0.02	0.82	0.11 (0.01-1.1)	1.2 (0.02-1.4)	0.12 (0.01-1.0)
Summary			0.63 (0.37-1.1)	1.1 (1.0-1.2)	0.60 (0.33-1.1) ^b
Fever and Cough					
No age restriction					
Monto et al, ²⁰ 2000	0.64	0.67	1.9 (1.8-2.1)	0.54 (0.50-0.57)	3.6 (3.1-4.2)
Only patients ≥ 60 y					
Govaert et al, ¹⁴ 1998	0.30	0.94	5.0 (3.5-6.9)	0.75 (0.66-0.84)	6.6 (4.2-10)
Fever and Cough and Acute Onset					
No age restriction					
Monto et al, ²⁰ 2000	0.63	0.68	2.0 (1.8-2.1)	0.54 (0.51-0.58)	3.6 (3.1-4.1)
Only patients ≥ 60 y					
Govaert et al, ¹⁴ 1998	0.27	0.95	5.4 (3.8-7.7)	0.77 (0.68-0.85)	7.1 (4.5-11)

Abbreviations: CI, confidence interval; DOR, diagnostic odds ratio; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

^aLR+ is the likelihood ratio when the finding is present; LR- is the likelihood ratio when the finding is absent; DOR is an indicator of the test's overall accuracy.

^bHomogeneous DOR ($P > .05$). When the DOR was heterogeneous, we assessed for homogeneity separately for the positive and negative LRs.

findings for distinguishing patients with influenza from those without the illness among the unrestricted age group. The combination of fever and cough, with or without acute onset, had an intermediate diagnostic OR value. The diagnostic ORs were somewhat higher for all of these characteristics, particularly the combined symptoms, among persons aged 60 years or older; malaise (diagnostic OR, 4.9; 95% CI, 3.3-7.1) also performed well in this group.

Fever, headache, myalgias, and cough are the classic symptoms associated with influenza. Unfortunately, these symptoms are frequently observed in patients presenting with other infections during influenza season, making the clinical diagnosis of influenza problematic to the primary care physician. These data suggest that the strongest predictor of influenza was the acute onset of both fever and cough in patients aged 60 years or older.

We included only those studies in which a laboratory confirmation of the influenza virus was performed for all patients, thus eliminating verification bias. However, not all studies used the same criterion standard diagnostic test; one study used culture data only, without supplementation by titer increase or polymerase chain reaction.¹⁷ This may have caused false-negative results and a decreased estimate of prevalence of disease. In addition, several of the studies assessed the type of influenza (A vs B), whereas others did not. One study found that different clinical presentations were associated with the influenza type,¹² but another study showed no difference.²⁰ In the 2 studies that presented data on both influenza types A and B, the proportion of patients diagnosed with influenza B was small (<10%).^{12,20} Data presented here reflect all diagnoses of influenza, regardless of type or subtype; we do not know whether the clinical presentation of disease varies according to type or subtype.

The patient populations in the 6 studies were very different but represented a broad spectrum of patients with influenza-like illnesses. Two of the study populations were derived from randomized controlled trials of treatment or vaccine,^{14,20} 3 were prospective cohorts of patients presenting to general practitioners,^{12,17,23} and 1 was a population-based cohort surveyed for symptoms weekly by telephone.²⁵ The studies were also from several countries: 2 from The Netherlands,^{14,23} 1 from France,¹² 1 from the United States,¹⁷ and 1 from England²⁵; in addition, one was a multinational study²⁰ including patients from North America, Europe, and the southern hemisphere. This variability in patient population may have led to less precision in the assessment of symptoms because of cultural and language differences. The different study populations may also have had different clinical characteristics owing to the pool from which they were drawn. The studies from Europe were more likely to include patients from home. It is conceivable that these patients were more or less ill, had more or fewer symptoms, and had a different prevalence of influenza compared with the populations from the United States or other countries.

Including the randomized controlled trials may lead to spectrum bias because patients who enroll in randomized controlled trials assessing either treatment or prevention of influenza may not represent the population of patients

presenting to a primary care office. Spectrum bias may be a particular issue in the study by Govaert et al,¹⁴ in that signs and symptoms were assessed in all the patients enrolling in the vaccine trial, even those without complaints of illness.¹⁴ This not only leads to spectrum bias but also is consistent with the 6.6% prevalence in this study, which is lower than the prevalences in the other studies (range, 8%-67%).

Other differences in the study populations include the age range within each study. This may be important because Cox and Subbarao³⁷ have observed that influenza presents differently among various age groups. Although most of the patients studied in these reports were adults, several of the studies did include children. Govaert et al¹⁴ and Nicholson et al²⁵ evaluated only individuals aged 60 years or older. The positive LRs for several of the signs and symptoms evaluated in these studies are higher than those in the other studies. One possible explanation for this is that the clinical findings are more diagnostic of influenza in the elderly population or in a population with a lower prevalence of disease.

Although all of the studies recruited patients only during influenza season, some were specifically undertaken during epidemics. Thus, the prevalence of disease varies considerably in the published reports of the clinical findings. It is possible that clinical characteristics of the disease change between seasons according to the strain of influenza. All of the studies were performed before the SARS epidemic.

Monto et al²⁰ suggested that the positive predictive value of clinical signs and symptoms increased with increasing duration from illness onset. The 6 studies presented in this article had various durations of symptoms. Data were not available from each of the studies to assess whether the other studies supported the results of Monto et al.²⁰

Approach to Influenza Diagnosis

The 2003 outbreak of influenza brought the diagnostic dilemmas regarding influenza to the forefront. The reduced availability of vaccine for 2004-2005 created the potential for increased incidence of disease. When faced with a person with influenza-like illness, clinicians struggle with the decision of whether to test or to empirically treat.

There are several laboratory-based procedures available for diagnosing influenza. Viral culture is the criterion standard for laboratory diagnosis, but it may take several days to see cytopathic effects or for virus to be detected by hemadsorption or hemagglutination. Rapid methods may shorten the time to identification but at some cost in sensitivity. Fluorescent antibody staining or other immunoassays are used to confirm and to type influenza virus in culture and are frequently used directly on respiratory specimens as part of a respiratory virus battery. Results from direct immunoassays may be available within hours. Molecular methods such as reverse-transcriptase polymerase chain reaction and hybridization-based arrays are likely to replace culture as the criterion standard because of their superior sensitivity and rapid turnaround time. However, the availability of technology is limited. For diagnostic dilemmas, research studies, and

epidemiologic purposes, influenza infection can also be detected by a 4-fold or greater increase in a variety of diagnostic antibody titers (eg, hemagglutination inhibition, complement fixation, or enzyme immunoassay) between specimens collected at least 10 days apart. Although these laboratory-based methods are highly sensitive and specific, clinicians are increasingly reliant on point-of-care rapid diagnostic tests, which are easier to handle, are less costly, and provide test results in fewer than 30 minutes.

A summary of the rapid diagnostic tests for influenza is provided by the CDC (<http://www.cdc.gov/flu/professionals/diagnosis/>; accessed June 1, 2008). These include Directigen Flu A and Directigen Flu A + B (Becton-Dickinson, Franklin Lakes, New Jersey), FLU OIA and FLU OIA A/B (Thermo Electron Corp, Waltham, Massachusetts), XPECT Flu A/B (Remel, Lenexa, Kansas), NOW Flu A Test and NOW Flu B Test (Binax Inc, Portland, Maine), QuickVue Influenza Test and Quick Vue Influenza A + B Test (Quidel Corp, San Diego, California), SAS Influenza A Test and SAS Influenza B Test (SA Scientific Ltd, San Antonio, Texas), and ZstatFlu (ZymeTx Inc, Oklahoma City, Oklahoma). The tests require specimens of throat swabs, nasopharyngeal swabs, nasal washes, or nasal aspirates. The sensitivity and specificity of these tests have been reported in manufacturers' reports to be between 40% and 100% and between 52% and 100%, respectively. Given the differences between older and younger persons in presenting symptoms of influenza,³⁷ the operating characteristics of these tests could differ among various age groups; however, we found no data confirming this. The QuickVue and ZstatFlu tests have waivers from the Clinical Laboratory Improvement Amendments and can be used in

any office setting. The Quick Vue A + B Test is the only amendment-waived test that distinguishes between influenza A and B.

Multiple studies have compared individual test kits vs the reference standard of viral culture (Table 26-4).³¹⁻³⁶ In 2002, Rodriguez et al³⁵ published a study that directly compared 4 of the most widely used rapid diagnostic test kits in children with influenza-like illness. During the 1999-2000 epidemic, the authors had patients provide specimens for viral culture and direct fluorescent antigen, as well as for testing with Directigen Flu A, FLU OIA, QuickVue Influenza Test, and ZstatFlu A/B. Influenza A was found in 49% of the patients; 17% of the cases were detected by viral culture only. Sensitivity and specificity of the 4 tests ranged from 72% to 95% and from 76% to 84%, respectively. For diagnosing influenza, these tests all had similar LRs ($P = .69$) when the results were positive, with a summary LR of 4.7 (95% CI, 3.6-6.2). The ZstatFlu test has a lower sensitivity than the other tests ($P < .001$); however, the remaining tests perform similarly ($P > .99$) and exceedingly well for ruling out influenza when the test result is negative, with a summary LR of 0.06 (95% CI, 0.03-0.12).³⁵

Two recent studies examined the cost-effectiveness of several influenza management strategies in adults, including several strategies in which rapid influenza diagnostic tests were used.^{38,39} The estimates used for the sensitivity of the rapid tests ranged from 59% to 81%; the estimates used for specificity ranged from 70% to 99%. The prior probability estimate of influenza was 35% in the analysis by Rothberg et al³⁸ and 60% in that by Smith and Roberts.³⁹ In both analyses, testing strategies were less effective than empirical treatment because of the low sensitivity of the tests. These analyses were sensitive to the

Table 26-4 Studies of the Performance of Rapid Diagnostic Tests for Influenza

Source, y	Study Period	Location	No. of Patients/ Specimens	Age Range	Design	Selection Criteria	Diagnostic Test
Marcante et al, ³² 1996	December 1994–February 1995	Padova, Italy	41	Children and adults ^a	Prospective cohort	Pediatric/adult patients with symptoms of influenza-like illness	Directigen Flu A
Noyola et al, ³³ 2000	December 19, 1997–April 13, 1998	Houston, TX	196	Children ^a	Prospective cohort	Children with respiratory illness	Zstat Flu A/B
Quach et al, ³⁴ 2002	February and March 2001	Montreal, Quebec	300	Children ^a	Prospective cohort	Children with influenza-like symptoms presenting to Children's Hospital, Montreal	QuickVue
Rodriguez et al, ³⁵ 2002	December 19, 1999–January 13, 2000	Virginia	152	Ages 3 y to adult ^a	Prospective cohort	Symptomatic patients seen in outpatient private practice	Directigen Flu A, Zstat Flu A/B, QuickVue Influenza Test A/B, FLU OIA A/B
Bellei et al, ³¹ 2003	May–October 2000	São Paulo, Brazil	33	18–56 y	Prospective/retrospective cohort	Adult volunteers, 24-h onset of symptoms of influenza-like illness with influenza A confirmation	QuickVue
Cazacu et al, ³⁶ 2004	January–April 2003	Houston, TX; Ft Myers, FL; Syracuse, NY	400	Children and adults ^a	Prospective cohort	Children and adults with respiratory or influenza-like symptoms	Xpect Flu A/B

^aSpecific ages not stated.

probability of influenza infection; the cost-effectiveness of empirical treatment improved relative to the testing strategy as the probability increased. In fact, in the study by Rothberg et al,³⁸ empirical treatment with a neuraminidase inhibitor in unvaccinated patients was more cost-effective at any probability of influenza greater than 14%. Testing was preferred only between a probability of 5% and 14% (in unvaccinated patients). The study by Smith and Roberts³⁹ yielded similar results, favoring rapid testing only at a lower prevalence of influenza. These studies highlight the importance of the physician's estimate of the likelihood of influenza.

The decision analytic model used by Rothberg et al³⁸ was sensitive to vaccination status. In a recent systematic review of the literature, the estimated reduction in serologically confirmed cases of influenza A by the live attenuated aerosol vaccines was 48%; the reduction with the use of inactivated parenteral vaccines was 68%.⁴⁰ Vaccine efficacy and effectiveness may be affected by epidemiologic characteristics such as age and institutionalization. At least 1 study showed a vaccine efficacy of 58% in older patients who were not institutionalized.⁴¹

From these analyses, if one is able to estimate the probability of influenza to be greater than 25% to 30%, rapid diagnostic testing does not add to the overall cost-effectiveness of treatment. Thus, clinicians must develop a pretest probability based on clinical signs and symptoms, vaccination history, and epidemiologic risk factors. During influenza season, the CDC publishes weekly online updates that contain information about the prevalence of visits for influenza-like illness, along with data about influenza outbreaks (<http://www.cdc.gov/flu/weekly/fluactivity.htm>; accessed March 28, 2008). The same information is generally available for each state through its own surveillance reporting systems. It is important that physicians understand the information available in the reports. The percentage of visits to sentinel providers for influenza-like illness for the week ending December 6, 2003 (week 49), was high (5.1%), and there was regional variation (Figure 26-1B and C). Among laboratory respiratory specimens submitted as part of the CDC surveillance system, 37% tested positive for influenza during week 49 (Figure 26-1A). At the beginning of the 2003-2004 influenza season, for the week ending October 4, 2003 (week 40), the percentage of office visits for influenza-like illness was only 0.9% (Figure 26-1B); only 1.4% of laboratory respiratory specimens tested positive for influenza during the same week.

Unfortunately, there is no linkage between the surveillance systems for monitoring influenza-like illness and laboratory results. The CDC surveillance systems are careful to note that the system is designed to report where, when, and what influenza viruses are circulating, but the data cannot be used by the clinician to determine the probability that an individual patient with an influenza-like illness actually has influenza. Although the likelihood of influenza may vary, along with the frequency of influenza-like illness, no data exist for clinically determining whether the threshold levels of the decision analytic model have been exceeded. Depending on the acuity of illness, vaccination status, and presence of comorbid conditions, some physicians might choose to treat empirically with medication, whereas some might choose testing.

CLINICAL SCENARIO—RESOLUTION

The patient came to the office during the usual influenza season with classic influenza-like symptoms. She had been ill for 24 hours, was not vaccinated, and was exposed to many children with influenza-like illnesses. A suspicion of influenza forces the decision of whether to treat her symptomatically, treat her with an antiviral agent, or test for influenza with a rapid test. Because she was fewer than 48 hours into the illness, treating her could allow her to return to work more quickly if she does indeed have influenza. The data most pertinent to this patient were released by the CDC on December 11, 2003 (<http://www.cdc.gov/flu/weekly/weeklyarchives2003-2004/weekly49.htm>; accessed March 28, 2008). The CDC data indicated regional outbreaks of influenza in her area of the country, with 5.1% of primary care visits for influenza-like illnesses (Figure 26-1B). In week 49, 37% of specimens submitted had laboratory confirmation of influenza (Figure 26-1A).

Once clinicians have used the symptoms in Tables 26-2 and 26-3 to establish that a patient has an influenza-like illness, they should use epidemiologic data to determine whether influenza virus is circulating. Clinicians must rely on their clinical judgment in deriving a pretest probability for influenza. If a rapid influenza test is obtained, strict adherence to the manufacturer's protocol is required for accurate interpretation. If the result is positive, the odds of disease increase almost 5-fold (summary LR, 4.7). A negative rapid influenza test result (summary LR, 0.06) decreases the probability of disease and could effectively rule out influenza if the prior probability is low. Astute clinicians will recognize that the decision to use rapid diagnostic testing can vary throughout the influenza season, depending on the age of their patient, the setting, and the prevalence of disease in their community.

CLINICAL BOTTOM LINE

Influenza presents with a constellation of symptoms, including cough, fever, malaise, myalgias, and headache. We reviewed the literature regarding signs and symptoms and their diagnostic accuracy for influenza. Unfortunately, no specific symptom or combination of symptoms is diagnostic of this common infection. Despite the variability in participant nationality, language, culture, and age, as well as in clinical setting and influenza type/subtype in the studies reviewed, the data indicate that, although not perfect, the combination of fever and cough during influenza season suggests a significantly increased likelihood of influenza among elderly individuals.

The usefulness of these signs and symptoms follows from their ability to identify a group with influenza-like illness. However, the prevalence of disease among this population varies from week to week and from year to year throughout the influenza season. Clinicians must pay attention to the surveil-

lance data to understand where, when, and what influenza viruses are circulating. As an example, the peak weeks during the 2002-2003 influenza season for influenza-like illnesses occurred later in the season and at lower rates than in 2003-2004. The role of rapid influenza tests has not been fully established, although it seems likely that clinicians will have many options for testing during future influenza seasons. In a randomized trial of the usefulness of rapid influenza tests in a pediatric emergency department, physicians provided with rapid test results ordered fewer laboratory tests and chest radiographs, prescribed fewer antibiotics but more antiviral agents, kept patients in the emergency department for shorter periods, and generated lower patient charges.⁴²

Once the clinical criteria are used to establish the presence of an influenza-like illness, there is little information other than epidemiologic data that is useful for guiding diagnostic and therapeutic decision making. During the current era of rapidly evolving infections with pathogens unfamiliar to most physicians, we do not know how well the symptoms, signs, and rapid diagnostic tests would perform if these newer infections were to become epidemic. Clinicians in the United States must pay particular attention to the weekly CDC and state reports regarding regional influenza patterns during influenza seasons. International clinicians should use data from the WHO International Influenza Program, the WHO Flunet, Health Canada, or the European Influenza Surveillance Scheme. The hyperlinks for all these sites are available at the CDC Web site (<http://www.cdc.gov/flu/weekly/intsurv.htm>; accessed June 1, 2008). It would be useful for clinicians if a formal linkage could be established between clinical and laboratory surveillance strategies, such as the collection of influenza virus cultures from a random sample of persons presenting with influenza-like illnesses, to allow more precise estimation of an individual's likelihood of disease. In the absence of such a system, physicians may consider point-of-care testing among patients in their individual practices to gain an estimate of the prevalence of influenza among their patients presenting with influenza-like illnesses.

Author Affiliations at the Time of the Original Publication

Department of Medicine, University of Louisville, Louisville, Kentucky (Drs Call, Vollenweider, Hornung, and McKinney); Louisville Veterans Affairs Medical Center, Louisville, Kentucky (Drs Call and McKinney); School of Public Health and Information Sciences, University of Louisville, Louisville, Kentucky (Drs Hornung and McKinney); and Durham Veterans Affairs Medical Center and Duke University Medical Center, Durham, North Carolina (Dr Simel). Dr Call is currently affiliated with the Department of Internal Medicine, Medical College of Virginia, Richmond.

Acknowledgments

The authors thank Lakshmana Pendyala, MD, for his assistance in the review of articles to determine whether they met the required criteria for relevance. Mark Ebell, MD, Matthew Muller, MD, and Christopher Woods, MD, provided excellent advice on earlier drafts of the manuscript.

REFERENCES

- Thompson WW, Shay DK, Weintraub E, et al. Mortality associated with influenza and respiratory syncytial virus in the United States. *JAMA*. 2003;289(2):179-186.
- Thompson WW, Shay DK, Weintraub E, et al. Influenza-associated hospitalizations in the United States. *JAMA*. 2004;292(11):1333-1340.
- Nichol KL. Cost-benefit analysis of a strategy to vaccinate healthy working adults against influenza. *Arch Intern Med*. 2001;161(5):749-759.
- Centers for Disease Control and Prevention. Update: influenza activity—United States, 2003-04 season. *MMWR Morb Mortal Wkly Rep*. 2003;52(49):1197-1202.
- Centers for Disease Control and Prevention. 2004-05 Flu vaccine shortage: interim influenza vaccination recommendations—2004-05 influenza season. <http://www.cdc.gov/flu/protect/vaccineshortage>. Accessed October 5, 2004.
- Lee PY, Matchar DB, Clements DA, Huber J, Hamilton JD, Peterson ED. Economic analysis of influenza vaccination and antiviral treatment for healthy working adults. *Ann Intern Med*. 2002;137(4):225-231.
- Booth CM, Matukas LM, Tomlinson GA, et al. Clinical features and short-term outcomes of 144 patients with SARS in the greater Toronto area. *JAMA*. 2003;289(21):2801-2809.
- Ebell MH, White LL, Casault T. A systematic review of the history and physical examination to diagnose influenza. *J Am Board Fam Pract*. 2004;17(1):1-5.
- Boivin G, Hardy I, Tellier G, Maziade J. Predicting influenza infections during epidemics with use of a clinical case definition. *Clin Infect Dis*. 2000;31(5):1166-1169.
- Carman WF, Wallace LA, Walker J, et al. Rapid virological surveillance of community influenza infection in general practice. *BMJ*. 2000;321(7263):736-737.
- Carrat F, Tachet A, Housset B, Valleron AJ, Rouzioux C. Influenza and influenza-like illness in general practice: drawing lessons for surveillance from a pilot study in Paris, France. *Br J Gen Pract*. 1997;47(417):217-220.
- Carrat F, Tachet A, Rouzioux C, Housset B, Valleron AJ. Evaluation of clinical case definitions of influenza: detailed investigation of patients during the 1995-1996 epidemic in France. *Clin Infect Dis*. 1999;28(2):283-290.
- de Arruda E, Hayden FG, McAuliffe JF, et al. Acute respiratory viral infections in ambulatory children of urban northeast Brazil. *J Infect Dis*. 1991;164(2):252-258.
- Govaert TM, Dinant GJ, Aretz K, Knottnerus JA. The predictive value of influenza symptomatology in elderly people. *Fam Pract*. 1998;15(1):16-22.
- Hak E, Moons KG, Verheij TJ, Hoes AW. Clinical signs and symptoms predicting influenza infection. *Arch Intern Med*. 2001;161(10):1351-1352.
- Hall CB, Long CE, Schnabel KC. Respiratory syncytial virus infections in previously healthy working adults. *Clin Infect Dis*. 2001;33(6):792-796.
- Hulson TD, Mold JW, Scheid D, et al. Diagnosing influenza: the value of clinical clues and laboratory tests. *J Fam Pract*. 2001;50(12):1051-1056.
- Lina B, Valette M, Foray S, et al. Surveillance of community-acquired viral infections due to respiratory viruses in Rhone-Alpes (France) during winter 1994 to 1995. *J Clin Microbiol*. 1996;34(12):3007-3011.
- Long CE, Hall CB, Cunningham CK, et al. Influenza surveillance in community-dwelling elderly compared with children. *Arch Fam Med*. 1997;6(5):459-465.
- Monto AS, Gravenstein S, Elliott M, Colopy M, Schweinle J. Clinical signs and symptoms predicting influenza infection. *Arch Intern Med*. 2000;160(21):3243-3247.
- Poland GA. Patterns of respiratory illness among elderly persons: the value of communitywide surveillance studies for influenza. *Arch Fam Med*. 1997;6(5):466-467.
- Schmid ML, Kudesia G, Wake S, Read RC. Prospective comparative study of culture specimens and methods in diagnosing influenza in adults. *BMJ*. 1998;316(7127):275.
- van Elden LJ, van Essen GA, Boucher CA, et al. Clinical diagnosis of influenza virus infection: evaluation of diagnostic tools in general practice. *Br J Gen Pract*. 2001;51(469):630-634.
- Zambon M, Hays J, Webster A, Newman R, Keene O. Diagnosis of influenza in the community: relationship of clinical diagnosis to confirmed virological, serologic, or molecular detection of influenza. *Arch Intern Med*. 2001;161(17):2116-2122.

25. Nicholson KG, Kent J, Hammersley V, Cancio E. Acute viral infections of upper respiratory tract in elderly people living in the community: comparative, prospective, population based study of disease burden. *BMJ*. 1997;315(7115):1060-1064.
26. Holleman DR Jr, Simel DL. Does the clinical examination predict airflow limitation? *JAMA*. 1995;273(4):313-319.
27. Glas AS, Lijmer JG, Prins MH, Bonsel GJ, Bossuyt PM. The diagnostic odds ratio: a single indicator of test performance. *J Clin Epidemiol*. 2003;56(11): 1129-1135.
28. Mantel N, Brown C, Byar DP. Tests for homogeneity of effect in an epidemiologic investigation. *Am J Epidemiol*. 1977;106(2):125-129.
29. Fleiss JL, Gross AJ. Meta-analysis in epidemiology, with special reference to studies of the association between exposure to environmental tobacco smoke and lung cancer: a critique. *J Clin Epidemiol*. 1991;44(2):127-139.
30. Deville WL, Buntinx F. Guidelines for conducting systematic reviews of studies evaluating the accuracy of diagnostic tests. In: Knotterus JA, ed. *The Evidence Base of Clinical Diagnosis*. London, England: BMJ Books; 2002:156-159.
31. Bellei N, Benfica D, Perosa AH, Carlucci R, Barros M, Granato C. Evaluation of a rapid test (QuickVue) compared with the shell vial assay for detection of influenza virus clearance after antiviral treatment. *J Virol Methods*. 2003;109(1):85-88.
32. Marcante R, Chiumento F, Palu G, Cavedon G. Rapid diagnosis of influenza type A infection: comparison of shell-vial culture, directigen flu-A and enzyme-linked immunosorbent assay. *New Microbiol*. 1996;19(2):141-147.
33. Noyola DE, Paredes AJ, Clark B, Demmler GJ. Evaluation of a neuraminidase detection assay for the rapid detection of influenza A and B virus in children. *Pediatr Dev Pathol*. 2000;3(2):162-167.
34. Quach C, Newby D, Daoust G, Rubin E, McDonald J. QuickVue influenza test for rapid detection of influenza A and B viruses in a pediatric population. *Clin Diagn Lab Immunol*. 2002;9(4):925-926.
35. Rodriguez WJ, Schwartz RH, Thorne MM. Evaluation of diagnostic tests for influenza in a pediatric practice. *Pediatr Infect Dis J*. 2002;21(3):193-196.
36. Cazacu AC, Demmler GJ, Neuman MA, et al. Comparison of a new lateral-flow chromatographic membrane immunoassay to viral culture for rapid detection and differentiation of influenza A and B viruses in respiratory specimens. *J Clin Microbiol*. 2004;42(8):3661-3664.
37. Cox NJ, Subbarao K. Influenza. *Lancet*. 1999;354(9186):1277-1282.
38. Rothberg MB, Bellantonio S, Rose DN. Management of influenza in adults older than 65 years of age: cost-effectiveness of rapid testing and antiviral therapy. *Ann Intern Med*. 2003;139(5 pt 1):321-329.
39. Smith KJ, Roberts MS. Cost-effectiveness of newer treatment strategies for influenza. *Am J Med*. 2002;113(4):300-307.
40. Demicheli V, Rivetti D, Deeks JJ, Jefferson TO. Vaccines for preventing influenza in healthy adults. *Cochrane Database Syst Rev*. 2004;(3): CD001269.
41. Govaert TM, Thijs CT, Masurel N, Sprenger MJ, Dinant GJ, Knotterus JA. The efficacy of influenza vaccination in elderly individuals: a randomized double-blind placebo-controlled trial. *JAMA*. 1994;272(21): 1661-1665.
42. Bonner AB, Monroe KW, Talley LI, Klasner AE, Kimberlin DW. Impact of the rapid diagnosis of influenza on physician decision-making and patient management in the pediatric emergency department: results of a randomized, prospective, controlled trial. *Pediatrics*. 2003;112(2):363-367.

Prepared by David L. Simel, MD, MHS

Reviewed by W. Paul McKinney, MD

CLINICAL SCENARIO

A 68-year-old woman calls for an urgent appointment before a midwinter trip to visit with her family and new grandchildren. She has a slight increase in her temperature (37°C [99.8°F]), but she has a cough and has felt gradually worse for 2 days. She wants to know whether she should cancel her trip. Her medical history reveals only hypertension, for which she takes a single medication.

The week before her call, you had not noticed any change in the steady number of 2 to 3 patients a day appearing in your office with similar symptoms. However, the day she called there were 5 similar calls from patients with influenza-like illness.

UPDATED SUMMARY ON INFLUENZA

Original Review

Call SA, Vollenwieder MA, Hornung CA, Simel DL, McKinney WP. Does this patient have influenza? *JAMA*. 2005;293(8):987-997.

The Update was prepared within 12 months of publication of The Rational Clinical Examination article, so the “Make the Diagnosis” section summarizes the findings published in the original article.

CLINICAL SCENARIO—RESOLUTION

This clinical scenario highlights the need to understand the local epidemiology of influenza and case definition for influenza-like illness. Once you notice a possible change in the number of patients with influenza-like illness, you should first check the Centers for Disease Control and Prevention (CDC) data to see whether the rate of influenza-like illness is increasing in your region and whether

influenza has been isolated in reference laboratories. When you look at the CDC Web site, you notice that there is a steady increase in influenza-like illness being reported, along with evidence of influenza being increasingly isolated in reference laboratories.

Her symptoms do not, however, identify her as having an influenza-like illness because she does not have a temperature higher than 38°C (100°F). The CDC definition of influenza-like illness requires the appropriate temperature, accompanied by either a cough or sore throat. If you decided that she may really have had a fever but simply did not capture it with self-measurement, you should recognize that only her malaise (likelihood ratio [LR], 2.6) increases the likelihood that her illness is influenza. A cough alone, lack of acute onset, and the absence of fever in a patient older than 60 years do not have LRs sufficiently different from 1 and therefore do not provide you enough information. Because of the increasing rate of influenza and her concern, you ask her to come to the office for a rapid influenza test.

The problem in deciding to order the rapid influenza test resides in the difficulty with estimating the prior probability. Cost-effectiveness studies show that testing is the appropriate strategy when the prior probability is between 5% and 14%; however, many clinicians will be uncertain. If you estimated that the probability of influenza was at the upper end of the testing strategy (say, 15%), the negative result lowers the likelihood of influenza to approximately 1%. However, if you had thought that the probability of influenza was as high as 50% and ordered a test (rather than treated empirically), the negative rapid influenza test result (LR, 0.06) decreases the probability to 5.7%. It is most likely that she has some other type of viral infection, which could also spread to her family, but the use of antivirals and advice about influenza should be based on the low probability estimate from the rapid influenza test.

INFLUENZA—MAKE THE DIAGNOSIS

PRIOR PROBABILITY

Clinicians must rely on informed clinical judgment to determine the prior probability of influenza, which requires an understanding of the epidemiology weekly reports for influenza, available from the CDC (for the United States, <http://www.cdc.gov/flu/weekly/>, and for International surveillance, <http://www.cdc.gov/flu/weekly/intsurv.htm>; accessed June 1, 2008) that applies to your population. For US data, the CDC reports the frequency of influenza-like illness during influenza season and whether influenza virus is present in your region. However, they do not report whether a given patient with an influenza-like illness is likely to have influenza.

POPULATION FOR WHOM INFLUENZA SHOULD BE CONSIDERED

The clinical evaluation is used to identify influenza-like illness among all patients during influenza season.

DETECTING THE LIKELIHOOD OF INFLUENZA

A few findings can lower the likelihood of influenza among patients (all adults or children) when they have influenza-like illness ([Table 26-5](#)).

Table 26-5 Likelihood Ratios for Findings Across All Age Groups That Lower the Probability of Influenza

Adults or Children	LR (95% CI)
Absence of fever	0.40 (0.25-0.66)
Absence of cough	0.42 (0.31-0.57)
Presence of nasal congestion	0.49 (0.42-0.59)

Abbreviations: CI, confidence interval; LR, likelihood ratio.

For older adults, the presence of a few findings increases the likelihood of influenza, whereas the presence of 1 finding (sneezing) makes influenza a little less likely ([Table 26-6](#)).

Table 26-6 Likelihood Ratios for Findings That Suggest Influenza in Older Adults

Adults > 60 y	LR (95% CI)
Fever and cough, combined with acute onset	5.4 (3.8-7.7)
Fever	3.8 (2.8-5.0)
Malaise	2.6 (2.2-3.1)
Chills	2.6 (2.0-3.2)
Sneezing	0.47 (0.24-0.92)

Abbreviations: CI, confidence interval; LR, likelihood ratio.

During influenza season, a negative result on a commercially available rapid influenza tests greatly decreases the likelihood that a patient has influenza ([Table 26-7](#)). These studies may be most useful as the rate of influenza-like illness is increasing on visits to sentinel providers and before data are available from the CDC that suggest influenza is circulating in your area. The CDC provides updated guidance on available rapid influenza tests and their role in screening for influenza (<http://www.cdc.gov/flu/professionals/diagnosis/>; accessed June 1, 2008). We cannot be absolutely certain that the operating characteristics of these tests will stay constant from one influenza season to the next.

Table 26-7 Likelihood Ratios for Some Rapid Influenza Tests

	LR+ (95% CI)	LR- (95% CI)
Rapid influenza tests ^a	4.7 (3.6-6.2)	0.06 (0.03-0.12)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

^aDirectgen Flue A (Becton-Dickinson, Franklin Lakes, New Jersey), FLU OIA (Thermo BioStar, Boulder, Colorado), Quick Vue Influenza Test (Quidel Corp, San Diego, California).

REFERENCE STANDARD TESTS

Viral culture is the reference standard test, often accompanied by polymerase chain reaction tests. Some studies use a 4-fold increase in viral antibody titers. Both of these reference standards are suitable only for epidemiologic surveillance since neither result would be available to help guide the treatment of an individual patient because of the long turnaround time required for results.

CHAPTER 27

Does This Patient Have a Torn Meniscus or Ligament of the Knee?

Daniel H. Solomon, MD, MPH

David L. Simel, MD, MHS

David W. Bates, MD, MSc

Jeffrey N. Katz, MD, MSc

Jonathan L. Schaffer, MD, MBA

CLINICAL SCENARIOS

CASE 1 A 20-year-old man presents to your office complaining of knee pain after playing basketball. During the game, as he came down after jumping for a rebound, another player fell on the back of his calf. He remembers hearing a pop and had pain on standing, preventing him from playing in the remainder of the game. While on the bench, he noticed that the pain improved, but his knee swelled. He iced the knee and was able to put some weight on it later that day. Today, putting all of his weight on the knee makes it feel as if it will buckle.

CASE 2 A 72-year-old woman observes that her left knee swells. She has had pain in the medial aspect of the knee for several years, but only recently did she notice the fullness. She fell several weeks ago. The knee hurts her constantly but feels worse going down stairs, especially early in the morning and late in the day. She finds acetaminophen helps, but the pain relief is not adequate for her to be fully active.

WHY IS THE DIAGNOSIS IMPORTANT?

Ten percent to 15% of adults in the community report knee symptoms, with more than 3.3 million new visits made to physicians annually.^{1,2} Overall, knee pain accounts for 3% to 5% of all visits to physicians, and a substantial proportion results in referrals for diagnostic imaging or specialty care.³ The history and physical examination can assist the examiner in determining whether the knee pain is part of a systemic condition or whether it represents a local musculoskeletal problem. When the knee pain is part of a local regional musculoskeletal disorder, the clinician must decide whether the pain represents a torn meniscal or ligamentous structure and then whether nonoperative or operative intervention is indicated. Because torn meniscal or ligamentous structures can cause significant pain and disability, injuries to these structures may require expeditious repair. The physical examination can aid the primary care clinician in assessing the likelihood of a torn meniscal or ligamentous structure and whether a referral will be beneficial.

Although musculoskeletal conditions are common and costly, physicians in training receive little instruction in musculoskeletal medicine.^{4,5} This educational deficit potentially leads to suboptimal treatment. Several studies have suggested that the musculoskeletal examination can be effectively taught through the use of small-group teaching and trained actors playing the role of the patient-educators.^{6,7} The purpose of this review is to analyze the diagnostic accuracy of the physical examination for meniscal and ligamentous injuries.

Anatomy of Meniscal and Ligamentous Knee Injuries and Their Relationship to Symptoms

The knee joint is the largest articulation in the body. It is a modified hinge with an extensive range of motion. The stability of the joint is provided by the soft tissue structures: the anterior cruciate ligament (ACL) and the posterior cruciate ligament (PCL), the medial collateral ligament (MCL) and the lateral collateral ligament (LCL), the menisci, the capsule, and the muscles (Figure 27-1). The ACL and PCL add stability to the joint and aid in proprioception. The subcutaneous location in a weight-bearing extremity, combined with the relatively long lever arms exerting forces on the joint, renders the knee susceptible to injury. All of the structures that compose the knee joint synchronously function through a normal, physiologic range of motion. Knee symptoms occur when any of these structures are altered, potentially creating interference with normal knee function.

An anatomic description of the knee provides a basis for understanding the various injury patterns. The ligaments passively limit the motion of the joint, thus providing stabilization. The ACL and PCL limit the anterior and posterior displacement of the tibia on the femur, respectively. Because the intact ACL prevents anterior motion of the tibia on the femur, an ACL injury leads to abnormal forward movement of the tibial plateau. This abnormal motion leads to relative internal rotation of the tibia during the terminal part of extension. Absence of a functioning ACL and the related anterolateral rotatory instability can result in the sensation that the knee is buckling or giving out. These symptoms occur with normal walking but may be most prominent during pivoting movements, such as those that occur with quick changes in direction. In the absence of knee buckling, patients with ACL disruption may express a loss of confidence in the stability of their knee, possibly because of the ACL's role in proprioception.⁸

The PCL provides stability to most forces regardless of knee position. Isolated disruption of the PCL permits the

tibia to displace posteriorly, decreasing the forces on the posterior horns of the menisci and increasing the forces directly on the articular surfaces of the medial compartment and patellofemoral joint. Although absence of the PCL may have no associated symptoms, it may result in hyperextension of the knee, posterior displacement of the tibia during knee flexion, and varus (bowlegged) and valgus (knock-kneed) angulation with the knee extended. Knee buckling may occur, especially during pivoting motions or when descending stairs. Symptomatic PCL lesions are more common in patients with chronic tears or with acute tears associated with other ligament injuries.

The meniscal fibrocartilages are semilunar, crescentic-shaped structures that are attached to the tibial plateau at the edge of the articulating surfaces of the femur and tibia. The menisci are wedge shaped, with a thin free edge at the inner margin and a wide base attaching to the tibia by the coronary ligaments. The surface is flat inferiorly and concave superiorly, providing a congruous surface for the transmission of 50% of the axial forces across the knee joint.⁹ The menisci increase joint stability, facilitate nutrition, and provide lubrication and shock absorption for the articular cartilage.¹⁰ The lateral meniscus is larger than the medial meniscus and less firmly attached to the tibia, resulting in a more mobile structure.¹⁰ The medial meniscus, firmly attached to the capsule and MCL, is relatively immobile. Because of its fixed nature, combined with the greater force transmission across the medial aspect of the knee, the medial meniscus is more susceptible to injury.¹⁰ Knee flexion forces the menisci posteriorly. In extreme flexion, the posterior portion of the meniscus is firmly compressed between the posterior portion of the tibial plateau and femoral condyle.

Mechanism of Meniscal and Ligamentous Knee Injuries

The position of the joint at the time of the traumatic force dictates which anatomic structures are at risk for injury;

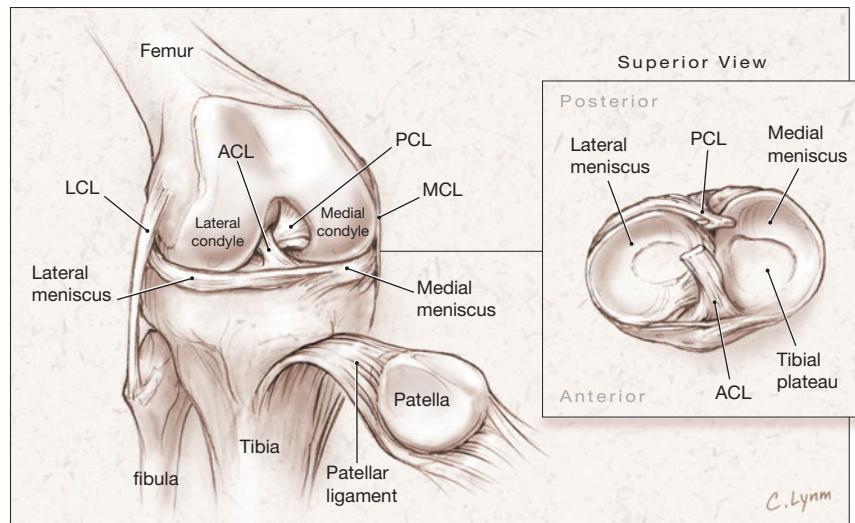


Figure 27-1 Anatomy of the Knee

Right knee shown. Abbreviations: ACL, anterior cruciate ligament; LCL, lateral collateral ligament; MCL, medial collateral ligament; PCL, posterior cruciate ligament.

hence, an important aspect of obtaining the patient's history for acute injuries is to allow him or her to describe the position of the knee and direction of forces at the time it was injured. In full knee extension, the ACL and PCL limit the anteroposterior motion of the tibia on the femur. The ACL is often injured during traumatic twisting injuries in which the tibia moves forward with respect to the femur, often accompanied by valgus stress. No direct blow to the knee or leg is required, but the foot is usually planted, and the patient may remember a popping sensation at the injury. Similar to the ACL, PCL injuries often occur during twisting with a planted foot, in which the force of the injury is directed posteriorly against the tibia with the knee flexed.

The most common collateral ligament injury results from an abduction and external rotation force applied on a knee in an extended or slightly flexed position. An intact MCL helps the ACL prevent posterior motion of the femur. An injury to the MCL may allow for anterior subluxation of the tibial plateau during flexion, especially in an ACL-deficient patient.

Meniscal injuries typically occur through application of specific forces while the knee joint is in certain positions. During flexion, if the tibia is rotated internally, the posterior horn of the medial meniscus is pulled toward the center of the joint. This movement can produce a traction injury of the medial meniscus, tearing it from its peripheral attachment and producing a longitudinal tear of the substance of the meniscus. With aging, the meniscal tissue degenerates and can delaminate, thus making it more susceptible to splitting from shear stress, resulting in horizontal cleavage tears. Without the menisci, the loads on the articular surfaces are increased significantly, leading to a greater potential for degenerative arthritis. Because the menisci are without pain fibers, it is the tearing and bleeding into the peripheral attachments, as well as traction on the capsule, that most likely produce a patient's symptoms of pain. In fact, 16% of asymptomatic patients have meniscal tears demonstrated on magnetic resonance imaging (MRI), with the incidence increasing to 36% for patients older than 45 years.¹¹ Older patients are more likely to have degenerative meniscal tears with fewer mechanical symptoms and an insidious onset.

With posterior horn tears, the meniscus can return to its anatomic position with extension. If the tear extends anteriorly beyond the MCL, creating a bucket-handle tear, then the unstable meniscus fragment cannot always move back into an anatomic position. Such a meniscal tear can result in locking of the knee in a flexed position. Locking of the knee is more common in younger patients with meniscal tears. The lateral meniscus, being more mobile, is less likely to be associated with locking when torn. With walking, traction on medial or lateral meniscal tears may create a clicking sensation.

Epidemiology of Meniscal and Ligamentous Knee Injuries

Injuries to the collateral or cruciate ligaments or to menisci are difficult to account for entirely because many are diagnosed without imaging or arthroscopic confirmation and many go undiagnosed. Data collected through surveys of

athletes participating in organized sports or information collected at sports medicine clinics provide the most reliable data but do not represent the true spectrum of meniscal and ligamentous knee injuries. In a 7-year study of trainees at the US Naval Academy, women had a relative risk of 2.44 for ACL injury compared with men.¹² A similar increased risk for ACL injuries was also observed among female competitive alpine skiers.¹³ A Norwegian study of soccer players with verified ACL injuries suggested that there were 0.063 injuries per 1000 game-hours; women had an almost 2-fold greater incidence of ACL injuries than did men.¹⁴ Although a number of other studies exist, there are few epidemiologic data regarding other meniscal and ligamentous knee injuries.^{15,16}

Clinical Examination for Internal Derangement of the Knee

The purpose of the examination is to make a correct anatomic diagnosis. The patient should be allowed to recite the history of the knee discomfort without interruption. After the history has been taken, the examiner inspects, palpates, and assesses function of the unaffected (or less affected) knee. Examining the healthy knee first creates trust that the examiner is not trying to cause pain and distracts the patient somewhat from the actual maneuvers, allowing greater relaxation. The knees should be examined with the patient in a position that makes him or her most comfortable.¹⁷ The healthy knee must be examined because an essential component of interpreting the findings in the affected knee is the comparison between knees. The following sections describe the cardinal features of a knee examination for a meniscal or ligamentous injury.

Inspection

After resolution of acute symptoms, a patient's gait should be observed. Patients will usually assume a position that provides them the most comfort. If the patient is seated on the examination table, the affected knee will be flexed and hanging off the edge. The quadriceps and calves should be evaluated for atrophy, often present after ligamentous injuries. The knees should be inspected for asymmetry that may indicate swelling. An early sign of effusion is the loss of the peripatellar groove on either side of the patella, observed best with the patient supine. Also, swelling over the medial or lateral aspect of the joint should be recorded and may indicate local inflammation over the collateral ligaments.

Palpation

Differences in temperature between the knees suggest inflammation. With the patient supine, the knees should be examined for an effusion and discomfort with patellar motion. An effusion can be detected by noticing the loss of the peripatellar groove and by palpation of the fluid. Smaller effusions may be detected by compressing the medial and superior aspects of the knee and then pressing or tapping the lateral aspect to create a fluid wave. A perceptible bulge on the medial aspect suggests a small effusion; this sign may not be present with larger effusions. Ballottement of the patella may also be a useful technique for detecting an effusion. The

examiner quickly pushes down on the patella. In the normal knee joint with minimal free fluid, the patella moves directly into the femoral condyle, and there is no tapping sensation underneath the examiner's fingertips. However, in the knee with excess fluid, the patella is floating; thus, ballottement causes the patella to tap against the femoral condyle. This sensation is transmitted to the examiner's fingertips. Localized swelling over specific knee structures, such as the MCL or LCL, can also be observed. Crepitus, a palpable grating sensation, may be produced during certain motions in joints with cartilage disruption.¹⁸ The maneuvers producing crepitus, the location of the crepitus, and any pain elicited should be recorded. Joint line tenderness can also be detected by palpating medial and lateral to the patella in the groove between the femoral condyle and the tibia.

Function

The Lachman test, anterior drawer test, and lateral pivot shift test are the 3 physical examination maneuvers commonly used to assess the integrity of the ACL (Figure 27-2).

Although the patient may be fearful, these functional tests should not cause pain with isolated ACL injuries in the subacute setting.

Lachman test is typically performed while the patient lies supine with the knee flexed to 20 to 30 degrees.¹⁹ The examiner stands to the side of the patient's leg, with the patient's heel on the examination table. The femur is grasped with one hand just above the knee. While the examiner grasps the femur firmly to prohibit motion of the upper leg and to relax the hamstrings, the other hand grasps the proximal tibia. The lower leg is then given a brisk forward tug, and a discrete end point should be felt. A positive test result is one in which the end point is not discrete or there is increased anterior translation of the tibia. The test is more difficult to perform when the examiner has small hands or the patient has large legs, both situations making it more difficult to completely grasp the legs. In this situation, the patient may be placed prone, with the knee at the same degree of flexion while the examiner attempts the same motion of the tibia.

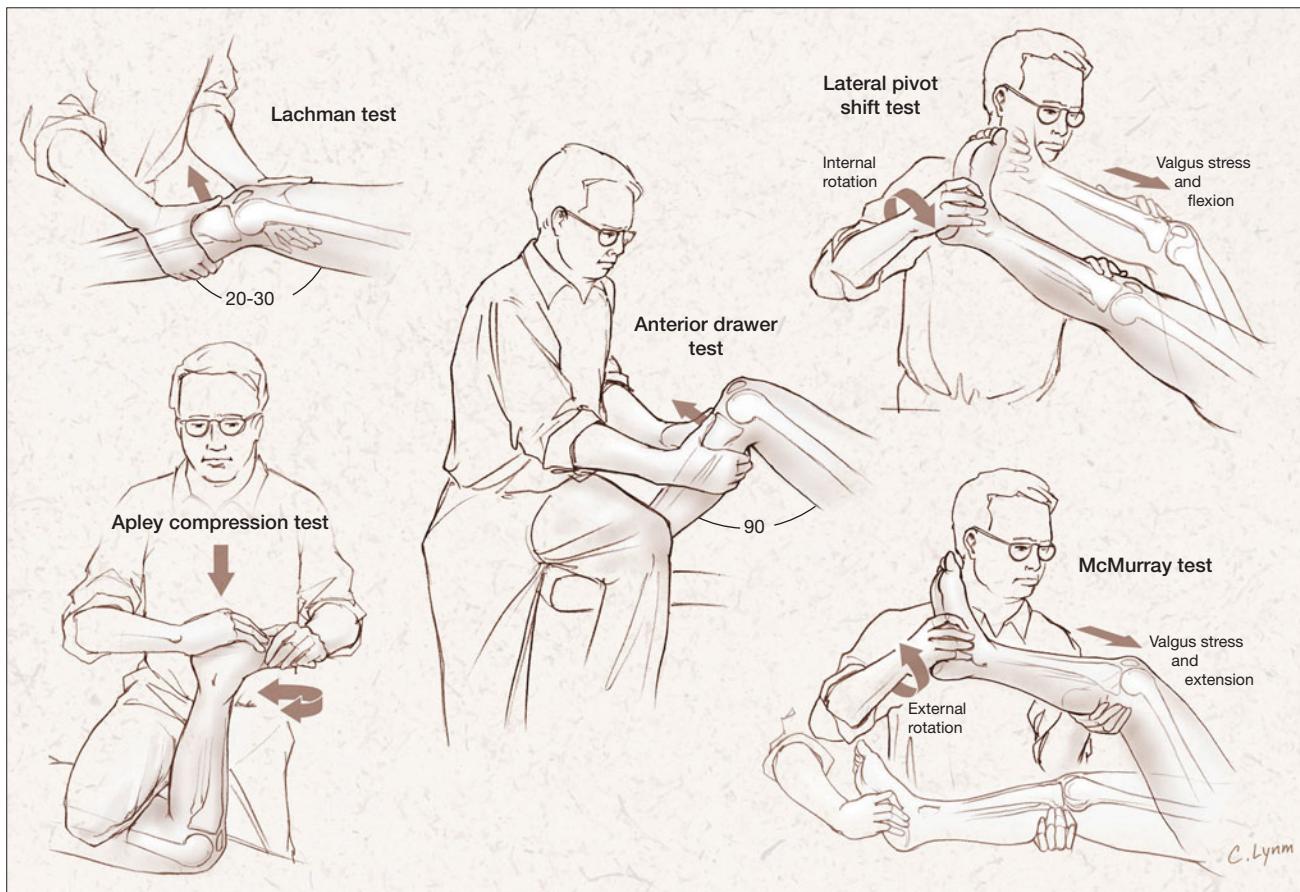


Figure 27-2 Examination Maneuvers

Right knee shown. Examination maneuvers include the Lachman, anterior drawer, lateral pivot shift, Apley compression, and McMurray tests. Lachman test, performed to detect anterior cruciate ligament (ACL) injuries, is conducted with the patient supine and the knee flexed to 20 to 30 degrees. The anterior drawer test detects ACL injuries and is performed with the patient supine and the knee in 90 degrees of flexion. The lateral pivot shift test is performed with the patient supine, the hip flexed 45 degrees, and the knee in full extension. Internal rotation is applied to the tibia while the knee is flexed to 40 degrees under a valgus stress (pushing the outside of the knee medially). The Apley compression test, used to assess meniscal integrity, is performed with the patient prone and the examiner's knee over the patient's posterior thigh. The tibia is externally rotated while a downward compressive force is applied over the tibia. The McMurray test, used to assess meniscal integrity, is performed with the patient supine and the examiner standing on the side of the affected knee.

The anterior drawer test is also performed with the patient supine and the knee in 90 degrees of flexion. The examiner quickly pulls the upper portion of the calf forward, using both hands. The tibia must not be rotated, and the hamstrings must be relaxed to properly assess the ACL. An intact ACL abruptly stops the tibia's forward motion as the ACL reaches its maximum length. If the tibia can be moved anteriorly without an abrupt stop, referred to as a discrete end point, this is considered a positive anterior drawer sign. It is often useful to perform this test on the uninjured knee to determine whether the amount of anterior translation differs between knees.

The lateral pivot shift test combines a valgus stress (pushing the outside of the knee medially) with a twisting force while the knee is being flexed (see Figure 27-2). In Losee's²⁰ version of the test, the patient rests on his or her back with the knee at 45 degrees' flexion. The examiner places a hand on the lateral aspect of the knee and pushes medially, creating a valgus strain. At the same time, the examiner's other hand supports and pulls the foot laterally. As the examiner slowly extends the knee, the tibia and foot begin to twist internally. A positive test result consists of an obvious thud or jerk at 10 to 20 degrees' flexion in the ACL-deficient knee, representing anterior subluxation of the tibia on the femur.

Posterior or PCL stability is generally assessed with the posterior drawer test, which is performed with the patient supine and the knee flexed to 90 degrees. The alignment of the knees is inspected; if the tibia of the affected knee is subluxed posteriorly (a posterior sag), then applying anterior pressure will correct the sag. If the subluxation can be corrected, it is considered a positive posterior drawer sign. Others consider a posterior drawer test result to be positive if a posterior force on the tibia encounters no discrete end point, the reverse of the anterior drawer test. A method of assessing whether a PCL injury is present in combination with an MCL injury is to perform the abduction (or valgus) stress test with the knee in 2 positions. First, with the knee in 30 degrees of flexion, the examiner supports the foot or ankle of the leg being examined and places the other hand along the lateral aspect of the knee. An inward or medial force is then applied to the knee while an opposite force is applied to the lower leg. The examiner grades the opening of the medial compartment of the knee. If the opening is larger on the injured side than on the opposite side, an MCL injury is suggested. The same test is then carried out with the knee held in full extension. Normally, the abduction stress test produces no opening of the medial compartment when the knee is fully extended in a patient with an intact PCL and MCL. If the opening of the medial compartment is similar with the knee in full extension, a combined PCL and MCL injury is suspected.

Finally, meniscal integrity is assessed with several specific examination maneuvers, including the McMurray test, the Apley compression test, and the medial-lateral grind test (Figure 27-2). The McMurray test is performed with the patient supine. The examiner stands on the side of the affected knee and places one hand on the heel and another along the medial aspect of the knee, providing a valgus force.

The knee is extended from a fully flexed position while the tibia is rotated internally. The test is repeated while the tibia is rotated externally. A positive sign is indicated by a popping and sensation of symptoms along the joint line, often accompanied by an inability to fully extend the knee.

The Apley compression test is performed with the patient lying in a prone position on a low examination table. The examiner applies his or her knee into the posterior thigh of the leg to be examined and then flexes and externally rotates the tibia while gripping the ankle. The examiner then compresses the tibia downward. If this compression produces an increase in pain, the test result is considered positive.

The medial-lateral grind test is performed with the patient supine on the examination table. The examiner cradles the affected leg's calf in one hand and places the index finger and thumb of the opposite hand over the joint line. Valgus and varus stresses are applied to the tibia during flexion and extension. If a grinding sensation is palpated by the hand placed over the joint line, the medial-lateral grind test result is deemed positive.

METHODS

Search Strategy

We conducted MEDLINE and HealthSTAR searches to retrieve articles pertaining to the physical examination of patients with suspected meniscal or ligamentous injury of the knee. The search of MEDLINE and HealthSTAR included all years from 1966 and 1975, respectively; both searches were extended through December 31, 2000. Keywords for searching included "knee," "physical examination," "internal derangement," "anterior cruciate ligament," "posterior cruciate ligament," "medial collateral ligament," "lateral collateral ligament," and "meniscus." Reference lists from relevant articles were also manually searched. Searching was limited to English-language articles describing human studies.

A total of 88 articles were retrieved. We included 26 articles that compared the performance of the physical examination of the knee to a reference standard, such as arthroscopy, arthrotomy, or MRI. Three articles were subsequently excluded because no primary data were reported, only aggregated sensitivities and specificities. Several categories of physical examination findings were included: widely available maneuvers, maneuvers requiring specialized equipment, and general knee examination without specific maneuvers. We did not include data examining the accuracy of arthrometry or examination under anesthesia because both of these examination techniques are not widely available. Two of the authors, a rheumatologist (D.H.S.) and an orthopedic surgeon (J.L.S.), graded each article for its methodologic quality, using a standardized scoring system.²¹ The scoring system included assembly of the study (consecutive or otherwise), the relevance of the patient enrolled, the appropriateness and completeness of the reference standard, and the blinding of the examiner.

Articles contained level 1 evidence if they used an independent, blind comparison of the examination with the refer-

ence standard among at least 50 consecutive and relevant patients. Level 2 articles were similar in their methods but contained fewer than 50 patients. If patients were not recruited consecutively, but the authors conducted an independent and blind comparison with the reference standard, then the article was considered level 3. Level 4 evidence came from articles that compared the examination with a reference standard, but patients were not collected consecutively, nor was the comparison independent.

We noted whether the patients included in each study had acute or chronic knee symptoms and whether the examiner was a specialist in musculoskeletal care. However, no studies reported data separately for nonspecialist examiners. Data were abstracted from each article to allow for calculation of the sensitivity and specificity of each physical examination finding. Several articles commented on the composite examination for meniscal or ligamentous injuries. These articles did not include data for specific examination maneuvers; rather, all aspects of the physical examination were combined in an unspecified manner.

Analysis

Sensitivity was calculated as the percentage of patients with a given lesion on the reference standard (usually arthroscopy or arthrotomy) who had an abnormal physical examination result; specificity was the percentage of patients without a given lesion who had normal results on the physical examination maneuver.²² When sensitivity and specificity were available, likelihood ratios (LRs) with 95% confidence intervals (CIs) were calculated according to the method of Simel et al²³ and Hasselblad and Hedges.²⁴ The LR for a positive test result was calculated as the sensitivity divided by (1 minus specificity); for a negative test result, the LR equaled (1 minus sensitivity) divided by specificity. When several studies provided data to calculate the LRs for the same examination maneuver, a summary LR was estimated from a random-effects model to provide conservative values.²⁵

RESULTS

No articles could be identified that adequately examined the diagnostic accuracy of the physical examination for MCL or LCL lesions. Hence, these structures will not be discussed further.

Anterior Cruciate Ligament Examination

Three researchers reported on the composite examination for ACL injuries without giving data for specific examination maneuvers (Table 27-1).²⁶⁻²⁸ These investigators found that the sensitivity of the examination for ACL injuries was more than 82% and the specificity was more than 94%. The summary LRs for these studies were 25 (95% CI, 2.1-306) for a positive examination result and 0.04 (95% CI, 0.01-0.48) for a negative examination result. Twelve other studies²⁹⁻⁴⁰ were included that examined the anterior drawer, lateral pivot shift, and the Lachman maneuver tests.

The methodologic quality of these studies was inconsistent; patients primarily had known ruptured ACLs and underwent subsequent arthroscopy or arthrotomy. Because only patients with known lesions were included, calculation of specificity and LRs in all but 4 studies was precluded.

The specificity of the anterior drawer test for ACL ruptures ranged from 23% to 100%, with a mean of 67% (SD, 42%).^{29,36,38} Likewise, the sensitivity of the anterior drawer test varied from 9% to 93%, with a mean of 62% (SD, 23%).^{29,31,32,35-40} Several of these studies were small, which may explain the variability in results. The summary LR (Table 27-2) for a positive anterior drawer test result was 3.8 (95% CI, 0.7-22) and for a negative test result was 0.30 (95% CI, 0.05-1.50). Only 1 study³⁸ reported on the specificity of Lachman test, and it found 100% specificity; however, the authors reported on a population of patients who underwent MRI and subsequent arthroscopy, thus limiting the generalizability of these findings. The sensitivity of Lachman test ranged from 60% to 100%, and the mean was 84% (SD, 15%) (Table 27-1).^{32,34,35,37-40} According to the one study³⁸ that reported both the specificity and sensitivity of Lachman test, the LR for a positive test result was 42 (95% CI, 2.7-651) and for a negative test result was 0.1 (95% CI, 0-0.4) (Table 27-2). The specificity of the lateral pivot shift test has not been reported. The sensitivity of this maneuver varied from 27% to 95%, with a mean of 38% (SD, 28%).^{30-32,35,39}

Posterior Cruciate Ligament Examination

Two studies of the composite examination for PCL injuries reported a mean sensitivity of 91% and specificity of 98% (Table 27-3).^{26,27} The summary LR for a positive general examination result was 21 (95% CI, 2.1-205) and for a negative general examination result, 0.05 (95% CI, 0.01-0.50). Three studies^{36,41,42} analyzed the diagnostic accuracy of specific examination maneuvers. The specificity of the posterior drawer test was not reported in any study. Two studies^{41,42} reported its sensitivity, which ranged from 51% to 86%, with a mean of 55%. The only other examination maneuver tested for PCL lesions was the abduction stress test, examined by the one investigator who originally described the test.³⁶ This test had a sensitivity of 94% and a specificity of 100%. The resulting LR for a positive test result was 94 (95% CI, 6-1487) and for a negative test result, 0.1 (95% CI, 0-0.4).

Meniscal Examination

Nine studies investigated the diagnostic accuracy of the examination for meniscal injuries (Table 27-4); all used arthroscopy as the reference standard.^{27,28,33,43-48} Five of these studies reported the accuracy of the composite examination; the mean sensitivity for the composite examination was 77% (SD, 7%), and the specificity was 91% (SD, 3%).^{27,28,43-45} Four other studies examined specific examination maneuvers.^{33,46-48} Joint line tenderness had a mean sensitivity of 79% (SD, 4%) and a specificity of 15% (SD, 22%).^{33,46-48} The summary LR for a positive test result was 0.9 (95% CI, 0.8-1.0) and for a negative test result, 1.1 (95% CI, 1.0-1.3) (Table 27-5). The mean sensitivity of the McMurray test was 53% (SD, 15%),

Table 27-1 Diagnostic Accuracy of the Physical Examination for Anterior Cruciate Ligament Injuries^a

Source, y	Level of Evidence	No. of Subjects	Patient Population	Reference Standard	Examination Maneuver	Sensitivity, %	Specificity, %
Simonsen et al, ²⁶ 1984	1	118	Consecutive patients with hemarthrosis; acute	Arthroscopy	General examination	62	56
O'Shea et al, ²⁷ 1996	1	156	Consecutive patients with chronic knee pain; acute and chronic	Arthroscopy	General examination	97	100
Rose and Gold, ²⁸ 1996	4	154	Nonconsecutive patients with knee pain; chronic	Arthroscopy	General examination	100	99
Braunstein, ²⁹ 1982	4	29	Consecutive patients who underwent arthrography and then arthroscopy	Arthroscopy	ADT	91	89
Dahlstedt and Dalen, ³⁰ 1989	2	41	Consecutive patients with hemarthrosis but no fracture on radiograph; acute	Arthroscopy	LPST	37	NA
DeHaven, ³¹ 1980	4	35	Consecutive athletes with knee injuries and hemarthrosis; acute	Arthroscopy	ADT	9	NA
					LPST	27	NA
Donaldson et al, ³² 1985	4	101	Nonconsecutive patients from sports medicine clinic found at surgery to have ACL tears; acute	Arthroscopy/arthroscopy	ADT	70	NA
					LPST	35	NA
					Lachman	99	NA
Fowler and Lubliner, ³³ 1989	1	24	Consecutive patients with knee pain; chronic	Arthroscopy	JLT	75	NA
Gurtler et al, ³⁴ 1987	1	75	Consecutive patients with ACL tears; acute and chronic	Arthroscopy	Lachman	100	NA
Hardaker et al, ³⁵ 1990	1	101	Consecutive patients with knee injuries and hemarthrosis presenting to sports medicine clinic; acute	Arthroscopy	ADT	18	NA
					LPST	29	NA
					Lachman	74	NA
Hughston et al, ³⁶ 1976	4	68	Consecutive patients with ruptures of the "medial compartment"; acute	Arthroscopy	ADT	65	23
Jonsson et al, ³⁷ 1982	4	107	Nonconsecutive patients found at surgery to have a ruptured ACL; acute and chronic	Arthroscopy	ADT	93	NA
					Lachman	60	NA
Lee et al, ³⁸ 1988	4	41	Nonconsecutive patients who underwent MRI and then arthroscopy	Arthroscopy	ADT	78	100
					Lachman	89	100
Liu et al, ³⁹ 1995	4	38	Nonconsecutive patients with proven ACL ruptures; acute	Arthroscopy	ADT	63	NA
					LPST	95	NA
					Lachman	72	NA
Mitsou and Vallianatos, ⁴⁰ 1988	4	144	Nonconsecutive patients with proven ACL ruptures; acute and chronic	Arthroscopy/arthroscopy	ADT	72	NA
					Lachman	91	NA

Abbreviations: ACL, anterior cruciate ligament; ADT, anterior drawer test; JLT, joint line tenderness; LPST, lateral pivot shift test; MRI, magnetic resonance imaging; NA, not applicable (no patients without lesions were included).

^aAcute patients refers to those treated within 3 months of their injury and chronic refers to beyond 3 months. If no mention is made of acute or chronic, the authors did not specify.

and the specificity was 59% (SD, 36%).^{33,46-48} The summary LR for a positive test result was 1.3 (95% CI, 0.9-1.7) and for a negative test result, 0.8 (95% CI, 0.6-1.1). Other maneuvers were not formally examined in more than 1 study and included the Apley compression test, the medial-lateral grind test, and the presence of a joint effusion. The Apley compression test had a sensitivity of 16%; no patients without meniscal lesions were tested.³³ The medial-lateral grind test had a sensitivity of 69% and a specificity of 86%.⁴⁶ A joint effusion was found to have a sensitivity of 35% and specificity of 100%; however, this last study only included patients admitted for arthroscopy.⁴⁷

Limitations of Data

Given the relative frequency and economic consequences of meniscal or ligamentous knee injuries, data about the accuracy

Table 27-2 Selected Physical Examination Maneuvers for Ligamentous Knee Injuries^a

Source, y	LR (95% CI)	
	Positive	Negative
Anterior Drawer Test		
Hughston et al, ³⁶ 1976	0.8 (0.6-1.2)	1.5 (0.6-3.8)
Braunstein, ²⁹ 1982	8.2 (2.2-31)	0.1 (0-0.7)
Lee et al, ³⁸ 1988	37 (2.3-576)	0.2 (0.1-0.5)
Summary ^b	3.8 (0.7-22)	0.30 (0.05-1.50)
Lachman Test		
Lee et al, ³⁸ 1988	42 (2.7-651)	0.1 (0-0.4)

Abbreviations: CI, confidence interval; LR, likelihood ratio.

^aIncludes all studies with data supplied to calculate both sensitivity and specificity.

^bCalculated with a random-effects model.

Table 27-3 Diagnostic Accuracy of the Physical Examination for Posterior Cruciate Ligament Injuries^a

Source, y	Level of Evidence	No. of Subjects	Patient Population	Reference Standard	Examination Maneuver	Sensitivity, %	Specificity, %
Simonsen et al, ²⁶ 1984	4	118	Consecutive patients with hemarthrosis; acute	Arthroscopy	General examination	91	80
O'Shea et al, ²⁷ 1996	1	156	Consecutive patients with chronic knee pain; acute and chronic	Arthroscopy	General examination	100	99
Hughston et al, ³⁶ 1976	4	68	Consecutive patients with ruptures of the medial compartment; acute	Arthrotomy	AST	94	100
Baker et al, ⁴¹ 1984	4	40	Nonconsecutive patients with known PCL tear; acute	Arthroscopy	PDT	86	NA
Loos et al, ⁴² 1981	4	59	Nonconsecutive patients with PCL tear; acute	Arthroscopy/arthrotomy	PDT	51	NA

Abbreviations: AST, abduction stress test; NA, indicates not applicable (no patients without lesions were included); PCL, posterior cruciate ligament; PDT, posterior drawer test.

^aAcute patients refers to those treated within 3 months of their injury and chronic refers to beyond 3 months. If no mention is made of acute or chronic, the authors did not specify.

Table 27-4 Diagnostic Accuracy of the Physical Examination for Meniscal Injuries^a

Source, y	Level of Evidence	No. of Subjects	Patient Population	Reference Standard	Examination Maneuver	Sensitivity, %	Specificity, %
Daniel, ⁴³ 1991	4	177	Nonconsecutive patients with suspected meniscal tears	Arthroscopy/arthrotomy	General examination	82	78
Gillies and Seligson, ⁴⁴ 1979	4	50	Nonconsecutive patients with known meniscal tears	Arthrotomy	General examination	64	NA
Miller, ⁴⁵ 1996	4	57	Nonconsecutive patients with known meniscal tears; acute and chronic	Arthroscopy	General examination	81	NA
O'Shea et al, ²⁷ 1996	1	156	Consecutive patients with knee pain; acute and chronic	Arthroscopy	General examination	73	84
Rose and Gold, ²⁸ 1996	4	154	Nonconsecutive patients with knee pain; chronic	Arthroscopy	General examination	79	79
Fowler and Lubliner, ³³ 1989	1	80	Consecutive patients with knee pain; chronic	Arthroscopy	JLT McMurray Apley	85 29 16	NA NA NA
Anderson and Lipscomb, ⁴⁶ 1986	4	100	Consecutive patients suspected of having meniscal tears presenting for arthroscopy; acute and chronic	Arthroscopy/arthrotomy	JLT McMurray MLGT	77 58 69	NA 29 86
Barry et al, ⁴⁷ 1983	4	44	Nonconsecutive patients presenting for meniscectomy	Arthroscopy/arthrotomy	JLT McMurray Effusion	76 56 35	43 100 100
Noble and Erat, ⁴⁸ 1980	4	200	Nonconsecutive patients presenting for meniscectomy; acute and chronic	Arthroscopy	JLT McMurray	79 63	11 57

Abbreviations: JLT, joint line tenderness; MLGT, medial-lateral grind test; NA, indicates not applicable (no patients without lesions were included).

^aAcute patients refers to those treated within 3 months of their injury and chronic refers to beyond 3 months. If no mention is made of acute or chronic, the authors did not specify.

of the physical examination were relatively limited. Although no specific examination maneuver has impressive test performance characteristics, the composite examination results for ACL, PCL, and meniscal lesions are reported to be reasonably sensitive and specific. One possible explanation for this finding is that a constellation of examination findings may be more useful than any one finding. No data were available to judge the accuracy of the physical examination results of the MCL and LCL.

The patient population was an important determinant of the accuracy of the examination. Some investigators included only acute injuries and others, only chronic injuries, whereas some did not specify injury type. The chronicity of the injury may affect the sensitivity and specificity of examination maneuvers. The examination for ACL injuries was less accurate if a hemarthrosis was present because the increased intra-articular volume causes pain that is increased with any examination maneuver. This is a good illustration of spectrum bias, in which the spectrum of patients included in a study affects the diagnostic accuracy of a given test,⁴⁹ and may have accounted for some of the variation in the results reported between articles.

Another potential source of variation was the experience of the examiner and the precise methods used for conducting the physical examination test. It is commonly believed that the examination for meniscal and ligamentous injuries is difficult to learn and that accuracy may therefore increase with experience. Although all of the studies included in this review used orthopedic surgeons, the reports did not describe the examiners' number of years in practice. If experience is an important determinant of accuracy, the data presented in this review should represent an upper limit for less experienced physicians. The definitions of an abnormal or positive physical examination result were not always clear from the articles. Also, the reproducibility of the physical examination was unclear and rarely reported. These sources of variation all contribute to heterogeneity between studies, illustrated by broad 95% CIs in the summary LRs.

The physical examination should be preceded by taking a careful history. Historical findings that may substantially improve the accuracy of the physical examination include the angle and force of impact if an injury occurred; whether the patient heard a pop at the injury; whether the patient has been experiencing catching, locking, or giving way of the knee; and whether the patient had noticed swelling around the knee. The sensitivity and specificity of historical items deserve attention, but we were unable to find published data regarding the sensitivity and specificity of commonly asked questions. Our review suggests that a combination of historical and physical examination findings may be more useful than any one specific item. Future studies must pay careful attention to recruiting an appropriate patient population, including subjects without pathologic lesions. They should also be careful in describing the physical examination, explicitly documenting criteria for abnormal; in calculating interobserver and intraobserver reliability; and in testing the diagnostic accuracy of clinically relevant clusters of historical and examination items.

How to Improve Your Physical Examination Skills

Improving your diagnostic skills for meniscal and ligamentous knee injuries takes practice. The physical examination

Table 27-5 Selected Physical Examination Maneuvers for Meniscal Knee Injuries^a

Source, y	LR (95% CI)	
	Positive	Negative
McMurray Test		
Noble and Erat, ⁴⁸ 1980	1.5 (1.1-2.1)	0.6 (0.5-0.9)
Barry et al, ⁴⁷ 1983	8.9 (0.6-132)	0.5 (0.3-0.7)
Anderson and Lipscomb, ⁴⁶ 1986	0.8 (0.5-1.3)	1.5 (0.4-4.9)
Summary ^b	1.3 (0.9-1.7)	0.8 (0.6-1.1)
Joint Line Tenderness		
Noble and Erat, ⁴⁸ 1980	0.9 (0.8-1.0)	1.9 (0.8-4.3)
Barry et al, ⁴⁷ 1983	1.3 (0.7-2.6)	0.6 (0.2-1.6)
Summary ^b	0.9 (0.8-1.0)	1.1 (1.0-1.3)
Joint Effusion		
Barry et al, ⁴⁷ 1983	5.7 (0.4-86)	0.7 (0.5-0.9)
Medial-Lateral Grind		
Anderson and Lipscomb, ⁴⁶ 1986	4.8 (0.8-30)	0.4 (0.2-0.6)

Abbreviations: CI, confidence interval; LR, likelihood ratio.

^aIncludes all studies with data supplied to calculate both sensitivity and specificity.

^bCalculated with a random-effects model.

can be practiced on healthy patients to develop an examination routine and gain a mental image of healthy anatomy. Patients with knee pain should be examined so that you can describe what you think is the anatomic lesion causing the pain. If you refer the patient, the referral letter should include your presumed anatomic diagnosis, which forces the examination to be more thorough, and it will aid the consultant in his or her evaluation. If the patient undergoes MRI or surgery, compare your assessment with the imaging or surgical findings.

CLINICAL SCENARIOS—RESOLUTIONS

The first case describes a young man with a probable ACL rupture. The angle of injury, the presence of a pop, the difficulty bearing weight, and the transient swelling support this diagnosis. He should be counseled about his prognosis, encouraged to begin a program of quadriceps strengthening, and given the option of pursuing surgical reconstruction if the symptoms are functionally limiting. The second case characterizes a common scenario in primary care practices, the older patient with degenerative joint disease and a probable superimposed degenerative meniscal tear. This patient's functional limitations need to be assessed carefully. If she is not too impaired, joint aspiration of the effusion, nonsteroidal anti-inflammatory drugs, quadriceps strengthening, and a cane may provide enough pain relief and mobility to make more invasive treatment unnecessary. If conservative management fails and her symptoms include locking or giving way, arthroscopic partial meniscectomy may be useful. Patients with substantial impairment and significant degenerative changes on weight-bearing radiographs may be candidates for total knee replacement.

THE BOTTOM LINE

According to our review of the literature and clinical experience, we suggest the medical history and physical examination for patients with possible meniscal or ligamentous lesions outlined in **Box 27-1**. Although there are scant specific data supporting each element of the medical history and physical examination we have outlined, these items are presumed to be part of the composite examination that was found to be useful in determining whether there is a possible meniscal or ligamentous injury. The composite examination for an ACL tear performed by orthopedic physicians is highly predictive (positive LR, 25; 95% CI, 2.1-306; negative LR, 0.04; 95% CI, 0.01-0.50), as is the composite examination for a PCL tear (positive LR, 21; 95% CI, 2.1-205;

Box 27-1 Recommended Basic Medical History and Physical Examination for Patients With Suspected Meniscal or Ligamentous Knee Injuries

HISTORICAL ITEMS

1. Where exactly is the knee pain (point to it with 1 finger)?
2. What is the duration of the pain?
3. Before the pain started, had there been a change in activities?
4. Was there an injury to the lower extremity; if so, what was the direction of the forces?
5. Was there a pop at the injury?
6. Was the knee swollen at the injury or anytime since?
7. Is there giving way or buckling of the knee?
8. Is the knee locking or catching in extension or flexion?
9. Is there pain in the hip, thigh, or back?

PHYSICAL EXAMINATION TESTS

1. Alignment: Are the femur, tibia, and patella in normal alignment during standing and walking?
2. Range of motion: Can the patient actively or passively flex and extend the knee?
3. Effusion: Is there a fluid wave or does ballottement of the patella produce a tapping sensation?
4. Joint line tenderness: Is the patient tender along the medial or lateral joint lines?
5. Lachman test: Is there a discrete end point when the tibia is anteriorly subluxed on the femur?
6. Anterior drawer test: Is there anterior subluxation of the tibia on the femur?
7. Posterior drawer test: Is there posterior sag or translation of the tibia on the femur?
8. Lateral pivot shift: Does the tibia jump anteriorly when extended or flexed with a valgus stress?
9. McMurray test: Is there a popping at the joint line when the knee is extended and rotated?

negative LR, 0.05; 95% CI, 0.01-0.50). The examination for meniscal tears is less efficient; the composite examination confers a positive LR of 2.7 (95% CI, 1.4-5.1) and a negative LR of 0.4 (95% CI, 0.2-0.7). If the medical history and physical examination do not allow the determination of a meniscal or ligamentous injury, consultation with a musculoskeletal specialist may obviate expensive and unnecessary diagnostic imaging.

Author Affiliations at the Time of the Original Publication

Division of Rheumatology, Immunology, and Allergy, Robert B. Brigham Multipurpose Arthritis and Musculoskeletal Diseases Center (Drs Solomon and Katz), Division of Pharmacoepidemiology and Pharmacoeconomics (Dr Solomon), Division of General Medicine, Departments of Medicine (Dr Bates) and Orthopedic Surgery (Dr Schaffer), Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts; and Durham Veterans Affairs Medical Center and Duke University Medical Center, Durham, North Carolina (Dr Simel). Dr Schaffer is now with The Cleveland Clinic, Ohio.

Acknowledgments

This work was supported by a grant from the American Academy of Orthopedic Surgeons. Dr Solomon holds an Investigator Award from the Arthritis Foundation. Dr Katz is supported in part by grants AR36308 and K24 AR02123 from the National Institutes of Health.

The authors thank Peter Blomgren, MD, Laurence Higgins, MD, Charles Brown, MD, and Eugene Oddone, MD, for the thoughtful comments on an earlier version of this article.

REFERENCES

1. Praemer A, Furner S, Rice DP. *Musculoskeletal Conditions in the United States*. Rosemont, IL: American Academy of Orthopedic Surgeons; 1999:27.
2. Cunningham LS, Kelsey JL. Epidemiology of musculoskeletal impairments and associated disability. *Am J Public Health*. 1984;74(6):574-579.
3. Katz JN, Solomon DH, Schaffer JL, Horsky J, Burdick E, Bates DW. Outcomes of care and resource utilization among patients with knee and shoulder disorders treated by general internists, rheumatologists, or orthopedic surgeons. *Am J Med*. 2000;108(1):28-35.
4. Wright V, Helliwell PS. Undergraduate education in musculoskeletal diseases. *Br J Rheumatol*. 1992;31(4):279-280.
5. Craton N, Matheson GO. Training and clinical competency in musculoskeletal medicine: identifying the problem. *Sports Med*. 1993;15(5):328-337.
6. Branch VK, Graves G, Hanczyc M, Lipsky PE. The utility of trained arthritis patient educators in the evaluation and improvement of musculoskeletal examination skills of physicians in training. *Arthritis Care Res*. 1999;12(1):61-69.
7. Lawry GV, Schuldt SS, Kreiter CD, Densen P, Albanese MA. Teaching a screening musculoskeletal examination: a randomized, controlled trial of different instructional methods. *Acad Med*. 1999;74(2):199-201.
8. Friden T, Roberts D, Zatterstrom R, Lindstrand A, Moritz U. Proprioceptive defects after an anterior cruciate ligament rupture—the relation to associated anatomical lesions and subjective knee function. *Knee Surg Sports Traumatol Arthrosc*. 1999;7(4):226-231.
9. Shrive NG, O'Connor JJ, Goodfellow JW. Load-bearing in the knee joint. *Clin Orthop*. 1978;131:279-287.
10. Rath E, Richmond JC. The menisci: basic science and advances in treatment. *Br J Sports Med*. 2000;34(4):252-257.

11. Boden SD, Davis DO, Dina TS, et al. A prospective and blinded investigation of magnetic resonance imaging of the knee: abnormal findings in asymptomatic subjects. *Clin Orthop.* 1992;282:177-185.
12. Gwinn DE, Wilckens JH, McDevitt ER, Ross G, Kao TC. The relative incidence of anterior cruciate ligament injury in men and women at the United States Naval Academy. *Am J Sports Med.* 2000;28(1):98-102.
13. Stevenson H, Webster J, Johnson R, Beynon B. Gender differences in knee injury epidemiology among competitive alpine ski racers. *Iowa Orthop J.* 1998;18:64-66.
14. Bjordal JM, Arnly F, Hannestad B, Strand T. Epidemiology of anterior cruciate ligament injuries in soccer. *Am J Sports Med.* 1997;25(3):341-345.
15. Curl WW, Krome J, Gordon ES, Rushing J, Smith BP, Poehling GG. Cartilage injuries: a review of 31 516 knee arthroscopies. *Arthroscopy.* 1997;13(4):456-460.
16. Marchi AG, Di Bello D, Messi G, Gazzola G. Permanent sequelae in sports injuries: a population based study. *Arch Dis Child.* 1999;81(4):324-328.
17. Feagin JA. *The Crucial Ligaments.* New York, NY: Churchill Livingstone; 1988:1-14.
18. Ike R, O'Rourke KS. Compartment-directed physical examination of the knee can predict articular cartilage abnormalities disclosed by needle arthroscopy. *Arthritis Rheum.* 1995;38(7):917-925.
19. Torg JS, Conrad W, Kalen V. Clinical diagnosis of anterior cruciate ligament instability in the athlete. *Am J Sports Med.* 1976;4(2):84-93.
20. Losee RE. Concepts of the pivot shift. *Clin Orthop.* 1983;172:45-51.
21. Holleman D, Simel D. Does the clinical examination predict airflow limitations? *JAMA.* 1995;273(4):313-319.
22. Sackett DL. A primer on the precision and accuracy of the clinical examination. *JAMA.* 1992;267(19):2638-2644.
23. Simel DL, Samsa GP, Matchar DB. Likelihood ratios with confidence. *J Clin Epidemiol.* 1991;44(8):763-770.
24. Hasselblad V, Hedges LV. Meta-analysis of screening and diagnostic tests. *Psychol Bull.* 1995;117(1):167-178.
25. Eddy DM, Hasselblad V, Schacter R. *Meta-analysis by the Confidence Profile Method.* Burlington, MA: Academic Press; 1992:92-93.
26. Simonsen O, Jensen J, Mouritsen P, Lauritzen J. The accuracy of clinical examination of injury to the knee joint. *Injury.* 1984;16(2):96-101.
27. O'Shea KJ, Murphy KP, Heekin RD, Herzwurm PJ. The diagnostic accuracy of history, physical examination, and radiographs in the evaluation of traumatic knee disorders. *Am J Sports Med.* 1996;24(2):164-167.
28. Rose NE, Gold SM. A comparison of accuracy between clinical examination and magnetic resonance imaging in the diagnosis of meniscal and anterior cruciate ligament tears. *Arthroscopy.* 1996;12(4):398-405.
29. Braunstein EM. Anterior cruciate ligament injuries: a comparison of arthrographic and physical diagnosis. *AJR Am J Roentgenol.* 1982;138(3):423-425.
30. Dahlstedt LJ, Dalen N. Knee laxity in cruciate ligament injury: value of examination under anesthesia. *Acta Orthop Scand.* 1989;60(2):181-184.
31. DeHaven KE. Diagnosis of acute knee injuries with hemarthrosis. *Am J Sports Med.* 1980;8(1):9-14.
32. Donaldson WF, Warren RF, Wickiewicz T. A comparison of acute anterior cruciate ligament examinations: initial vs examination under anesthesia. *Am J Sports Med.* 1985;13(1):5-10.
33. Fowler PJ, Lubliner JA. The predictive value of five clinical signs in the evaluation of meniscal pathology. *Arthroscopy.* 1989;5(3):184-186.
34. Gurtler RA, Stine R, Torg JS. Lachman test evaluated: quantification of a clinical observation. *Clin Orthop.* 1987;(216):141-150.
35. Hardaker WT Jr, Garrett WE Jr, Bassett FH III. Evaluation of acute traumatic hemarthrosis of the knee joint. *South Med J.* 1990;83(6):640-644.
36. Hughston JC, Andrews JR, Cross MJ, Moschi A. Classification of knee ligament instabilities, I: the medial compartment and cruciate ligaments. *J Bone Joint Surg Am.* 1976;58(2):159-172.
37. Jonsson T, Althoff B, Peterson L, Renstrom P. Clinical diagnosis of ruptures of the anterior cruciate ligament: a comparative study of the Lachman test and anterior drawer sign. *Am J Sports Med.* 1982;10(2):100-102.
38. Lee JK, Yao L, Phelps CT, et al. Anterior cruciate ligament tears: MR imaging compared with arthroscopy and clinical tests. *Radiology.* 1988;166(3):861-864.
39. Liu SH, Osti L, Henry M, Bocchi L. The diagnosis of acute complete tears of the anterior cruciate ligament. *J Bone Joint Surg Br.* 1995;77(4):586-588.
40. Mitsou A, Vallianatos P. Clinical diagnosis of ruptures of the anterior cruciate ligament: a comparison between the Lachman test and the anterior drawer sign. *Injury.* 1988;19(6):427-428.
41. Baker CL, Norwood LA, Hughston JC. Acute combined posterior cruciate and posterolateral instability of the knee. *Am J Sports Med.* 1984;12(3):204-208.
42. Loos WC, Fox JM, Blazina ME, Del Pizzo W, Friedman MJ. Acute posterior cruciate ligament injuries. *Am J Sports Med.* 1981;9(2):86-92.
43. Daniel DM. Assessing the limits of knee motion. *Am J Sports Med.* 1991;19(2):139-147.
44. Gillies H, Seligson D. Precision in the diagnosis of meniscal lesions: a comparison of clinical evaluation, arthrography, and arthroscopy. *J Bone Joint Surg Am.* 1979;61(3):343-346.
45. Miller GK. A prospective study comparing the accuracy of the clinical diagnosis of meniscal tear with magnetic resonance imaging and its effect on clinical outcome. *Arthroscopy.* 1996;12(4):406-413.
46. Anderson AF, Lipscomb AB. Clinical diagnosis of meniscal tears. *Am J Sports Med.* 1986;14(4):291-293.
47. Barry OCD, Smith H, McManus F, MacAuley P. Clinical assessment of suspected meniscal tears. *Ir J Med Sci.* 1983;152(4):149-151.
48. Noble J, Erat K. In defense of the meniscus: a prospective study of 200 meniscectomy patients. *J Bone Joint Surg Br.* 1980;62-B(1):7-11.
49. Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *N Engl J Med.* 1978;299(17):926-930.

This page intentionally left blank

Prepared by Daniel H. Solomon, MD, MPH, Jeff Katz, MD,
David Bates, MD, and Jonathan L. Schaffer, MD, MBA

Reviewed by Richard Riedel, MD

CLINICAL SCENARIO

A 55-year-old man presents with 6 months of knee pain. He observes that the pain recently intensified after a weekend of skiing. He denies any specific trauma during his ski trip. He feels increased pain when squatting or walking down stairs. An occasional click has been audible when he is walking.

UPDATED SUMMARY ON THE RATIONAL KNEE EXAMINATION

Original Review

Solomon DH, Bates DW, Katz JN, Simel DL, Schaffer JL. Does this patient have a torn meniscus of the knee? the value of the physical examination in determining whether a patient has a meniscal or ligamentous injury. *JAMA*. 2001;286(13):1610-1620.

UPDATED LITERATURE SEARCH

Our literature search used the parent search strategy for The Rational Clinical Examination series, combined with the subject headings “exp knee,” “exp ligament,” and “exp meniscus,” published in English from 2002 to July 2004. The search yielded 12 articles. We reviewed all the titles and abstracts, identifying 4 articles for additional review. None of these original articles are included in the update. Two did not meet the quality review criteria (examiner blinded to the criterion standard or nonselected patient population) that we originally established. The other 2 did not provide adequate data for combining with the previous studies.

We did find 1 new nonsystematic review that addresses the sensitivity and specificity of some of the key examination maneuvers for meniscal and ligamentous injuries of the knee.¹ The review did not include a methods section for identifying the literature or a methodologic assessment of the referenced articles. Thus, some conclusions from the recently published review may seem clinically sensible but be incorrect because the articles included were not all methodologically rigorous.

CHANGES IN THE REFERENCE STANDARD

The reference standard is the examination of the knee structure of interest (ligament or meniscus) at surgery. However, for patients who do not undergo surgery, the magnetic resonance imaging (MRI) results are the reference standard.

NEW FINDINGS

A nonsystematic review found no additional evidence to alter the following conclusions.

- The Lachman test is the best maneuver for detecting anterior cruciate ligament tears.
- The McMurray test has inadequate sensitivity for ruling out a meniscal tear.

IMPROVEMENTS IN THE DATA PRESENTED IN THE ORIGINAL PUBLICATION

There are no changes in the original data presented on the rational examination for the meniscus and ligaments of the knee. A JAMAInteractive displays the anatomy and maneuvers of the knee examination (<http://jama.ama-assn.org/cgi/content/full/286/13/1610/DC1>; accessed June 1, 2008).

Results of Literature Review

No data suggest new validated examination items for injuries to the meniscus or ligaments of the knee. Symptoms common in patients with meniscal injuries include clicking, locking, and pain. With anterior cruciate ligament injuries, patients have pain and giving way of the knee. A nonsystematic review concluded that the Lachman test is the best test for anterior cruciate ligamentous injuries. The anterior drawer test has been studied more frequently (see Figure 27-2).

Evidence From Guidelines

No government guidelines explicitly address the diagnosis of injuries of the meniscus or ligaments.

CLINICAL SCENARIO—RESOLUTION

The patient should be asked about key symptoms of meniscal and ligamentous injuries, including clicking, locking, and giving way. The examination should include the anterior drawer and Lachman tests for anterior cruciate ligament injuries and at least the medial lateral grind test for meniscal injuries.

Further medical history and evaluation reveal that the patient has pain with squatting and a positive medial lateral grind test result (positive likelihood ratio [LR], 4.8).

Both of these findings, together with negative anterior drawer (negative LR, 0.3) and Lachman test results (negative LR, 0.1), suggest that his injury is likely meniscal and not of the anterior cruciate ligament.

If the patient is not functionally disabled, a trial of anti-inflammatory medicines and physical therapy should be attempted for 8 to 12 weeks. At that point, further evaluation will determine the need for more testing or treatment.

KNEE EXAMINATION—MAKE THE DIAGNOSIS

PRIOR PROBABILITY FOR A LIGAMENTOUS OR MENISCAL TEAR

The physical examination can help in determining which patients are likely to have meniscal or ligamentous injuries of the knee. However, no data exist that allow us to establish reliable prior probability estimates. Among patients with knee pain referred by primary care providers or rheumatologists to an orthopedist, the orthopedist will clinically diagnose meniscal tears in about 25% of patients and ligamentous injuries in about 10%.² We do not know the underlying distribution of these conditions in patients who do not require referral. Because the mechanism of an injury predicts the actual anatomic defect, experts probably can predict (better than chance) the most likely injury when they either observe the trauma or get a reliable medical history.

POPULATION FOR WHOM LIGAMENTOUS OR MENISCAL INJURIES OF THE KNEE SHOULD BE CONSIDERED

Adults with knee pain associated with an injury or with mechanical symptoms, including clicking, catching, locking, or giving way.

DETECTING THE LIKELIHOOD OF A LIGAMENTOUS OR MENISCAL INJURY OF THE KNEE

The best physical examination maneuvers for ligamentous tears or meniscal injuries are shown in Table 27-6. A JAMAInteractive displays the anatomy and some of the maneuvers described in Table 27-6 (<http://jama.ama-asn.org/cgi/content/full/286/13/1610/DC1>; accessed June 1, 2008).

Table 27-6 Physical Examination Maneuvers for Ligamentous and Meniscal Injuries of the Knee

Symptom (No. of Studies)	LR+ (95% CI)	LR- (95% CI)
Anterior Cruciate Ligament Injuries		
Lachman test (1)	42 (2.7-651)	0.1 (0.0-0.4)
Anterior drawer test (3)	3.8 (0.65-22)	0.3 (0.05-1.5)
Meniscal Injuries		
Joint effusion (1)	5.7 (0.4-86)	0.7 (0.5-0.9)
Medial lateral grind test (1)	4.8 (0.8-30)	0.4 (0.2-0.6)
McMurray test (3)	1.3 (0.9-1.7)	0.8 (0.6-1.1)
Joint line tenderness (2)	0.9 (0.8-1.0)	1.1 (1.0-1.3)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

REFERENCE STANDARD TESTS

Serial clinical examinations performed by a specialist are a pragmatic reference standard.

MRI is used to rule in or rule out ligamentous tears. Arthroscopy may be required to rule out meniscal tears.

REFERENCES FOR THE UPDATE

- Malanga GA, Andrus S, Nadler SF, McLean J. Physical examination of the knee: a review of the original test description and scientific validity of common orthopedic tests. *Arch Phys Med Rehabil.* 2003;84(4):592-603.^a
- Solomon DH, Avorn J, Warsi A, et al. Which patients with knee problems are likely to benefit from nonarthroplasty surgery? *Arch Intern Med.* 2004;164(5):509-513.

^aFor the Evidence to Support the Update for this topic, see <http://www.JAMAevidence.com>.

EVIDENCE TO SUPPORT THE UPDATE:

27

Knee Ligaments and Menisci

TITLE Physical Examination of the Knee: Review of the Original Test Description and Scientific Validity of Common Orthopedic Tests.

AUTHORS Malanga GA, Andrus S, Nadler SF, McLean J.

CITATION *Arch Phys Med Rehabil.* 2003;84(4):592-603.

QUESTION How were the common physical examination maneuvers for the knee described originally, and what is their sensitivity and specificity?

DESIGN Qualitative systematic review (ie, systematic review without meta-analysis) of articles that examined physical examination items for injuries of the meniscus and ligaments of the knee. There was no attempt to pool data across studies, and there were no explicit criteria for which studies were included.

DATA SOURCES MEDLINE and bibliographies of all publications included for review and recent review articles.

STUDY SELECTION AND ASSESSMENT There were no clear selection criteria and no formal assessment of the articles included in the review. MEDLINE was searched for articles published between 1970 and 2000. Search terms included each of the involved knee structures and examination maneuvers.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

The authors included studies of the anterior cruciate ligament, posterior cruciate ligament, medial and collateral ligaments, patellofemoral disorders, and the meniscus. Some

studies allowed examiners to conduct the studies under anesthesia. The diagnostic standards were many, including magnetic resonance imaging or arthroscopic findings.

MAIN OUTCOME MEASURES

Sensitivity and specificity.

MAIN RESULTS

For each physical examination maneuver, the authors provide the original description of the examination technique.

CONCLUSION

LEVEL OF EVIDENCE Narrative review.

STRENGTHS This review addresses specific examination maneuvers and includes the original descriptions of common examination maneuvers in detail sometimes not provided in original studies.

LIMITATIONS There was no clear method for selecting the included articles and no attempt was made to pool the results.

The review does not add new information to the current understanding of the physical examination for meniscal or ligamentous injuries of the knee. The sensitivity and specificity values reported are not associated with quality scores, so the clinician cannot confidently apply the results. However, this narrative is useful in providing descriptions for how each examination technique is performed.

Reviewed by Daniel H. Solomon, MD, MPH

This page intentionally left blank

CHAPTER 28

Is This Adult Patient Malnourished?

Allan S. Detsky, MD, PhD

Philip S. Smalley, MD

Jose Chang, MD

CLINICAL SCENARIOS

CASE 1 Ten days before being treated, a 65-year-old man experienced a Wallenberg stroke involving the lateral medulla, which left him with difficulty swallowing. Since then, he had been treated with intravenous fluids, as attempts at eating led to mild aspiration with pneumonia. In that period, he lost 6% of his usual body weight and was continuing to lose weight. He felt weak and was able to ambulate only with difficulty because of his stroke-related ataxia and generalized weakness. On physical examination, there was an obvious squared-off appearance to his shoulders from subcutaneous tissue and muscle wasting. There was no edema.

CASE 2 A 63-year-old man was admitted to the hospital for gastric resection of an obstructing gastric carcinoma. He was well until 6 weeks before admission, when he began to notice the rapid onset of early satiety. This progressed to the point where he began to vomit virtually all food and fluids. He had lost 15% of his body weight and was continuing to lose weight. He was ambulatory but felt weak and was no longer able to carry on his usual daily activities because of this weakness. On physical examination, there was muscle wasting. There was obvious subcutaneous tissue loss in the triceps and thoracic regions, as well as muscle loss in the deltoids. There was edema in his ankles but no ascites.

CASE 3 A 70-year-old man was admitted to the hospital for resection of his descending colon because of an adenocarcinoma detected on investigation for bright-red blood in his bowel movements. Between 6 and 3 months before admission, he had lost 10% of his body weight for reasons that he could not explain. However, his weight had stabilized in the 2 months before admission, and in fact, he had gained back 4% of his weight. His dietary intake had been slightly below normal but had recently improved. He reported no significant gastrointestinal (GI) symptoms other than the bleeding and a mild change in his bowel habits. He had his usual level of energy. On physical examination, there was no evidence of subcutaneous tissue loss, muscle wasting, edema, or ascites.

WHY PERFORM NUTRITIONAL STATUS ASSESSMENT?

Malnutrition occurs among patients either because of their primary diseases (eg, malignancy) or because the procedures they undergo to treat the primary disease prevent them from receiving adequate nutritional intake for prolonged periods (eg, surgery).

There are 2 components of nutritional status assessment. The first is body composition analysis, which is the determination of the mass of body components, such as total body protein and

total body fat. These components are measured by in vivo neutron activation analysis and tritiated water dilution technique, which represents the criterion standard (also known as the gold standard) for measures of body composition. The second component is physiologic function, defined by some as changes in cellular and organ function, measured in a variety of ways, such as skeletal muscle strength, respiratory function, protein synthesis, and tissue repair.

During the past 3 decades, clinicians have become increasingly aware of the prevalence of malnutrition among hospitalized patients.¹⁻⁴ Clinicians have recognized that malnourished patients are at a higher risk of developing complications while undergoing treatment. These complications include death, sepsis, abscess formation, other infections such as pneumonia, wound healing difficulties postoperatively, and respiratory failure. Some have used the term *nutrition-associated complications*^{5,6} to highlight the relationship between malnutrition and these adverse events. The increased risk for malnourished patients is thought to be caused more by functional impairment than changes in body composition,⁷ although in studied subjects there is clearly a correlation between the 2 components of nutritional status.

Investigations in the 1970s^{1,2} estimated that the prevalence of malnutrition among hospitalized patients was as high as 40%. Studies^{4,8} on patients undergoing general GI surgery showed that the prevalence of either mild or severe malnutrition was 48%³ and 31%, respectively. Detsky et al⁴ confirmed the relationship between malnutrition and the risk of nutrition-associated complications. In their series of 202 patients undergoing general GI surgery at 2 Toronto (Ontario, Canada) teaching hospitals, 10% of the total series of patients had major nutrition-associated complications, including 6 deaths related to sepsis, 2 nonfatal episodes of sepsis, 3 subphrenic or intra-abdominal abscesses, 2 anastomotic breakdowns, 2 wound dehiscences, and 5 major wound abscesses. However, among those who were assessed to be severely malnourished preoperatively, this major complication rate was 67%. Windsor and Hill,⁷ using a slightly different system of nutritional status assessment in 102 patients undergoing major GI surgery, also showed that severely malnourished patients had a higher risk of major complications than patients designated as having normal nutritional status. These results confirm the usefulness of nutritional status assessment as a predictor of high risk for postoperative complications. Thus, it becomes both a method of assessing prognosis and a method of diagnosing a particular health state. Furthermore, assessing nutritional status identifies patients who may benefit from enteral or parenteral nutritional repletion to reduce the risk of these complications.⁹⁻¹¹ Although patients with chronic medical conditions also are thought to be at higher risk of developing complications, such as respiratory failure or infection, most of what we know comes from patients undergoing surgical procedures.

THE ANATOMIC/PHYSIOLOGIC ORIGIN OF FINDINGS IN THIS AREA

Syndromes of undernutrition of calories and protein have been studied most extensively in children of developing

nations and are not frequently observed in North America. Two extremes of protein-energy malnutrition have been defined: marasmus, caused primarily by deficiency of calories, resulting in stunted growth in children, loss of adipose tissue, and generalized wasting of lean body mass without edema; and kwashiorkor, a primary deficiency of protein manifested by edema but in which adipose tissue is preserved.

Many individuals who are malnourished will have elements of both protein and calorie deficiencies. The complex metabolic processes that result from protein-energy malnutrition are beyond the scope of this overview. However, in North America, nutritional assessment is used as a predictor of future complications in patients and therefore may go beyond the traditional measurement of pure malnutrition resulting from inadequate intake of protein, calories, or micronutrients. Nutritional assessment, particularly if it encompasses or focuses on physiologic function, may be an overall marker of illness that is not caused solely by inadequate intake or reversed by nutritional supplementation. This may explain why the clinical trials of total parenteral nutrition in a variety of clinical circumstances have in some cases produced disappointing results in improving outcomes.¹²

Nutritional deficiency syndromes involving vitamins and micronutrients evolve through 3 stages because most micronutrients are stored in tissues, and a temporary reduction in intake is buffered by a reduction in body stores. The second stage involves metabolic changes without symptoms, whereas severe depletion will result in the final stage of clinical signs and symptoms. They will not be discussed in this article.

HOW TO PERFORM NUTRITIONAL ASSESSMENT

This article primarily describes features of the medical history and physical examination for assessing overall nutritional status.

The relevant features of a patient's medical history and physical examination can be elicited by a technique known as the subjective global assessment (SGA) of nutritional status.⁸ The application of this technique divides patients into 3 classes: class A, well nourished; class B, moderately (or suspected of being) malnourished; and class C, severely malnourished. The components of this technique are described in Table 28-1. There are 4 elements of the medical history.

1. Weight Loss in the 6 Months Before the Examination, Expressed as a Proportionate Loss From Previous Weight

A weight loss of less than 5% is considered small. A weight loss between 5% and 10% is considered potentially significant, and a weight loss of more than 10% is considered definitely significant. In addition to considering the amount of weight loss, it is important to note the pattern of the weight loss. For example, suppose a patient lost 12% of his or her weight in the 6 months to 1 month before the examination and then regained half of that weight in the subsequent month, resulting in a net loss of 6% for the entire period. This patient would be considered

Table 28-1 Features of Subjective Global Assessment^a

Medical History

1. Weight change

Overall loss in past 6 months: amount = _____ kg; _____ %

Change in past 2 weeks: _____ increase

_____ no change

_____ decrease

2. Dietary intake change (relative to normal)

_____ no change

_____ change duration = _____ weeks

_____ type: _____ suboptimal solid diet _____ full liquid diet

_____ hypocaloric liquids _____ starvation

3. Gastrointestinal symptoms (that persisted for > 2 weeks)

_____ none _____ nausea _____ vomiting _____ diarrhea _____ anorexia

4. Functional capacity

_____ no dysfunction (eg, full capacity)

_____ dysfunction duration = _____ weeks

_____ type: _____ working suboptimally

_____ ambulatory

_____ bedridden

Physical (for each trait specify: 0 = normal, 1+ = mild, 2+ = moderate, 3+ = severe)

_____ loss of subcutaneous fat (triceps, chest)

_____ muscle wasting (quadriceps, deltoids)

_____ ankle edema

_____ sacral edema

_____ ascites

Subjective global assessment rating (select one)^a

_____ A = well nourished

_____ B = moderately (or suspected of being) malnourished

_____ C = severely malnourished

^aClass A indicates individuals with less than 5% weight loss or more than 5% total weight loss but recent gain and improvement in appetite; class B, those with 5%-10% weight loss without recent stabilization or gain, poor dietary intake, and mild (1+) loss of subcutaneous tissue; and class C, ongoing weight loss of more than 10%, with severe subcutaneous tissue loss and muscle wasting, often with edema.

Derived from Detsky et al.⁸

better nourished than a patient who had lost 6% progressively in the 6 months, with continued weight loss in the recent weeks, before the examination. Patients can be considered well nourished despite significant proportions of weight loss if there has been a recent stabilization or increase in weight. In eliciting the history of weight pattern from patients, we recommend asking the patient what his or her maximum weight was and what it was 1 year ago, 6 months ago, 1 month ago, and at present. If patients report substantial weight loss that we cannot confirm with prior records, we ask for confirming history of a change in clothing size or whether their clothes now fit very loosely. Finally, we ask for the pattern of the weight loss during the past few weeks (continued loss, stabilization, or gain).

2. Dietary Intake in Relation to the Patient's Usual Pattern

Patients are classified as having either normal or abnormal (decreased) intake in the weeks to months before the exami-

nation. The duration and degree of abnormality are also noted (eg, starvation, hypocaloric liquids, full liquid diet, or suboptimal solid diet). For example, patients with strokes resulting in swallowing difficulties may have been starved, simply receiving intravenous or hypocaloric fluids for several weeks before the examination. Patients with lesions that obstruct the outflow from the stomach, such as cancer or severe ulcers, may have been receiving pure liquid diets. In eliciting this history, we recommend asking patients whether their eating patterns have changed during the past few weeks and then ask if their pattern has changed during the past few months. Has the amount of food eaten decreased? If so, by how much? Are there certain kinds of foods that they used to eat that they can no longer eat? Why are they eating less (intentional reduction, unintentional reduction, ordered by clinician)? What happens if they try to eat more? Ask for an example of a typical breakfast, lunch, and dinner and a comparison with typical meals 6 to 12 months ago.

3. Presence of Significant Gastrointestinal Symptoms: Anorexia, Nausea, Vomiting, and Diarrhea

By significant we mean that these symptoms must have persisted on virtually a daily basis for a period longer than 2 weeks. Short-term diarrhea or intermittent vomiting is not considered significant. Daily or twice-daily vomiting secondary to obstruction is considered significant.

4. The Patient's Functional Capacity or Energy, Ranging From Full Capacity to Bedridden

Patients who are unable to eat will often complain of fatigue and weakness to the point at which they are bedridden.

There are 3 features of the physical examination that are recorded as normal (0), mild (1+), moderate (2+), or severe (3+).

1. Loss of Subcutaneous Fat

There are several locations where one can look for loss of subcutaneous fat, and the best are the triceps region of the arms, the midaxillary line at the costal margin, the interosseous and palmar areas of the hand, and the deltoid regions of the shoulder (Figures 28-1 and 28-2). Positive findings are loss of fullness or 1 or more areas where the skin fits too loosely over the deeper tissues; this latter sign may be falsely positive in elderly individuals who may

appear to have lost subcutaneous tissue without being clinically malnourished.

2. Muscle Wasting

The best muscles to examine are the quadriceps femoris and deltoids. In the deltoid region, malnourished patients have a squared-off appearance to their shoulders from the combination of muscle and subcutaneous tissue loss (Figure 28-3). In severe malnutrition, the quadriceps will have loss of bulk and tone. Obviously, neurologic lesions (that may present with unilateral wasting) may produce false-positive findings here.

3. Loss of Fluid From the Intravascular to Extravascular Space, Namely, Ankle or Sacral Edema and Ascites

The first 2 signs are best assessed by inspection and then by palpation, remembering that some features are best inspected from a distance, eg, squared-off shoulders. Edema is assessed by pressing the ankle (leg) or sacrum, feeling the fluid move out of the subcutaneous tissue, and then observing “pitting,” persistent depression of the area pressed (more than 5 seconds).

There is no explicit numeric weighting scheme described for combining these features of the history and physical examination into an SGA. Rather, they are combined subjectively into an overall or global assessment. In the study that established the precision and accuracy of SGA,^{4,8} clinicians placed greatest importance on the following variables: weight loss of more than 10%, poor dietary intake, loss of subcutaneous tissue, and muscle wasting. Patients suspected of being malnourished or judged to have moderate malnourishment (class B) had lost at least 5% of their body weight in the weeks before examination without stabilization or weight gain, had a definite history of reduction in dietary intake, and exhibited mild (1+) loss of subcutaneous tissue. When patients had considerable edema, ascites, or tumor mass, less attention was paid to the amount of weight loss. The other historical features helped the clinicians confirm the patient's self-report of weight loss or dietary change but received less weight in the ranking system.

If, on the other hand, a patient had a recent weight gain that did not appear to be merely fluid retention, clinicians designated that patient well nourished (class A), even if the net weight loss was between 5% and 10% and there was mild loss of subcutaneous tissue. The assignment of a class A rank also should occur in settings in which the patient has had an improvement in the other historical features of SGA, such as appetite.

To be classified as severely malnourished (class C), patients should demonstrate obvious physical signs of malnutrition, such as severe (3+) subcutaneous tissue loss and muscle wasting, often with edema, in the presence of a clear and convincing pattern of ongoing weight loss of at least 10%.

By design, this system is less sensitive and more specific. That is, few well-nourished patients will receive a false-positive diagnosis of malnourishment, but some patients with mild degrees of malnutrition may be missed.

Windsor and Hill⁷ describe a slightly different system of nutritional status that focuses more on physiologic function. Their system has 2 components: weight loss and functional status. Preoperative percentage of weight loss is defined as



Figure 28-1 Loss of Subcutaneous Tissue in the Arm and Chest Wall



Figure 28-2 Loss of Subcutaneous Tissue Overlying the Fifth Metacarpal

Hand with tissue loss (left) vs healthy hand (right).

(recalled well weight minus current measured weight) divided by well weight. A weight loss of more than 10% during the preceding 3 months was considered significant. Confirmation of weight loss is sought in the physical examination by palpating skin folds for loss of fat and muscles in a manner similar to that just described, functional impairment of overall activity levels (by observing the patient on the ward), overall mood (alertness, ability to concentrate, and irritability), skeletal muscle function (having the patient squeeze the examiner's hand), respiratory function (effort and sound of coughing and shortness of breath), wound healing (unhealed wounds and sores or scratches or skin sepsis), and serum albumin level of less than 3.2 g/dL. If patients have weight loss of less than 10%, with no evidence of abnormal physiologic function, then they are placed in group 1. With weight loss of more than 10% but no abnormal physiologic function, patients are placed in group 2, and with both features, they are placed in group 3.

READER PARTICIPATION

Before you read further, we suggest that you return to the patient scenarios that opened this overview and decide whether you judge them to be well nourished, moderately malnourished, or severely malnourished using SGA. After doing so, read on.

The patient in case 1 was moderately malnourished (class B). This ranking was determined by his continuing loss of weight, the limitation of nutritional intake to hypocaloric fluids for 2 weeks, and the mild loss of subcutaneous tissue and muscle.

The patient in case 2 was severely malnourished (class C). This judgment was most influenced by his continuing large weight loss, change in dietary intake, and positive physical findings.

The patient in case 3 was well nourished (class A). Although he had experienced considerable weight loss at some time before admission, his weight had stabilized and increased just before admission.

PRECISION OF THE ASSESSMENT OF NUTRITIONAL STATUS

Investigators at the University of Toronto studied 202 patients at 2 teaching hospitals who were undergoing major GI surgery.⁸ A nurse and 3 residents learned the technique of nutritional status described herein by examining a series of patients and reviewing their assessments with those of a senior clinician. The emphasis was on combining the symptoms and signs of malnutrition to minimize the false-positive diagnosis of malnutrition (high specificity) at the expense of increasing false-negative results (lower sensitivity). After reviewing several patients together, the nurse and one of the 3 residents performed duplicate, independent assessments of 109 patients. There was perfect agreement in 100 (91%) of 109 patients on the SGA rankings. This was 78% above the agreement that could be expected by chance alone (the κ statistic was 0.78, with SE = 0.08 and 95% confidence interval ranging from 0.62 to 0.94). The κ statistics for the 3 pairings

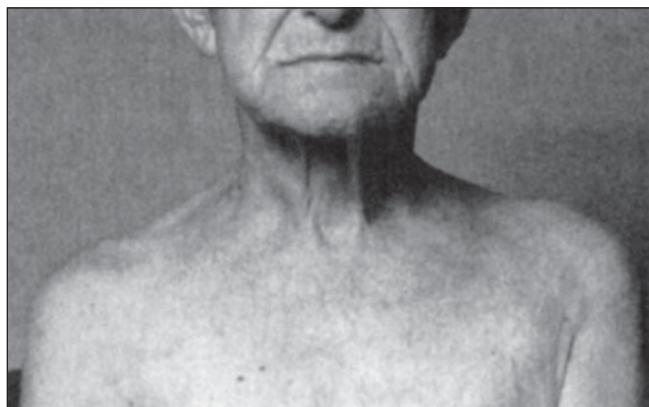


Figure 28-3 Loss of Subcutaneous Tissue in the Shoulders, Giving a Squared-off Appearance

of the nurse with the individual residents were 0.60, 0.81, and 1.0, respectively, revealing some variation in agreement between different clinicians. Hirsch et al¹³ also documented 79% concordance between SGA rankings of residents and specialists in clinical nutrition.

ACCURACY OF NUTRITIONAL ASSESSMENT

Because there is no criterion standard for the diagnosis of malnutrition that incorporates body composition and physiologic function (the in vivo neutron activation analysis and titrated water technique are the criterion standards of body composition alone), studies of the accuracy of techniques of nutritional status assessment have related it to the development of complications judged to result from malnutrition. Therefore, patients are sorted into the columns of the usual 2×2 table based on whether they develop malnutrition-associated complications.

The study by Detsky et al⁴ provides useful data on the accuracy of SGA (Table 28-2). Nineteen patients (10% of the total studied) were classified as severely malnourished (class C), 44 (21%) were classified as moderately (or suspected of being) malnourished (class B), and 139 (69%) were classified as well nourished (class A). The likelihood ratios in this table show that the SGA is a powerful predictor of postoperative complications. The likelihood ratio greater than 4 for severely malnourished patients means that this designation (class C) was more than 4 times as likely to be found in patients with, as opposed to patients without, postoperative complications. Patients designated as moderately (or suspected of being) malnourished (class B) generated a likelihood ratio of close to unity, indicating no clinically important change between the preexamination and postexamination probability of postoperative complications (20/202, or 10%). Finally, well-nourished patients (class A) generated likelihood ratios of 0.66 for their admission SGA and only 0.38 for their minimum SGA, indicating a lower than average risk of postoperative complications.

SGA performed better than objective measurements of the physical examination, such as percentage of ideal weight on admission and percentage of body fat calculated from

Table 28-2 Relationship Between Subjective Global Assessment (SGA) and Major Postoperative Complications^a

SGA Class	Patients Assigned Class on Admission, No. (%)	Major Complications ^b Occurring in This Class, No. (%)	Likelihood Ratio for Admission SGA	Likelihood Ratio for Minimum SGA During Hospitalization
Severely malnourished	19 (10) ^b	8 (42)	4.4	4.1
Moderately (or suspected of being) malnourished	44 (21)	4 (9)	0.96	0.93
Well nourished	139 (69)	8 (5)	0.66	0.38

^aDerived from Detsky et al.⁴^bOf the 20 complications, there were 6 deaths related to sepsis, 2 nonfatal episodes of sepsis, 3 subphrenic or intra-abdominal abscesses, 2 anastomotic breakdowns, 2 wound dehiscences, and 5 major wound infections (abscesses).**Table 28-3** Predictive Properties of Unpromising Techniques^a

Likelihood Ratio	
Ideal Weight on Admission, %	
≤79	1.5
80-99	0.62
≥100	1.2
Admission Body Fat, %	
≤9	1.0
10-14	0.83
>20	0.99

^aDerived from Detsky et al.⁴

anthropometric measurements. The range of likelihood ratios for these variables displayed considerably less accuracy than those associated with SGA and the combination of SGA (Table 28-3).

Laboratory determination of serum albumin level was also shown to be an accurate predictor of complications, associated with a progression of likelihood ratios that is similar to that of SGA. Moreover, the combination of SGA and albumin provided slightly improved accuracy compared with either method alone. However, other objective methods that are frequently said to be useful techniques of assessing nutritional status (serum transferrin level, creatinine-height index, and total lymphocyte count) were not shown to be accurate predictors of complications.⁴

The study by Windsor and Hill⁷ provides similar data demonstrating the predictive validity of their system. Of the 102 patients, 43 (42%) were in group 1 (analogous to SGA class A), 17 (17%) were in group 2 (analogous to SGA class B), and 42 (41%) were in group 3 (analogous to SGA class C). The rate of major complications, septic complications, and pneumonia in the 3 groups was significantly different, and the likelihood ratios for predicting major complications showed a similar progression to SGA of 0.53, 0.69, and 1.8 for groups 1, 2, and 3, respectively.

Finally, the predictive validity of SGA was also reported by the Veterans Affairs perioperative total parenteral nutrition randomized trial¹¹ that enrolled only patients with various

degrees of malnutrition. Among the control patients, those in SGA class C had higher rates of major infectious complications and noninfectious complications.

Some have also reported the high correlation between SGA and other measures of nutritional status assessments that are thought to be more objective, such as anthropometry,^{3,13} albumin level,^{3,13} total serum protein level,^{3,13} and criterion standard measures of body composition. Windsor and Hill⁷ also show good correlations between their system and anthropometry, body composition, and objective measures of the physiologic functions in their method (eg, grip strength and respiratory muscle index).

ARE THESE SYMPTOMS OR SIGNS EVER NORMAL?

Many individuals are thin, and this in itself does not constitute malnutrition. However, we should note that obesity, defined as an excess of adipose tissue or by the degree to which a patient's weight exceeds that which is judged ideal by some anthropometric formula, is also a common problem in hospitalized patients. Epidemiologic studies have shown that a 20% excess over ideal weight imparts a health risk. Similarly, obesity has been shown to place patients at a high risk of experiencing surgical complications, such as poor wound healing and venous thrombosis.

SPECIAL WAYS TO LEARN, TEST YOURSELF, AND CORRECT DEFICIENCIES IN THE ELICITATION OF THESE SYMPTOMS AND SIGNS

Clinicians who wish to become competent at nutritional assessment can do so by applying the following strategies: First, they should undergo a training period with other learners, in which they discuss each of the features of the technique together and review a series of patients for each of the findings. In particular, the group should review methods of eliciting the medical history, performing the inspection, and standardizing terms such as *normal*, *mild*, *moderate*, and *severe*. Next, they should rank several patients together and reach consensus about what constitutes an A, B, or C ranking. Finally, they should perform their own tests of clinical reproducibility by treating a series of (perhaps 10) patients

independently and comparing their rankings. To improve the precision and validity of their elicitation of the individual features of the SGA, they should consider verification strategies, such as asking the patient's spouse about the features of the history, examining physician records for previous weights, asking whether the patient's clothes now fit loosely, and examining recent and old pictures of the patient.

THE BOTTOM LINE

Clinicians can learn to perform SGA of nutritional status with precision. The features of the medical history and physical examination are shown in Table 28-1. We recommend the group approach to standardize the definitions of the features of the history and physical examination contained in SGA and to gain competency in their application. In doing so, we recommend that clinicians train themselves to be less sensitive and more specific in labeling patients as malnourished. Because there is no criterion standard for malnutrition that incorporates body composition and physiologic function, this clinical skill should be used as a prognostic instrument to identify patients who are at high risk of developing complications and who may benefit from nutritional repletion and support. The technique is an accurate predictor of patients who are at higher risk of developing complications such as infection or poor wound healing.

Author Affiliations at the Time of the Original Publication

From the Department of Medicine (Drs Detsky and Smalley) University of Toronto, Toronto, Ontario, Canada; Division of General Internal Medicine and Clinical Epidemiology, The Toronto Hospital, Toronto, Ontario, Canada (Dr Detsky); Department of Medicine, Oshawa General Hospital, Oshawa, Ontario, Canada (Dr Chang); and Division of General Internal Medicine, The Wellesley Hospital, Toronto, Ontario, Canada (Dr Smalley).

Acknowledgments

This study was supported in part by a National Health Research Scholar Award from Health and Welfare Canada, No. 6606-2849-48, to Dr Detsky.

REFERENCES

- Bistrian BR, Blackburn GL, Hallowell E, et al. Protein status of general surgical patients. *JAMA*. 1974;230(6):858-860.
- Bistrian BR, Blackburn GL, Vitale J, et al. Prevalence of malnutrition in general medical patients. *JAMA*. 1976;235(15):1567-1570.
- Baker JP, Detsky AS, Wesson DE, et al. Nutritional assessment: a comparison of clinical judgment and objective measurements. *N Engl J Med*. 1982;306(16):969-972.
- Detsky AS, Baker JP, O'Rourke K, et al. Predicting nutrition-associated complications for patients undergoing gastrointestinal surgery. *JPEN J Parenter Enteral Nutr*. 1987;11(5):440-446.
- Detsky AS, Baker JP, Mendelson RA, Wolman SL, Wesson DE, Jeejeebhoy KN. Evaluating the accuracy of nutritional assessment techniques applied to hospitalized patients: methodology and comparisons. *JPEN J Parenter Enteral Nutr*. 1984;8(2):153-159.
- Detsky AS, Mendelson RA, Baker JP, Jeejeebhoy KN. The choice to treat all, some, or no patients undergoing gastrointestinal surgery with nutritional support: a decision analysis approach. *JPEN J Parenter Enteral Nutr*. 1984;8(3):245-253.
- Windsor JA, Hill GL. Weight loss with physiologic impairment: a basic indicator of surgical risk. *Ann Surg*. 1988;207(3):290-296.
- Detsky AS, McLaughlin JR, Baker JP, et al. What is subjective global assessment of nutritional status? *JPEN J Parenter Enteral Nutr*. 1987;11(1):8-13.
- Detsky AS, Baker JP, O'Rourke K, Goel V. Perioperative parenteral nutrition: a meta-analysis. *Ann Intern Med*. 1987;107(2):195-203.
- Buzby GP, Williford WO, Peterson OL, et al. A randomized clinical trial of total parenteral nutrition in malnourished surgical patients: the rationale and impact of previous clinical trials and pilot study on protocol design. *Am J Clin Nutr*. 1988;47(2 suppl):357-365.
- Veterans Affairs Total Parenteral Nutrition Cooperative Study Group. Perioperative nutrition in surgical patients. *N Engl J Med*. 1991;325:525-532.
- Detsky AS. Parenteral nutrition: is it helpful? *N Engl J Med*. 1991;325(8):573-575.
- Hirsch S, de Obaldia N, Petermann M, et al. Subjective global assessment of nutritional status: further validation. *Nutrition*. 1991;7(1):35-37.

This page intentionally left blank

Prepared by David L. Simel, MD, MHS
Reviewed by Alan Detsky, MD, PhD, and Amy Rosenthal, MD

CLINICAL SCENARIO

A 68-year-old man with advanced emphysematous lung changes has been hospitalized 3 times during the past winter for episodes of dyspnea. His diet is not as good as usual, but his weight has not changed during the past 2 months. Because of his breathing difficulty, he spends much of his day in and out of a reclining chair. You notice that his arms are a bit thin, with loss of muscle. There is some peripheral edema. Overall, his weight is down about 2 kg from what he considers his baseline (a 3% loss). During his last hospitalization, a serum albumin level was 3.4 g/dL, and you see that his total lymphocyte count was 1525 cells/ μ L. Is he appropriately nourished?

UPDATED SUMMARY ON MALNUTRITION IN ADULTS

Original Review

Detsky AS, Smalley PS, Chang J. The rational clinical examination: is this patient malnourished? *JAMA*. 1994;271(1):54-58.

UPDATED LITERATURE SEARCH

Our literature search used the parent search strategy for The Rational Clinical Examination series, combining the subject headings "malnutrition/di," "protein-energy malnutrition/di," and "nutritional disorders/di," published in English from 1993 to September 2004. The focus was on macronutrient rather than micronutrient deficiency (vitamins and minerals). The search yielded 96 titles for review, of which 39 articles appeared to have promising abstracts. Two nonsystematic reviews on malnutrition in the elderly helped us focus on simpler nutritional screening assessments, performed by physicians. We reviewed the reference lists from these 2 reviews.^{1,2}

We reviewed studies of adults with more than 100 study subjects. We were interested only in original studies that prospectively assessed adult malnutrition compared with an appropriate reference standard and that contained data allowing us to estimate the sensitivity and specificity of clinical symptoms, signs, or screening instruments. In addition, we focused on screening instruments that are simple and require little additional training. We retained only 3 articles

for detailed review. We used a qualitative approach to sum up the main features of the other identified studies and non-systematic reviews.³

NEW FINDINGS

Details of the Update

The majority of studies on adult malnutrition include either the elderly subject (healthy, hospitalized, or institutionalized) or patients with malignancy. Some of the specific screening instruments for the elderly lack generalizability to other populations because they include questions concerning dementia and deficits in the activities of daily living that will be less of a concern in younger patients.

Virtually all screening instruments emphasize the importance of quantifying weight loss and assessing changes in appetite. In general, a change in weight of 5% is small, whereas a change more than 10% is definitely significant. However, clinical judgment is still required and can be highlighted by 2 simple examples: (1) a patient undertaking a diet may have more than 10% weight loss and not be malnourished, or (2) a patient with cirrhotic ascites may be severely malnourished, despite a stable weight, when extracellular fluid replaces weight lost from decreasing muscle mass.

Incorporation bias affected many of the newer studies of adult malnutrition because the results of the screening tests were also part of the reference standard. This seems inevitable in nutritional research because the reference standard requires the combination of objective findings (medical history, anthropometric measures, and biochemical measures) and clinical impression. In retaining articles for specific review, we identified those that seemed less affected by the bias. For example, one study compared a discriminant analysis equation using quantitative variables. Although the variables might have been available for some patients, it seemed unlikely that the total score from the equation would have been readily available.⁴

IMPROVEMENTS IN THE DATA PRESENTED IN THE ORIGINAL PUBLICATION

There are no changes in the performance characteristics of the recommended subjective global assessment (SGA) for adult

malnutrition. An additional study of observer variability for the SGA in a different patient population (women with gynecologic malignancies) found a weighted κ of 0.80 (95% confidence interval [CI], 0.67-0.92),⁵ almost identical to that reported in the original publication ($\kappa = 0.78$). This provides us with a high degree of confidence in the reliability of the SGA.

CHANGES IN THE REFERENCE STANDARD

No single test serves adequately as a reference standard for malnutrition. The assessment of adult malnutrition requires a com-

bination of the patient medical history, physical examination results, biochemical and anthropometric measures, and an expert's opinion. From a pragmatic viewpoint and from the viewpoint of a clinical investigator, most physicians would accept the opinion of an expert (eg, a clinical dietitian or a physician with expertise in nutritional assessment) who used these variables as part of their assessment. The reference standard for determining malnutrition goes beyond identifying the patient with current protein-energy deficiency. A more relevant issue is identifying the patient at risk for nutrition-related complications.

In the original Rational Clinical Examination article on malnutrition, the SGA was proposed as the best screening measure (Table 28-1). Since publication of that article, newer studies have used the SGA itself as the reference standard in an attempt to find other screening approaches that require less expertise, fewer variables, or less time. This is a reasonable approach in that the SGA has been validated for its reliability⁶ ($\kappa = 0.71$ or higher) and accuracy in predicting outcomes from malnutrition.

Table 28-4 Likelihood Ratio of a Low Albumin Level for Malnutrition

Finding	Reference Standard	LR+ (95% CI)	LR- (95% CI)
Serum albumin < 3.0 g/dL	SGA	3.3 (1.6-6.9)	0.88 (0.79-0.95)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio; SGA, subjective global assessment.

Table 28-5 Multivariate Findings for Adult Malnutrition

Malnutrition Screening Tool ^{12,13}	Item Score
1. Have you lost weight without trying?	
No	0
Unsure	2
Yes	Use question 2 instead
2. If 1 is yes, use the question, How much weight (kg) have you lost?	
None	0
1-5	1
6-10	2
11-15	3
>15	4
Unsure	2
3. Have you been eating poorly because of a decreased appetite?	
No	0
Yes	1
Malnutrition screening score	Sum of above

Table 28-6 Likelihood Ratios of Combinations of Findings for Malnutrition

Combination of Findings	Reference Standard	LR (Factor Present)	LR (Factor Absent)
Malnutrition screening tool (score ≥ 2) ^{12,13} (2 studies)	SGA	13 (2.9-61)	0.27 (0.19-0.39)
LAW criteria (discriminant function using lymphocyte count, albumin, percentage weight loss) (1 study) ⁴	Expert assessment by a dietitian	6.1 (4.0-9.6)	0.10 (0.03-0.25)

Abbreviations: LAW criteria, lymphocyte count, albumin, percentage weight loss; LR, likelihood ratio; SGA, subjective global assessment.

RESULTS OF LITERATURE REVIEW

The SGA was independently compared to a single biochemical measure, the serum albumin level, among a population of hospitalized, older, general medical patients.⁷ A serum albumin level less than 3.0 g/dL increases the likelihood of moderate or severe malnutrition (likelihood ratio [LR], 3.3; 95% CI, 1.6-6.9), although other conditions could be associated with hypoalbuminemia (Table 28-4). However, using a value as extreme as 3.0 g/dL will miss many patients, and the area under the receiver operating characteristic curve for albumin is only 0.58. These data support the continued use of the SGA for assessing patients.

A patient-generated subjective global assessment (PG-SGA) has been developed.⁶ This modification assigns values to explicit items on the physical examination, underlying conditions, metabolic stress, and amount of weight loss. The hope for the PG-SGA was that less-experienced observers might be able to use it because the scores are explicit for the various items. Although the accuracy was high compared with the SGA, the PG-SGA requires an independent comparison to the SGA and assessment of its interobserver variability. Given the large number of items on the PG-SGA vs the SGA, it may have lower interobserver variability.

Two shorter instruments have been developed and compared to the SGA, using the SGA as the reference standard (Tables 28-5 and 28-6). The Malnutrition Screening Tool^{12,13} shortens the SGA to the information collected in its first 3 questions. Summary likelihood ratios (Table 28-6) suggest that it performs well. A second approach creates a score from a discriminant model that combines the percentage weight loss with the serum albumin and the total lymphocyte count.⁴

For elderly patients, the Mini Nutritional Assessment (MNA) has been validated in a variety of ways and compared to physicians with expertise in clinical nutrition as the reference standard, along with dietary changes, anthropometry, and biochemical measures.⁸ The MNA has been applied to healthy, hospitalized, housebound, and institutionalized elderly patients. It requires about 10 minutes for an expert to complete the ques-

tionnaire. The screen shows excellent reliability, with a κ of 0.78 at a cut point of 18.⁹ The items in the questionnaire limit the applicability to elderly patients. Before it can be accepted as a reference standard for elderly patients, additional work needs to be done. One study, using an independent, blind application of the MNA to a clinical expert, found it to be only 62% accurate.¹⁰ Clinicians who care primarily for a geriatric population may find useful a compilation of review articles from a symposium on the MNA in the elderly.¹¹

LAW (Lymphocyte Count, Albumin, Percentage Weight Loss), Discriminant Model⁴

$$\begin{aligned} & 0.07242 \times (\text{total lymphocyte count, } \mu\text{L}) \\ + & 238.664 \times (\text{albumin, g/dL}) \\ - & 24.657 (\% \text{ weight change, expressed as } 15\% = \\ & 15 \text{ rather than } 0.15) \\ = & \text{score } \underline{\hspace{2cm}} \end{aligned}$$

Score < 747.2 = positive for malnutrition

Score > 747.2 = negative for malnutrition

EVIDENCE FROM GUIDELINES

No government guidelines address a preferred screen for the nutritional assessment of adults. The Joint Commission

requires nutritional assessment, when warranted by the patient's condition, in all health care settings.

CLINICAL SCENARIO—RESOLUTION

A major goal of nutritional assessment in adults is not only diagnosing current protein-energy deficiency but also identifying the patient at risk of nutrition-associated complications.

You have the data to use the LAW (lymphocyte count, albumin, percentage weight loss) criteria, but the discriminant function gives you a value of 848 and does not indicate moderate or at-risk malnutrition. The single value of albumin does not change the likelihood of malnutrition much because an albumin level more than 3.0 g/dL has an LR of only 0.88. These results expose the fallacy of relying too much on biochemical measures. The patient has lost weight attributable to a change in his appetite (malnutrition screening score of 2), which puts him at risk for moderate malnutrition. In addition, your physical examination results suggest the loss of muscle mass that could be quantified through caliper measurement of his triceps skinfold thickness. It is appropriate to use the items of the SGA that factor in his weight loss, change in diet, loss of functional capacity, and loss of subcutaneous fat in the triceps that together put him in a category of suspected moderate malnutrition.

ADULT MALNUTRITION—MAKE THE DIAGNOSIS

PRIOR PROBABILITY

The prior probability for adult malnutrition has a broad range. Among hospitalized medical or surgical patients, the prevalence is 10% to 40%. The prevalence among healthy patients, by definition, will be much lower.

POPULATION FOR WHOM ADULT MALNUTRITION SHOULD BE CONSIDERED

- Disorders, conditions, or treatments affecting appetite
- Malignancy
- Psychiatric illness
- Gastrointestinal tract illness
- Conditions requiring a change to a suboptimal solid diet (eg, liquid diets, tube diets)
- Disorders affecting metabolism
- Elderly patients
- Patients with unintentional weight loss of more than 5%, a major category of individuals for whom additional testing is warranted

IDENTIFYING THE MALNOURISHED ADULT

Determine whether the patient has lost weight, the amount of weight loss, and his or her appetite to get a malnutrition score (Table 28-7).

Table 28-7 Detecting the Likelihood of Adult Malnutrition

	LR+ (95% CI)	LR- (95% CI)
Malnutrition screening tool ^a (score ≥ 2)	13 (2.9-61)	0.27 (0.19-0.39)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

^aSee Table 28-5 for components of the malnutrition screening tool.

REFERENCE STANDARD TESTS

- Expert evaluation (dietitian or physician trained in nutritional care and assessment) using a combination of historical features, anthropometry, weight change, and biochemical measures.
- SGA by a trained clinician for identifying patients at risk of complications related to malnutrition.

REFERENCES FOR THE UPDATE

1. Omran ML, Morley JE. Assessment of protein energy malnutrition in older persons, part I: history, examination, body composition, and screening tools. *Nutrition*. 2000;16(1):50-63.
2. Omran ML, Salem P. Diagnosing undernutrition. *Clin Geriatr Med*. 2002;18(4):719-736.
3. Slavin RE. Best evidence synthesis: an intelligent alternative to meta-analysis. *J Clin Epidemiol*. 1995;48(1):9-18.
4. Elmore MF, Wagner DR, Knoll DM, et al. Developing an effective adult nutrition screening tool for a community hospital. *J Am Diet Assoc*. 1994;94(10):1113-1118.^a
5. Santoso JT, Cannada T, O'Farrell B, Alladi K, Coleman RL. Subjective versus objective nutritional assessment study in women with gynecological cancer: a prospective cohort trial. *Int J Gynecol Cancer*. 2004;14(2): 220-223.
6. Bauer J, Capra S, Ferguson M. Use of the scored patient-generated subjective global assessment (PG-SGA) as a nutrition assessment tool in patients with cancer. *Eur J Clin Nutr*. 2002;56(8):779-785.
7. Covinsky K, Covinsky M, Palmer RM, Sehgal A. Serum albumin concentrations and clinical assessments of nutritional status in hospitalized older people: different sides of different coins? *J Am Geriatr Soc*. 2002; 50(4):631-637.
8. Vellas B, Guigoz Y, Garry PJ, et al. The Mini Nutritional Assessment (MNA) and its use in grading the nutritional state of elderly patients. *Nutrition*. 1999;15(2):116-122.
9. Bleda MJ, Bolíbar I, Parés R, Salvà A. Reliability of the Mini Nutritional Assessment (MNA) in institutionalized elderly people. *J Nutr Health Aging*. 2002;6(2):134-137.
10. Azad N, Murphy J, Amos S, Toppan J. Nutrition survey in an elderly population following admission to a tertiary care hospital. *CMAJ*. 1999;161(5):511-515.
11. Vellas B, Garry PJ, Guigoz Y, eds. *Mini Nutritional Assessment (MNA): Research and practice the elderly*, Nestlé Nutrition Workshop Series Clinical & Performance Programme. Vol 1, 1999.
12. Ferguson M, Capra S, Bauer J, Banks M. Development of a valid and reliable malnutrition screening tool for adult acute hospital patients. *Nutrition*. 1999;15(6):458-464.^a
13. Ferguson ML, Bauer J, Gallagher B, Capra S, Christie DRH, Mason BR. Validation of a malnutrition screening tool for patients receiving radiotherapy. *Australas Radiol*. 1999;43(3):325-327.^a

^aFor the Evidence to Support the Update for this topic,
see <http://www.JAMAevidence.com>.

EVIDENCE TO SUPPORT THE UPDATE:

28

Malnourishment, Adult

TITLE Developing an Effective Adult Nutrition Screening Tool for a Community Hospital.

AUTHORS Elmore MF, Wagner DR, Knoll D, et al.

CITATION *J Am Diet Assoc.* 1994;94(10):1113-1118.

QUESTION Do 3 variables, identified through discriminant analysis, predict malnutrition?

DESIGN A 3-variable discriminant model (nutrition screening equation [NSEq]) was developed in one hospital and then tested prospectively in a second hospital.

SETTING Community hospital.

PATIENTS Randomly selected patients ($n = 151$) from a different hospital where the discriminant model was developed.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

The serum albumin level and total lymphocyte count were obtained from the first scheduled blood draw for the patient after admission. The percentage of weight loss was by self-report of the patient. The reference standard was a full nutritional assessment that incorporated a history, review of systems, current status of the patient, and biochemical and anthropometric measures by a trained dietitian. For some patients, variables included in the model might have been available to the clinician. However, the full nutritional assessment and screening tests were applied independently to develop the model. The patients were assigned to levels of not at nutritional risk vs at risk.

MAIN OUTCOME MEASURE

Comparison of the discriminant model to the reference standard diagnosis of malnutrition.

Discriminant model:

$$\begin{aligned} & 238.664 \times (\text{albumin, g/dL}) \\ + & 0.07242 \times (\text{total lymphocyte count, mm}^3) \\ - & 24.657 (\% \text{ weight change, expressed as } 15\% = 15 \\ & \text{rather than } 0.15) \\ = & \text{score} \\ \text{Score } < 747.2 & = \text{positive for malnutrition} \\ \text{Score } > 747.2 & = \text{negative for malnutrition} \end{aligned}$$

MAIN RESULTS

The Nutrition Screening Equation (Table 28-8) can be remembered from the acronym LAW (lymphocyte count, albumin, percentage weight loss).

Table 28-8 Likelihood Ratios for the Nutrition Screening Question for Malnutrition

Test	LR+	LR-	DOR (95% CI)
NSEq	6.1 (4.0-9.6)	0.10 (0.03-0.25)	64 (18-220)

Abbreviations: CI, confidence interval; DOR, diagnostic odds ratio; LR+, positive likelihood ratio; LR-, negative likelihood ratio; NSEq, nutrition screening equation.

CONCLUSION

LEVEL OF EVIDENCE Level 4.

STRENGTHS Score was developed in one setting and then validated in another.

LIMITATIONS The full nutritional assessment by the expert clinician included biochemical values.

The addition of the albumin to the subjective global assessment (SGA) studied by Detsky et al¹ found that the addition of the serum albumin to the SGA provided additional information that was better than the SGA alone or the albumin level alone. The addition of the total lymphocyte count was not useful when added to the SGA. The investigators used a reference standard for malnutrition that most primary care clinicians would accept. Although the investigators did not use the actual SGA, they used

a multimodal approach that incorporated the medical history, clinical evaluation, and biochemical and anthropometric measures. We could not separate out the relative contribution of percentage of weight loss vs the laboratory parameters.

The actual discriminant model used variables previously shown as important in combination and given as the LAW criteria.²

In the presence of incorporation bias (ie, the biochemical and weight loss characteristics were used as part of the reference standard), what is the value of this information? As in the studies by Ferguson et al,^{3,4} these results tell us more about how the expert clinicians incorporated these characteristics into their assessment than the independent value of these variables. Some clinicians, especially those less versed in assessment of malnutrition, might choose to use these results to justify obtaining a serum albumin level and total lymphocyte count when they are considering the presence of malnutrition in a patient with weight loss. However, it can be easily inferred from the model that intentional weight loss could lead to false-positive model results. For that reason, the clinical variables in the Malnutrition Screening Tool (MST) developed by Ferguson et al^{3,4} make more sense. Indeed, the higher diagnostic odds ratio of the MST suggests a greater accuracy and supports the need for clinically assessing the context of the patient's weight loss.

REFERENCES FOR THE EVIDENCE

- Detsky AS, Smalley PS, Chang J. Is this patient malnourished? *JAMA*. 1994;271(1):54-58.
- Omran ML, Morely JE. Assessment of protein energy malnutrition in older persons, part I: history, examination, body composition, and screening tools. *Nutrition*. 2000;16(1):50-63.
- Ferguson M, Capra S, Bauer J, Banks M. Development of a valid and reliable malnutrition screening tool for adult acute hospital patients. *Nutrition*. 1999;15(6):458-464.
- Ferguson ML, Bauer J, Gallagher B, Capra S, Christie DRH, Mason BR. Validation of a malnutrition screening tool for patients receiving radiotherapy. *Australas Radiol*. 1999;43(3):325-327.

Reviewed by David L. Simel, MD, MHS

TITLE Development of a Valid and Reliable Malnutrition Screening Tool for Adult Acute Hospital Patients.

AUTHORS Ferguson M, Capra S, Bauer J, Banks M.

CITATION *Nutrition*. 1999;15(6):458-464.

QUESTION Can a brief screening tool be developed that has high accuracy compared with the subjective global assessment (SGA)?

DESIGN Prospective convenience sample. A variety of questions served as candidate variables for a model that predicts the SGA. The SGA was not performed independently of the potential screening questions.

SETTING Inpatients at a Brisbane, Australia, hospital.

PATIENTS Adult, newly admitted, general medical patients (n = 408).

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

A multitude of individual questions were asked of the patient. The patients self-reported their heights and weights. An expert dietitian performed the SGA.¹ Once the model was developed, the interrater reliability of the model was tested by 2 of 3 observers in a prospective fashion. In addition, the results of the reduced model were compared with anthropometric and biochemical variables, along with hospital length of stay. The Malnutrition Screening Tool (MST) was compared with the SGA that was categorized into well-nourished vs moderately or severely malnourished.

MAIN OUTCOME MEASURES

Diagnostic accuracy of the MST compared with the SGA; 17% of the patients were moderately or severely malnourished.

MAIN RESULTS—UNIVARIATE

The simple questions performed well when interpreted in isolation (Table 28-9).

Table 28-9 Likelihood Ratios for Simple Questions Compared With the Subjective Global Assessment Tool^a

Questions	Sensitivity	Specificity	LR+	LR-	DOR
Are you eating poorly because of a decreased appetite?	0.87	0.93	12	0.14	89
Has your appetite/food intake been less than usual lately?	0.83	0.90	8.3	0.19	44
Have you lost weight recently without trying?	0.98	0.83	5.8	0.02	239

Abbreviations: DOR, diagnostic odds ratio; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

^aConfidence intervals not provided. The likelihood ratio represented calculated values from the sensitivity and specificity. Diagnostic odds ratios are all highly significantly different from 1, with P < .001.

When taken together, the simple questions form the MST (Table 28-10).

Table 28-10 Malnutrition Screening Tool

	Item Score
1. Have you lost weight without trying?	
No	0
Unsure	2
Yes	Use question 2 instead
2. If 1 is yes, use the question, How much weight (kg) have you lost?	
None	0
1-5	1
6-10	2
11-15	3
>15	4
Unsure	2
3. Have you been eating poorly because of a decreased appetite?	
No	0
Yes	1

MAIN RESULTS—MULTIVARIATE

The MST, a shortened version of the SGA, has high accuracy for malnutrition (Table 28-11).

Table 28-11 Malnutrition Screening Tool Compared With the Subjective Global Assessment Tool

	LR+ (95% CI)	LR- (95% CI)	DOR (95% CI)
MST $\geq 2^a$	47 (20-110)	0.28 (0.19-0.38)	168 (63-446)

Abbreviations: CI, confidence interval; DOR, diagnostic odds ratio; LR+, positive likelihood ratio; LR-, negative likelihood ratio; MST, Malnutrition Screening Tool.

^aValues represent those predicted by the model, using a cut point of ≥ 2 to predict a subjective global assessment of moderately or severely malnourished. The area under the curve for overall accuracy for the model was 0.97 (95% CI, 0.95-0.99).

The interobserver variability was almost perfect, with a $\kappa = 0.88$ for the MST. The reduced questionnaire showed statistically significant correlations with all the anthropometric variables, all the biochemical variables (except for the total lymphocyte count), and the hospital length of stay (4.9 days for well-nourished vs 9.5 days for those at risk of malnutrition; $P < .001$).

CONCLUSIONS

LEVEL OF EVIDENCE Level 4.

STRENGTHS Large patient population. The interobserver variability was assessed, providing confidence that the tool is reproducible. The study gives some insight into how the clinician might intuitively weight the variables of the SGA.

LIMITATIONS The same person collecting the candidate nutritional screening questions also did the SGA.

Some investigators have used the SGA as the reference standard for malnutrition in adult inpatients. The SGA combines features of the patient medical history and the physical examination with the clinical assessment to sort patients into well-nourished, moderately malnourished, or severely malnourished. Because it requires training and good judgment, it is reasonable to assess whether a smaller set of questions might convey the same answer as the SGA. The multivariate model had a diagnostic odds ratio that was not as good as the single question alone about unintended weight loss (239 for single question vs 169 for the model). The information in the 3 questions of the MST contains similar information to the first 3 questions of the SGA—we can infer that the SGA is highly dependent on these questions. As a data reduction step for less trained clinicians, it makes obvious sense to ask the adult general medical inpatient whether he or she has had unintended weight loss and how much, or a decreased appetite. This simple tool requires validation in adult outpatients who should have a lower prevalence of being at risk of malnutrition than inpatients.

REFERENCE FOR THE EVIDENCE

- Detsky AS, Smalley PS, Chang J. Is this patient malnourished? *JAMA*. 1994;271(1):54-58.

Reviewed by David L. Simel, MD, MHS

TITLE Validation of a Malnutrition Screening Tool (MST) for Patients Receiving Radiotherapy.

AUTHORS Ferguson M, Bauer J, Capra S, Christie DRH, Mason BR.

CITATION *Australas Radiol*. 1999;43(3):325-327.

QUESTION Does a brief malnutrition screening tool (MST) compare well with the subjective global assessment (SGA) for assessing malnourishment?

DESIGN Prospective, independent sample of all patients on designated study days.

SETTING Radiotherapy center at 2 Australian hospitals.

PATIENTS All adult patients who were undergoing radiotherapy on designated study days and agreed to participate ($n = 106$). The patients had a variety of sites affected by carcinoma—32% breast, 19% prostate, 11% gastrointestinal, 9% head and neck—and the rest had a variety of sites. Only 14 patients declined participation (enrollment rate, 88% of potentially eligible patients).

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

The MST was performed by separate dietitians, independent of the SGA. The MST developed in this study was compared

with the SGA that categorized patients into well nourished vs moderately or severely malnourished.

MAIN OUTCOME MEASURE

Diagnostic accuracy of a clinical prediction model compared with the SGA.

MAIN RESULTS

The MST, a shortened version of the SGA, has high accuracy for malnutrition (Table 28-12).

Table 28-12 Malnutrition Screening Tool Compared With the Subjective Global Assessment Tool

Test	LR+ (95% CI)	LR- (95% CI)	DOR (95% CI)
Malnutrition Screening Tool score ≥ 2	5.2 (3.2-7.5)	0 (0-0.76)	101 (9.6-1074)

Abbreviations: CI, confidence interval; DOR, diagnostic odds ratio; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

CONCLUSIONS

LEVEL OF EVIDENCE Level 2.

STRENGTHS The MST was applied by a separate dietitian, independently of the SGA. Lower prevalence of malnourishment compared with that of adult inpatients.

LIMITATIONS Homogenous patient population (cancer patients), although the mix likely included inpatients and outpatients.

This same group of authors developed the MST, and in this study, they applied the tool to a different group of patients. Compared with the model development study and validation, this group of patients had a lower prevalence of malnutrition (making them more comparable to outpatients rather than inpatients). A strength of this study is the independent application of the MST and SGA. The results confirm the diagnostic accuracy of the MST and suggest that the questions in the SGA pertaining to the amount of weight loss and anorexia carry a large amount of the information. It seems intuitive that patients without weight loss and without diminished appetite are less likely to be malnourished, although these data show that seemingly normal weight and appetite do not rule out malnourishment. The results do help clarify the magnitude of the information provided by these simple questions. The screen requires validation in general medical outpatients to assess its generalizability to a less sick population.

Reviewed by David L. Simel, MD, MHS

CHAPTER 29

Does This Patient Have a Mole or a Melanoma?

John D. Whited, MD

James M. Grichnik, MD, PhD

CLINICAL SCENARIO

A patient presenting to your office informs you that he is concerned about a mole on his arm. Although he is not sure how long the mole has been present, he tells you that recently it has enlarged and looks different. As you begin the examination, you also notice the presence of several other moles and ask yourself, is this lesion a benign mole or a malignant melanoma?

WHY IS THIS AN IMPORTANT QUESTION TO ANSWER?

Epidemiology

The incidence rate of malignant melanoma, once considered a rare malignancy, has increased dramatically in recent decades. In 1930, the lifetime risk of an individual in the United States developing melanoma was 1 in 1500. Estimates placed the lifetime risk in 1996 at 1 in 87, with 1 in 75 by 2000.¹ This increased incidence is important because, unlike the more common non-melanoma skin cancers (basal cell carcinoma and squamous cell carcinoma), melanoma is much more likely to cause death. Six of 7 skin cancer deaths are from melanoma.² Although risk generally increases with age, melanoma often occurs in young adulthood. The median age of onset for superficial spreading melanoma, which is by far the most common type of melanoma, is 44 years.³ Thus, a deadly melanoma can strike during early adulthood, resulting in decades of potential life lost.

Early Detection and Prognosis

Metastatic potential and death from melanoma are related to the tumor's level of invasion. The prognosis of melanoma is approximated by relating it to the thickness of the tumor at excision. Melanoma that is confined to the epidermis (*in situ*) is greater than 99% curable.⁴ Patients with thin lesions (thickness < 0.75 mm) have a 5-year survival rate of greater than 98%, whereas those with thicker lesions (> 4 mm) have a less than 50% survival rate.⁵ The prognosis is grim for metastatic disease. Nodal metastatic disease has a 36% 5-year survival rate, which decreases to only 5% with the presence of distant metastases.⁶ Thus, the importance of the physical examination is clear: If thin melanomas are detected and excised, a cure is likely, whereas undetected progression of the tumor markedly decreases a patient's chance of survival.

ANATOMIC AND PHYSIOLOGIC ORIGINS OF THE SIGNS AND SYMPTOMS USED TO EXAMINE THE SKIN FOR MELANOMA

Benign moles and melanoma arise from a cell normally present in the basal layer of the epidermis, called a melanocyte.

The melanocyte produces melanin, which results in skin pigmentation. Alterations in melanin production result in different pigmentary characteristics.

Whereas melanocytes normally exist as solitary units, in moles or nevi, they exist as collections of cells. These include junctional nevi, which are grouped collections of epidermal melanocytes; compound nevi, which are epidermal and dermal collections of melanocytes; and intradermal nevi, which are dermal collections of melanocytes. There also exists a spectrum of nevi that have various degrees of atypia, termed *atypical* or *dysplastic* nevi.

Melanomas lose normal growth controls, change their features, and tend to grow in an irregular manner, leading to asymmetry, irregular borders, and haphazard coloration. This contrasts with benign nevi, which are characteristically more stable, more symmetric, have well-defined borders, and have even color distribution. However, these features are not absolute, and caution is warranted, particularly when change has been noted.

HOW TO EXAMINE THE SKIN FOR MELANOMA

Historical Feature Assessment

History plays an important role in the examination of the skin for melanoma. Patients should be asked whether they have noted any lesions of concern, particularly any new moles or a change in size, shape, color, or sensation of a preexisting mole. This is critical information because approximately one-half of melanomas are initially discovered by the patient.⁷ Changes in size or color are the 2 most common patient-reported features of melanoma.⁸⁻¹⁰ Bleeding, tenderness or pain, and itching are also reported, although these features occur in more invasive lesions.⁸ Patients should also be asked about a personal or family history of melanoma. The results of previous skin biopsies and any history of nonmelanoma skin cancer should also be assessed. The patient's tendency to sunburn and a history of sunburns may also help assess risk. The presence of focal or systemic symptoms or the presence of any lumps or bumps under the skin should be addressed, particularly in a patient with a history of a cutaneous malignancy.

Physical Examination Technique

To examine for melanoma, the entire skin surface should be inspected. Melanoma can occur anywhere on the skin and may develop in sun-protected areas. Patients who undergo complete cutaneous examinations are 6.4 times more likely to have a melanoma detected than patients receiving only a partial examination.¹¹ The patient should be examined head to toe in a well-lit room. A gown may be used and removed incrementally to evaluate various regions of the patient's entire body surface. The examination of the patient's scalp may be aided by sequentially parting the hair or by the use of a handheld hair dryer. The oral mucosa, genital area, nails, and the skin between the toes should be included in the inspection for evidence of pigmented lesions.

When the patient is examined, it is also important to make a global assessment of his or her skin. For example, if your patient has multiple nevi, those nevi may have relatively uniform characteristics. However, if one of the moles has unusual features that are dissimilar to those of other nevi, that lesion should be more closely examined. In the same manner, a single pigmented lesion occurring in a patient without other nevi should be evaluated with an increased level of concern. Patients at high risk for melanoma appear to be those with numerous nevi, those with nevi with atypical features, and particularly those with a family history of melanoma.¹²⁻¹⁴ In patients with an increased risk of melanoma, regular skin examinations may result in melanoma detection at an earlier, thinner stage.¹⁵⁻¹⁸

Checklists as a Diagnostic Aid

In the United States, the ABCD checklist for detecting cutaneous melanoma is recommended as a means for distinguishing benign lesions from melanoma.¹⁹ The criteria making up the ABCD checklist are all physical examination features: (1) when the lesion is bisected, half is not identical to the other half: *asymmetry* (A); (2) when the border is uneven or ragged as opposed to smooth and straight: *border irregularity* (B); (3) when more than 1 shade of pigment is present: *color variegation* (C); and (4) when the lesion is greater than 6 mm in diameter (D).

Lesions that have these features should raise suspicions of a melanoma. Friedman et al¹⁹ state that, although not incorporated into the ABCD checklist, historical features of a changing, preexisting pigmented nevus or the development of a new pigmented lesion should alert the physician to the possibility of malignant melanoma. An amendment to the checklist that adds an (E), representing an elevation above the skin surface, was proposed in 1988.²⁰ See the Update to this chapter for additional details.

A second checklist is the revised 7-point checklist used in the United Kingdom.²¹ Three major criteria, all historical features, and 4 minor criteria, primarily physical examination features, are used to evaluate lesions suggestive of melanoma. The checklist was developed mainly for use by primary care physicians to assist them in making referral decisions. The major criteria are change in size, shape, and color; the minor criteria are inflammation, crusting or bleeding, sensory change, and a diameter 7 mm or greater.

One interpretation of this guideline states that the major criteria are the basis for determining referral decisions.²¹ Any patient with at least 1 major sign should be referred to a dermatologist. The revised 7-point checklist has also been given a slightly different interpretation, with change in shape replaced by irregular shape (or appearance of irregularity in an old lesion), change in color replaced by irregular color, and a greater importance placed on the minor criteria.^{22,23} A scoring system assigns 2 points for each major criterion and 1 point for each minor criterion. If a score of 3 points or more is noted, then a referral for lesion evaluation was suggested.

Criterion Standard for Diagnosing Melanoma

The criterion standard for the diagnosis of melanoma is the histopathologic evaluation of excised tissue.

METHODS

Search Strategy and Quality Filter

A literature search was performed using MEDLINE for 1966 through 1996. Medical Subject Heading terms “melanoma” and “skin neoplasms” were combined with “physical examination,” “sensitivity,” “specificity,” “observer variation,” “mass screening,” and “self-examination,” yielding approximately 713 citations. In addition, a MEDLINE search was performed with the search strategy developed for this series of articles, which yielded 659 citations. Titles, abstracts, and relevant articles were reviewed in their entirety. Current Contents (Institute for Scientific Information) were reviewed with the terms “melanoma,” “skin cancer,” and “mass screening” to search for more current articles. References for articles found by the search strategy and other manuscripts pertaining to melanoma were systematically reviewed for additional literature sources.

The quality of the published articles was evaluated as previously described.²⁴ For studies that assessed accuracy, 20 articles were reviewed. The 95% confidence intervals (CIs) reported here, when not reported in the original articles, were calculated from the available data when possible for test performance characteristics. Studies were included if the level of evidence was graded as C or above. Lack of independence between the reference standard and the test, leading to verification bias, occurs in the existing literature. Another methodologic issue relates to the nature of the reference standard, namely, histologic tissue obtained by biopsy. The decision to perform a skin biopsy requires clinical judgment because a biopsy specimen is not obtained for all patients with skin lesions. This requires using follow-up examinations, multiple examiners, or even consensus opinion to ascertain the diagnosis. No existing studies were given a quality score of A or B; thus, all 12 studies graded as C were included.

RESULTS

Precision of the Skin Examination for Melanoma

Two studies evaluated examiners’ precision for specific features of benign pigmented lesions, which include 4 of the features found in the ABCD(E) checklist. Physicians examining the most atypical pigmented lesion found on patients recently diagnosed as having malignant melanoma displayed a moderate level of interobserver agreement.²⁵ Among 3 examiners (medical oncologist, internist/epidemiologist, and dermatologist/dermatopathologist), the intraclass correlation coefficient was highest for degree of macularity, corresponding with elevation (E) at 0.56, asymmetry (A) at 0.46, haphazard color (C) at 0.44, and border irregularity (B) at 0.40. A second study assessed interobserver and intraobserver agreement among 3 physicians using photographs of melanocytic nevi.²⁶ After establishing criteria to be assessed for each feature, the level of agreement was similar to that found in the first study, although less precision was noted for rating asymmetry. Interobserver agreement for physician pairs, as measured by the κ statistic, ranged from 0.41 to 0.55 for macular vs papular lesions (E), 0.38 to 0.53 for color variegation (C), 0.29 to 0.53 for border irregularity (B), and 0.05 to 0.26 for contour irregularity (A). The level of intraobserver agreement was, overall, similar to interobserver agreement. However, intraobserver agreement was measured according to a 4-point scale, which graded the degree of each feature, rather than the presence or absence of each feature. These precision estimates are considered fair to moderate.²⁷ Because only benign pigmented lesions were assessed, observer agreement for these features found in actual malignant melanoma lesions may be higher than reported in these studies. Precision estimates for global assessments of the presence or absence of melanoma are not available.

Accuracy of Skin Examination for Melanoma With ABCD(E) and Revised 7-Point Checklists

Two studies have assessed accuracy of the ABCD(E) checklist (Table 29-1). Different features of the checklist were assessed,

Table 29-1 Operating Characteristics for the ABCD(E) Checklist

Source, y	Examiners	Setting	No. With Disease/ No. Without Disease	Sensitivity, % (95% CI)	Specificity, % (95% CI)	LR for a Positive Test Result (95% CI)	LR for a Negative Test Result (95% CI)
Healsmith et al, ²⁸ 1994 ^a	Dermatologists	Pigmented lesion clinic	65/0	92 (82-96)	... ^b
McGovern and Litaker, ²⁹ 1992 ^c	Chart review of dermatologists' examinations	Dermatology clinic	6/186	100 (54-100)	98 (95-99)	62 (19-170)	0 (0-0.5)

Abbreviations: CI, confidence interval; LR, likelihood ratio.

^aA positive test result required the lesion to have 1 or more of the ABCD(E) criteria: A, symmetry; B, border irregularity; C, irregular color; D, diameter greater than 6 mm; and E, elevation.

^bEllipses indicate data not available.

^cA positive test result required the lesion to have border irregularity (B), color irregularity (C), and diameter greater than 6 mm (D).

and the interpretation of a positive test result was not the same in both studies. Features of the ABCD(E) checklist were prospectively recorded for patients with pigmented lesions who had been referred to a clinic for pigmented lesions.²⁸ A total of 65 histologically confirmed melanomas were included in the analysis. Only 5 lesions were not identified by the ABCD(E) portion of the checklist, resulting in a sensitivity of 92% (95% CI, 82%-96%). The ABCD(E) checklist was considered positive when a lesion had 1 or more of the 5 features. Specificity was not reported.

A second study used a retrospective design to assess 3 features of the ABCD(E) checklist; border irregularity (B), color variegation (C), and diameter (D).²⁹ The checklist was applied by reviewing charts and pathology reports among patients who had undergone biopsies of pigmented lesions during a 1-year period. Pigmented lesion biopsy specimens were included when the dermatologist's pathology submission form indicated clinical diagnoses of dysplasia, lentigo maligna, or malignant melanoma. All 6 histologically confirmed melanomas had all 3 features on the checklist. The sensitivity was therefore 100% (95% CI, 54%-100%). Only 3 lesions that were benign had all 3 features, resulting in a specificity of 98% (95% CI, 95%-99%).

In the study by Healsmith et al,²⁸ all 5 of the melanomas that were not identified had a diameter of less than 6 mm, although a change in size was observed.²⁸ Because of concerns that requiring lesions to be larger than 6 mm may lower the sensitivity of the ABCD(E) checklist, resulting in missed lesions, 1150 melanomas that underwent biopsy during a 27-month period in Australia were retrospectively analyzed for their size.³⁰ Three hundred fifty-eight (31%) of 1150 of the melanomas were 6 mm or less in diameter. This indicates that requiring a diameter of greater than 6 mm in this sample of lesions would have lowered the sensitivity considerably.

More data exist for differentiating benign lesions from melanoma with the revised 7-point checklist than with the ABCD(E) checklist (Table 29-2). The revised 7-point checklist was also analyzed against the 65 histologically confirmed melanomas that were found during a 38-month period in the aforementioned prospective analysis by Healsmith et al.²⁸

This same checklist was applied to 100 randomly selected benign pigmented lesions—68 were considered benign according to clinical characteristics and 32 were histologically confirmed to be benign. The sensitivity of the revised 7-point checklist was 100% (95% CI, 94%-100%), because all melanomas had at least 1 major feature. The specificity was lower, at only 37% (95% CI, 28%-46%).

A second study reported the sensitivity of the revised 7-point checklist applied to 100 patients with histologically proven malignant melanoma to be 79% (95% CI, 70%-85%).²² In this study, the alternative interpretation of the revised 7-point checklist was used; features of the checklist were assigned scores, 2 points for each major feature and 1 point for each minor feature present. A score of 3 or more was considered a lesion that should be referred to a dermatologist because of its malignant potential. The checklist was prospectively applied to patients presenting to a clinic for pigmented lesions with lesions suggestive of melanoma.

The specificity of the revised 7-point checklist, again using the scoring system, was estimated by applying it to a consecutive series of 100 histologically benign lesions.²³ Seventy of the benign lesions achieved a score indicative of malignancy, resulting in a specificity of 30% (95% CI, 21%-39%). This is the only study that has assessed accuracy of patient assessments, reporting a specificity comparable to the physician evaluations of 32% (95% CI, 23%-41%).

Studies assessing accuracy of the checklists have not applied and interpreted the criterion standard independently with the checklists and should be interpreted with some discretion. Additionally, both the revised 7-point checklist and the ABCD(E) checklist have been subject to various interpretations of the requirements for positive and negative test results. With that in mind, existing evidence suggests that both checklists result in a sensitive diagnostic test. A highly sensitive test is desirable for a disease such as melanoma, which if left undetected can result in death. When the ABCD(E) checklist is used, requiring a lesion to be greater than 6 mm in diameter may lower the sensitivity, which could result in missed lesions. It appears that the checklists' high sensitivity may come at the expense of low specificity,

Table 29-2 Operating Characteristics for the Revised 7-Point Checklist

Source, y	Examiners	Setting	No. With Disease/ No. Without Disease	Sensitivity, % (95% CI)	Specificity, % (95% CI)	LR for a Positive Test Result (95% CI)	LR for a Negative Test Result (95% CI)
Healsmith et al, ²⁸ 1994 ^a	Dermatologists	Pigmented lesion clinic	65/100	100 (94-100)	37.0 (28-46)	1.6 (1.4-1.9)	0 (0-0.2)
Du Vivier et al, ²² 1991 ^b	Dermatologists	Pigmented lesion clinic	100/0	79 (70-85)	... ^c
Higgins et al, ²³ 1992 ^b	Dermatologists	Pigmented lesion clinic	0/100	...	30 (21-39)

Abbreviations: CI, confidence interval; LR, likelihood ratio.

^aA positive test result required the presence of 1 major feature: change in size, change in color, or change in shape.

^bA positive test result required a score of 3 points, with 2 points being assigned to a major criterion (change in size, irregular shape, or irregular color) and 1 point for a minor criterion (presence of inflammation, diameter ≥ 7 mm, crusting or bleeding, and sensory change).

^cEllipses indicate data not available.

especially when the revised 7-point checklist is used. No conclusions can be drawn about the specificity of the ABCD(E) checklist from the available data.

Accuracy for Detecting the Presence or Absence of Melanoma

Accuracy studies of global assessments for detecting melanoma use 2 methods of examination: actual patient examination and image evaluation through the use of pictures, slides, or digitized images of lesions. Accuracy assessments have included dermatologist and nondermatologist examiners.

Dermatologists have primarily been the examiners in studies using patient examinations. Estimates for sensitivity range widely from 50% to 97%, whereas specificity estimates have been more consistent, ranging from 96% to 99% (**Table 29-3**).³¹⁻³⁵ Data from existing studies generally do not allow for calculation of likelihood ratios. However, the positive predictive value, although influenced by prevalence, is often reported. In the largest series of patients followed after completion of a screening skin examination, the positive predictive value was found to be 17%.³⁶ Other estimates of the positive predictive value vary greatly, from 35% to 86%.^{32-34,37}

Lesions presented as pictures, slides, and digitized computer images rather than patient evaluations have been an alternative mode of evaluation used to assess accuracy and have often been used to compare nondermatologists' examinations to those performed by dermatologists (**Tables 29-4** and **29-5**). One study presented melanoma lesions in both a 35-mm slide and a digitized computer image format to nondermatologists (general internal medicine and family practice

residents) and dermatologists (resident and attending physicians).³⁸ Nondermatologists provided the correct diagnosis 60% of the time compared with dermatologists, who correctly diagnosed the lesions 74% of the time, a difference that was statistically significant. The correct treatment option (defined as recognition of the need for a biopsy and the type of biopsy required) was selected by nondermatologists significantly less often (52%) than by dermatologists (67%). The correct diagnosis and treatment options were determined by biopsy results and consensus opinion of 2 dermatologists. In another study that compared nondermatologist examiners (first-year internal medicine residents and practicing physicians) to dermatologists (third-year residents and practicing dermatologists), 100% of the dermatologists correctly identified at least 3 of the 6 melanoma lesions compared with 70% of the nondermatologists.³⁹ In an additional study that compared practicing primary care physicians and internal medicine residents with dermatology faculty, 88% of the nondermatologists correctly identified melanoma compared with 100% of the dermatologist examiners.⁴⁰ General practitioners in Australia who were shown pictures, which included 2 early melanoma lesions and 1 late melanoma lesion, correctly identified all 3 as melanoma 41% of the time.⁴¹ However, they made the correct decision to perform a biopsy on the lesion 83% of the time. Another study of similar design from New Zealand found that a high proportion of correct diagnoses and biopsy decisions was made by general practitioners.⁴² A correct diagnosis was made by general practitioners in 81% of cases compared with 90% by dermatologists. Recognizing the need for a biopsy was similar for both groups, with the correct biopsy decision being made greater than 95% of the time.

Table 29-3 Operating Characteristics for Global Assessments of the Presence or Absence of Melanoma

Source, y	Examiners	Setting	No. With Disease/No. Without Disease	Sensitivity, % (95% CI)	Specificity, % (95% CI)	LR for a Positive Test Result (95% CI)	LR for a Negative Test Result (95% CI)
DeCoste and Stern, ³¹ 1993	Pathology report review of specimens submitted by dermatologists	Dermatology clinic	Unknown	50	... ^a
Grin et al, ³² 1990	Pathology report review of specimens submitted by dermatologists	Oncology section/skin cancer unit	265/10436	81 (75-85)	99.2 (99.1-99.4)	107 (85-134)	0.2 (0.15-0.25)
Koh et al, ³³ 1990	Dermatologists	Melanoma/skin cancer screening clinic	9/0	97 ^b
McMullan and Hubener, ³⁴ 1956	Pathology report review of specimens submitted by dermatologists	Unknown	87/0	51 (40-60)
Curley et al, ³⁵ 1989	Physicians experienced in managing melanocytic lesions	Pigmented lesion clinic	3/114	...	96-99 ^c

Abbreviations: CI, confidence interval; LR, likelihood ratio.

^aEllipses indicate data not available.

^bCalculated from estimated false-negative rates.

^cRange of results from 3 examiners.

Table 29-4 Proportion of Correct Diagnoses by Specialty

Source, y	Format of Lesion Presentation	No. of Unique Melanoma Lesions Reviewed	Nondermatologist Participants	Dermatologist Participants	Correct Diagnosis, %	
					Nondermatologists	Dermatologists
Gerbert et al, ³⁸ 1996	35-mm Slides and digitized computer images	12	Internal medicine and family practice residents	Dermatology residents and attending physicians	60	74
Cassileth et al, ³⁹ 1986	35-mm Slides	6	Medical students, internal medicine residents, fellows, and practicing physicians	Dermatology residents and practicing physicians	70 ^a	100 ^a
Ramsay and Fox, ⁴⁰ 1981	Slides	1	Participants of review courses and internal medicine residents	Dermatology residents and faculty	88	100
Paine et al, ⁴¹ 1994	Photographs	3	General practitioners (Australia)	... ^b	41	...
McGee et al, ⁴² 1994	Photographs	3	General practitioners (New Zealand)	Physicians registered as dermatologists	81	90

^aProportion correctly identifying at least 3 of 6 lesions.^bEllipses indicate data not available.

Specificity is high, at least among dermatologist examiners, when patient populations are examined for melanoma. The sensitivity of patient examinations for melanoma is less clear, and better study designs are needed to provide more accurate sensitivity assessments. Studies that have used images of lesions rather than patient examinations have indicated that nondermatologists' examinations are less sensitive than examinations performed by dermatologists. Studies using examinations on patient populations, with rigorous application of the test and classification of the disease state, which include dermatologists and nondermatologists as examiners, are needed to provide better assessments of operating characteristics.

THE BOTTOM LINE

Returning to the clinical scenario, a concerned patient presents with an enlarging mole on his arm that has changed in appearance. According to the existing literature, the usefulness

of the ABCD(E) checklist or revised 7-point checklist to distinguish melanoma from benign lesions is not fully established. If a positive test result does not require that all 4 features of the ABCD(E) checklist be present, misdiagnosing a melanoma as a benign lesion appears to be unlikely. The accuracy of using the ABCD(E) checklist to predict the disease state when a positive test result requires the presence of all 4 features has not been described. However, early melanoma lesions may be small (<6 mm in diameter), and requiring a lesion to be greater than 6 mm in diameter when using the checklist may result in some early lesions to be falsely classified as benign. It is unclear how often benign lesions would be considered to have malignant potential with the ABCD(E) checklist. The patient's report of the lesion enlarging and changing in appearance incorporates the primary criteria used in the revised 7-point checklist. When the revised 7-point checklist is used, misdiagnosing a melanoma as benign would also be unlikely, although it appears the checklist may classify many benign lesions as malignant.

Table 29-5 Proportion of Correct Treatment Options by Specialty

Source, y	Format of Lesion Presentation	No. of Unique Melanoma Lesions Reviewed	Nondermatologist Participants	Dermatologist Participants	Correct Treatment, %	
					Nondermatologists	Dermatologists
Gerbert et al, ³⁸ 1996	35-mm Slides and digitized computer images	12	Internal medicine and family practice residents	Dermatology residents and attending physicians	52 ^a	67 ^a
Paine et al, ⁴¹ 1994	Photographs	3	General practitioners (Australia)	... ^b	83	...
McGee et al, ⁴² 1994	Photographs	3	General practitioners (New Zealand)	Physicians registered as dermatologists	96	97

^aBiopsy requirement and type of biopsy considered in correct treatment options.^bEllipses indicate data not available.

In summary, malignant melanoma is an increasingly common malignancy, with an incidence rate that is projected to increase. The medical history and physical examination play a unique role in the secondary prevention of cutaneous malignant melanoma. It is the sole means of identifying lesions that require excision for histopathologic evaluation. Because of the growth characteristics of melanoma, examinations that detect earlier stages of melanoma can result in a better prognosis. The utility of the ABCD(E) and revised 7-point checklists for distinguishing melanoma from benign skin lesions is not conclusively described. The ABCD(E) checklist (when a positive test result does not require all 4 features to be present) and the revised 7-point checklist appear to be sensitive diagnostic aids in evaluating individual lesions and therefore rarely classify a melanoma as a benign lesion. However, the revised 7-point checklist lacks specificity, resulting in benign lesions being classified as potentially malignant. The specificity of the ABCD(E) checklist is less well described. A change in lesion characteristics is frequently reported by patients with melanoma and is an important feature to assess during an examination. Better study designs are necessary to define the operating characteristics of physicians' examinations for detecting the presence or absence of melanoma. Existing evidence suggests that examinations are highly specific, at least among dermatologist examiners, but sensitivity estimates are less clear. Data regarding nondermatologists' examinations suggest that their examinations are less sensitive than those performed by dermatologists.

Author Affiliations at the Time of the Original Publication

Center for Health Services Research in Primary Care, Durham Veterans Affairs Medical Center (Dr Whited), and Divisions of General Internal Medicine (Dr Whited) and Dermatology (Dr Grichnik), Duke University Medical Center, Durham, North Carolina.

Acknowledgments

We thank Claude S. Burton III, MD, Sheri A. Keitz, MD, PhD, Frank J. Spence, Jr, MD, and Robert J. Sullivan, Jr, MD, MPH, for their review of drafts of the manuscript.

REFERENCES

- Rigel DS, Friedman RJ, Kopf AW. The incidence of malignant melanoma in the United States: issues as we approach the 21st century. *J Am Acad Dermatol*. 1996;34(5 pt 1):839-847.
- American Academy of Dermatology. *Program and Screening Guidelines: National Skin Cancer Prevention Education Program*. Schaumburg, IL: American Academy of Dermatology; 1996:37.
- Barnhill RL, Mihm MC Jr, Fitzpatrick TB, Sober AJ. Neoplasms: malignant melanoma. In: Fitzpatrick TB, Eisen AZ, Wolff K, eds. *Dermatology in General Medicine*. 4th ed. New York, NY: McGraw-Hill Inc; 1993: 1078-1115.
- NIH Consensus Development Panel on Early Melanoma. Diagnosis and treatment of early melanoma. *JAMA*. 1992;268:1314-1319.
- Day CL Jr, Mihm MC Jr, Lew RA, Kopf AW, Sober AJ, Fitzpatrick TB. Cutaneous malignant melanoma: prognostic guidelines for physicians and patients. *CA Cancer J Clin*. 1982;32(2):113-122.
- Koh HK. Cutaneous melanoma. *N Engl J Med*. 1991;325(3):171-182.
- Koh HK, Miller DR, Geller AC, Clapp RW, Mercer MB, Lew RA. Who discovers melanoma: patterns from a population-based survey. *J Am Acad Dermatol*. 1992;26:914-919.
- Wick MM, Sober AJ, Fitzpatrick TB, et al. Clinical characteristics of early cutaneous melanoma. *Cancer*. 1980;45(10):2684-2686.
- Keele M, Dick DC, Wakeel RA. A study of the value of the seven-point checklist in distinguishing benign pigmented lesions from melanoma. *Clin Exp Dermatol*. 1990;15(3):167-171.
- Elwood JM, Gallagher RP. The first signs and symptoms of melanoma: a population-based study. *Pigment Cell*. 1988;9:118-130.
- Rigel DS, Friedman RJ, Kopf AW, et al. Importance of complete cutaneous examination for the detection of malignant melanoma. *J Am Acad Dermatol*. 1986;14(5 pt 1):857-860.
- Swerdlow AJ, English J, MacKie RM, et al. Benign melanocytic naevi as a risk factor for malignant melanoma. *Br Med J (Clin Res Ed)*. 1986; 292 (6535):1555-1559.
- Greene MH, Clark WH, Tucker MA, Kraemer KH, Elder DE, Freaser MC. High risk of malignant melanoma in melanoma-prone families with dysplastic nevi. *Ann Intern Med*. 1985;102(4):458-465.
- Tucker MA, Halpern A, Holly EA, et al. Clinically recognized dysplastic nevi: a central risk factor for cutaneous melanoma. *JAMA*. 1997;277 (18):1439-1444.
- Rigel DS, Rivers JK, Kopf AW, et al. Dysplastic nevi: markers for increased risk for melanoma. *Cancer*. 1989;63(2):386-389.
- Tiersten AD, Grin CM, Kopf AW, et al. Prospective follow-up for malignant melanoma in patients with atypical-mole dysplastic-nevus syndrome. *J Dermatol Surg Oncol*. 1991;17(1):44-48.
- Masri GD, Clark WH, Guerry D, Halpern A, Thompson CJ, Elder DE. Screening and surveillance of patients at high risk for malignant melanoma result in detection of earlier disease. *J Am Acad Dermatol*. 1990; 22(6 pt 1):1042-1048.
- Titus-Ernstoff L, Ernstuff MS, Kirkwood JM, Barnhill RL, Fine J, Duray PH. Usefulness of frequent skin examination for the early detection of secondary primary cutaneous melanoma. *Cancer Detect Prev*. 1989;13(5-6):317-321.
- Friedman RJ, Rigel DS, Kopf AW. Early detection of malignant melanoma: the role of physician examination and self-examination of the skin. *CA Cancer J Clin*. 1985;35(3):130-151.
- Fitzpatrick TB, Rhodes AR, Sober AJ, Mihm MC Jr. Primary malignant melanoma of the skin: the call for action to identify persons at risk; to discover precursor lesions; to detect early melanomas. *Pigment Cell*. 1988;9:110-117.
- MacKie RM. Clinical recognition of early invasive melanoma. *BMJ*. 1990;301(6759):1005-1006.
- Du Vivier AWP, Williams HC, Brett JV, Higgins EM. How do malignant melanomas present and does this correlate with the seven-point checklist? *Clin Exp Dermatol*. 1991;16(5):344-347.
- Higgins EM, Hall P, Todd P, Murhti R, Du Vivier AWP. The application of the seven-point check-list on the assessment of benign pigmented lesions. *Clin Exp Dermatol*. 1992;17(5):313-315.
- Holleman DR, Simel DL. Does the clinical examination predict airflow limitation? *JAMA*. 1995;273(4):313-319.
- Barnhill RL, Roush GC, Ernstuff MS, Kirkwood JM. Interclinician agreement on the recognition of selected gross morphologic features of pigmented lesions: studies of melanocytic nevi V. *J Am Acad Dermatol*. 1992;26(2 pt 1):185-190.
- Meyer LJ, Piepkorn M, Goldgar DE, et al. Interobserver concordance in discriminating clinical atypia of melanocytic nevi, and correlations with histologic atypia. *J Am Acad Dermatol*. 1996;34(4):618-625.
- Landis JR, Koch GG. The measurement of observer agreement in categorical data. *Biometrics*. 1977;33(1):159-174.
- Healsmith MF, Bourke JF, Osborne JE, Graham-Brown RAC. An evaluation of the revised seven-point checklist for the early diagnosis of cutaneous melanoma. *Br J Dermatol*. 1994;130(1):48-50.
- McGovern TW, Litaker MS. Clinical predictors of malignant pigmented lesions: a comparison of the Glasgow seven-point checklist and the American Cancer Society's ABCDs of pigmented lesions. *J Dermatol Surg Oncol*. 1992;18(1):22-26.
- Shaw HM, McCarthy WH. Small-diameter malignant melanoma: a common diagnosis in New South Wales, Australia. *J Am Acad Dermatol*. 1992;27(5 pt 1):679-682.

31. DeCoste SD, Stern RS. Diagnosis and treatment of nevomelanocytic lesions of the skin: a community-based study. *Arch Dermatol.* 1993;129(1):57-62.
32. Grin CM, Kopf AW, Welkovich B, Bart RS, Levenstein MJ. Accuracy in the clinical diagnosis of malignant melanoma. *Arch Dermatol.* 1990;126(6):763-766.
33. Koh HK, Caruso A, Gage I, et al. Evaluation of melanoma/skin cancer screening in Massachusetts: preliminary results. *Cancer.* 1990;65(2):375-379.
34. McMullan FH, Hubener LF. Malignant melanoma: a statistical review of clinical and histological diagnoses. *Arch Dermatol.* 1956;74(6):618-619.
35. Curley RK, Cook MG, Fallowfield ME, Marsden RA. Accuracy in clinically evaluating pigmented lesions. *BMJ.* 1989;299(6690):16-18.
36. Koh HK, Norton LA, Geller AC, et al. Evaluation of the American Academy of Dermatology's national skin cancer early detection and screening program. *J Am Acad Dermatol.* 1996;34(6):971-978.
37. Krol S, Keijser LMT, Van Der Rhee HJ, Welvaart K. Screening for skin cancer in the Netherlands. *Acta Derm Venereol.* 1991;71(4):317-321.
38. Gerbert B, Maurer T, Berger T, et al. Primary care physicians as gatekeepers in managed care: primary care physicians' and dermatologists' skills at secondary prevention of skin cancer. *Arch Dermatol.* 1996;132(9):1030-1038.
39. Cassileth BR, Clark WH Jr, Lusk EJ, Frederick BE, Thompson CJ, Walsh WP. How well do physicians recognize melanoma and other problem lesions? *J Am Acad Dermatol.* 1986;14(4):555-560.
40. Ramsay DL, Fox AB. The ability of primary care physicians to recognize the common dermatoses. *Arch Dermatol.* 1981;117(10):620-622.
41. Paine SL, Cockburn J, Noy SM, Marks R. Early detection of skin cancer: knowledge, perceptions, and practices of general practitioners in Victoria. *Med J Aust.* 1994;161(3):188-195.
42. McGee R, Elwood M, Sneyd MJ, Williams S, Tilyard M. The recognition and management of melanoma and other skin lesions by general practitioners in New Zealand. *N Z Med J.* 1994;107(982):287-290.

Prepared by David L. Simel, MD, MHS, and James M. Grichnik, MD, PhD

Reviewed by John D. Whited, MD

CLINICAL SCENARIO

A 40-year-old non-Hispanic white patient presents to your clinic with concern about some skin lesions. He has no personal history of dysplastic nevi and no one in his family has melanoma. He confesses that he is not particularly worried about them, but his girlfriend is worried. There are 2 lesions on the upper back, neither of which the patient can see directly. He can feel one and observes that perhaps it has changed in size. As a child, he typically went without a shirt during much of the summer and did not use sunscreen. Sometimes, he sunburned with prolonged exposure. He has not experienced sunburns since his teenage years. The lesions are shown in Figures 29-1 and 29-2.



Figure 29-1 Lesion 1

UPDATED SUMMARY ON MELANOMA

Original Review

Whited JD, Grichnik JM. Does this patient have a mole or a melanoma? *JAMA*. 1998;279(9):696-701.

UPDATED LITERATURE SEARCH

Details of the Update

Our literature search used the parent search strategy for The Rational Clinical Examination series, combined with the subject “melanoma/di,” published in English from 1997 to 2004. We also crossed the clinical subject headings with “meta-analysis,” “ROC curve,” and the textwords “ABCDE,” “7-point,” and “seven-point” in the MEDLINE database. The results yielded 179 titles, for which we reviewed the titles and abstracts; 48 were selected for additional review. These articles were reviewed to identify articles that assessed the sensitivity and specificity of the medical history or physical examination features of nevi for melanoma. We required that the studies be done on actual patients (as opposed to pictures), involve prospectively collected data, and use basic observational skills used by general practitioners as opposed to examinations requiring special equipment. Because our



Figure 29-2 Lesion 2

focus was on the actual features of the examination itself rather than the overall accuracy of the examination, we eliminated studies without data on individual findings or the use of standardized checklists. We retained only 1 article on the ABCDE criteria. We found no additional studies on the 7-point checklists mentioned in the original publication.

NEW FINDINGS

- The “E” of the ABCDE checklist now represents “enlargement” as reported by the patient, rather than “elevation” as determined by the clinician (A = asymmetry in 2 axes, B = border irregularity, C = more than 1 color, D = dimension ≥ 6 mm).
- The patient reports that a lesion has enlarged is the single most powerful finding of the ABCDE criteria.
- Any single positive finding of the ABCDE criteria may justify a biopsy or referral to rule out a melanoma. The greater the number of findings, the greater the likelihood of melanoma. Dermatologists can use other techniques to help distinguish melanomas from atypical nevi.

IMPROVEMENTS IN THE DATA PRESENTED IN THE ORIGINAL PUBLICATION

A high quality study of the revised ABCDE criteria supports this screening paradigm.

CHANGES IN THE REFERENCE STANDARD

There are 3 issues involved in determining a reference standard. One acceptable reference standard is the result of histopathology. This standard can apply to all examiners. The reference standard can be stratified by whether the patient has a melanoma or combined to create a composite of melanoma

or dysplasia. A second reference standard is not precisely one of diagnostic accuracy, but instead assesses whether a primary care provider made the right “diagnosis” of a patient’s lesion, as evidenced by the decision to refer or biopsy. These studies use expert dermatologists who evaluate the patient or photographs of the lesion against a set of criteria for appropriateness. Finally, the inability to biopsy all lesions on a patient means that in a research study, the only patients enrolled are those with a suspicious lesion. This creates verification bias that could be avoided by following patients who do not undergo a biopsy. With a reasonable follow-up period, an unchanging lesion evaluated through direct observation and serial photographs could be accepted as proof of a nonmalignant melanocytic lesion.

RESULTS OF LITERATURE REVIEW

A large study,¹ conducted by dermatologists, represents the largest prospective evaluation of the revised ABCDE criteria that also uses a group of patients without melanoma (see Tables 29-6 and 29-7). Although the “E” initially represented elevation of the lesion, in this study it represented the patient’s report that the lesion had enlarged. No study has evaluated these criteria in a large population of patients treated initially by primary care providers. Dermatologists typically have greater sensitivity for accurately diagnosing melanoma than primary care providers, but the specificity of primary care providers has not been studied well.² We infer that dermatologists’ sensitivity for the individual ABCDE criteria would be higher than the sensitivity for primary care physicians.

A variety of studies on dermoscopy, including a systematic review, were identified.³ Dermoscopy, variably called dermatoscopy or epiluminescence microscopy, involves viewing a lesion through a handheld microscope that is similar to an otoscope. Dermoscopes provide $\times 10$ or higher magnification of the lesion through immersion oil (or cross-polarized light) to reduce surface reflection and allow the visualization of colors and patterns not easily seen with the naked eye. In general, trained dermatologists use this procedure as an examination secondary to basic clinical observations. The intent of dermoscopy is to provide additional information and improve diagnostic skill. Training is required, and the utility of dermoscopy for primary care providers remains under study and is not deemed sufficiently developed at this point for this review.

Table 29-6 Univariate Findings for Melanoma From the ABCDE Criteria

Finding	LR+ (95% CI)	LR- (95% CI)
A (asymmetry)	2.1 (1.9-2.5)	0.59 (0.52-0.66)
B (border)	2.1 (1.8-2.4)	0.59 (0.53-0.67)
C (color)	1.6 (1.5-1.8)	0.59 (0.52-0.68)
D (dimension)	2.3 (2.1-2.5)	0.17 (0.13-0.22)
E (enlargement)	11 (8.5-14)	0.18 (0.15-0.22)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

Table 29-7 Multivariate Findings for Melanoma

No. of Positive ABCDE Findings	LR (95% CI)
5 Positive	98 (31-303)
≥ 4	8.3 (6.2-11)
≥ 3	3.3 (2.8-3.9)
≥ 2	2.6 (2.3-2.9)
≥ 1	1.5 (1.4-1.6)
0	0.07 (0.04-0.13)

Abbreviations: CI, confidence interval; LR, likelihood ratio.

EVIDENCE FROM GUIDELINES

The US Preventive Services Task Force found the benefits of screening for melanoma unproven.⁴ However, the recommendation addressed screening with a total body skin examination. The task force recommends that clinicians be aware of the ABCD criteria or rapidly changing lesions that become apparent for whatever reason and that lesions with 1 or more abnormalities be biopsied. The evidence report behind those

Table 29-8 Skin Type Risk Factors for Melanoma⁶

Skin Type	Do You Burn in the Sun?	Do You Tan After Having Been in the Sun?
I	Always	Seldom
II	Usually	Sometimes
III	Sometimes	Usually
IV	Seldom	Always
V	Naturally brown skin	
VI	Naturally black skin	

recommendations did not address the accuracy of the individual criteria.⁴ Using the data of the US Preventive Services Task Force, the Canadian Task Force on Preventive Health concluded that the evidence is conflicting for the total body skin examination in the general population.⁵ The World Health Organization makes no specific recommendations on screening. They do suggest stratifying patients by risk factor. Patients with skin types I to II are at the highest risk of melanoma, types III to IV confer intermediate risk, and types V to VI have the lowest risk (see Table 29-8).⁶

MELANOMA—MAKE THE DIAGNOSIS

PRIOR PROBABILITY

The lifetime risk of melanoma is 1.7% for US white men and 1.3% for white women,⁷ which means that the probability of melanoma at any given point is less than 1%. Thus, for the US patient at average risk, the prior probability of melanoma can be conservatively taken as 1%, although the true point prevalence is likely much lower. Other populations have different risks. For example, the incidence of melanoma in the United States was 13.3 per 100 000 in 1995 vs 55 per 100 000 in Queensland, Australia.⁴

POPULATION FOR WHOM MELANOMA SCREENING SHOULD BE CONSIDERED

- Atypical (dysplastic) nevi
- Family history of melanoma
- Personal history of melanoma
- Multiple nevi
- Fair-skinned patients, especially those with sun sensitivity or proclivity to sunburns
- Large congenital nevus
- Immunosuppression

CLINICAL SCENARIO—RESOLUTION

LESION 1

This lesion is symmetric (A), has a discrete border (B), is of one color (C), is small (D), and has not enlarged (E) according to the patient. Thus, it has none of the ABCDE characteristics and the likelihood ratio (LR) for melanoma is low (0.07).

LESION 2

This lesion is asymmetric (A), has an irregular border (B), is of at least 2 colors (C), is larger than 6 mm (D) (although this might not be apparent from the photomicrograph), and is the one the patient believes might have changed size (E). Only 1 of the ABCDE criteria would have justified an appropriate biopsy, but this lesion fulfills all criteria and has a LR of 98 for melanoma.

Acknowledgment

Dr Grichnik is a founder and major shareholder in Digital Derm, Inc (MoleMapCD; total body photography) and has consulting and grant support from Electro-Optical-Systems, Inc (Melafind; melanoma detection device).

DETECTING THE LIKELIHOOD OF MELANOMA

See Table 29-9.

Table 29-9 Likelihood Ratios for Any One Finding From the ABCD(E) Criteria and the Individual Findings

Finding	LR+ (95% CI)	LR- (95% CI)
Any one finding from the ABCD criteria ^a	1.5 (1.4-1.6)	0.07 (0.04-0.13)
A (asymmetry)	2.1 (1.9-2.5)	0.59 (0.52-0.66)
B (border)	2.1 (1.8-2.4)	0.59 (0.53-0.67)
C (color)	1.6 (1.5-1.8)	0.59 (0.52-0.68)
D (dimension)	2.3 (2.1-2.5)	0.17 (0.13-0.22)
E (enlargement)	11 (8.5-14)	0.18 (0.15-0.22)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

^aThe LR+ increases rapidly with the number of abnormalities. Zero positive findings has a very low LR.

REFERENCE STANDARD TESTS

Biopsy with histopathology.

REFERENCES FOR THE UPDATE

- Thomas L, Tranchand P, Berard F, Secchi T, Colin C, Moulin G. Semiological value of the ABCDE criteria in the diagnosis of cutaneous pigmented tumors. *Dermatology* 1998;197(1):11-17.^a
- Chen SC, Bravata DM, Weil E, Olkin I. A comparison of dermatologists' and primary care physicians' accuracy in diagnosing melanoma: a systematic review. *Arch Dermatol*. 2001;137(12):1627-1634.
- Mayer J. Systematic review of the diagnostic accuracy of dermatoscopy in detecting malignant melanoma. *Med J Aust*. 1997;167(4):206-210.

4. Berg AO; US Preventive Services Task Force. Screening for skin cancer. *Am J Prev Med.* 2001;20(S3):44-46.
5. Feightner JW. Prevention of skin cancer. In: Canadian Task Force on the Periodic Health Examination, ed. *Canadian Guide to Clinical Preventive Health Care.* Ottawa, Ontario, Canada: Health Canada; 1994:850-859. http://www.ctfphc.org/Full_Text/Ch70full.htm. Accessed April 30, 2008.
6. World Health Organization. Ultraviolet radiation and the INTERSUN Programme. <http://www.who.int/uv/faq/skincancer/en/index2.html>. Accessed June 2, 2008.
7. Helfand M, Mahon SM, Eden KB, Frame PS, Orleans CT. Screening for skin cancer: a summary of the evidence. *Am J Prev Med.* 2001;20(S3):47-58.

^aFor the Evidence to Support the Update for this topic,
see <http://www.JAMAEvidence.com>.

EVIDENCE TO SUPPORT THE UPDATE:

29

Melanoma

TITLE Semiological Value of the ABCDE Criteria in the Diagnosis of Cutaneous Pigmented Tumors.

AUTHORS Thomas L, Tranchand P, Berard F, Secchi T, Colin C, Moulin G.

CITATION *Dermatology*. 1998;197(1):11-17.

QUESTION What is the effect of adding the (E) criteria (enlargement) to the traditional ABCD criteria for melanoma?

DESIGN All data were collected prospectively on consecutive patients undergoing a biopsy for a pigmented lesion.

SETTING Dermatology department. All patients were examined by dermatologists.

PATIENTS From the database, all patients with melanoma and prospectively recorded data ($n = 460$) were analyzed, along with 680 patients with the same prospectively recorded data who were found to have nonmalignant melanocytic tumors.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

A standardized form with 86 items was recorded for each patient. All patients had histopathology done after the examination was recorded.

- A. Geometrical asymmetry in 2 axes of the pigmented tumor
- B. Irregular borders
- C. At least 2 colors with the exception of darkening in the central lesion
- D. Diameter greater than or equal to 6 mm
- E. Enlargement of the surface (not height) as reported by the patient

MAIN OUTCOME MEASURES

Each item in the ABCDE criteria was assessed independently. In addition, a score consisting of the sum of positive results was compared with the reference standard.

MAIN RESULTS

See Tables 29-10, 29-11, and 29-12.

Table 29-10 Likelihood Ratio of the Components of the ABCDE Scale for Melanoma

Test	LR+ (95% CI)	LR- (95% CI)	DOR (95% CI)
A (asymmetry)	2.1 (1.9-2.5)	0.59 (0.52-0.66)	3.7 (2.9-4.7)
B (border)	2.1 (1.8-2.4)	0.59 (0.53-0.67)	3.5 (2.7-4.4)
C (color)	1.6 (1.5-1.8)	0.59 (0.52-0.68)	2.8 (2.2-3.5)
D (dimension)	2.3 (2.1-2.5)	0.17 (0.13-0.22)	14 (9.7-19)
E (enlargement)	11 (8.5-14)	0.18 (0.15-0.22)	60 (41-88)

Abbreviations: CI, confidence interval; DOR, diagnostic odds ratio; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

Table 29-11 Serial Likelihood Ratios for the Number of Positive Findings From the ABCDE Scale for Melanoma

No. of Positive Findings From ABCDE Criteria	LR (95% CI) ^a
5 Positive	98 (31-303)
≥4	8.3 (6.2-11)
≥3	3.3 (2.8-3.9)
≥2	2.6 (2.3-2.9)
≥1	1.5 (1.4-1.6)
0	0.07 (0.04-0.13)

Abbreviations: CI, confidence interval; LR, likelihood ratio.

^aThe area under the curve for these data is 0.85 (SE, 0.01).

Table 29-12 Likelihood Ratio of the Components of the ABCDE Scale for Melanoma or Atypical Dysplastic Nevus

Test	LR+ (95% CI)	LR- (95% CI)	DOR (95% CI)
A (asymmetry)	2.7 (2.3-3.2)	0.52 (0.47-0.59)	5.1 (4.0-6.6)
B (border)	2.6 (2.2-3.0)	0.52 (0.47-0.59)	4.9 (3.8-6.4)
C (color)	2.0 (1.7-2.2)	0.49 (0.43-0.57)	4.0 (3.1-5.1)
D (dimension)	2.5 (2.2-2.8)	0.17 (0.13-0.22)	15 (10-20)
E (enlargement)	13 (9.2-17)	0.25 (0.21-0.29)	51 (35-76)

Abbreviations: CI, confidence interval; DOR, diagnostic odds ratio; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

CONCLUSIONS

LEVEL OF EVIDENCE

Level 1.

STRENGTHS Large data set with all variables entered prospectively before biopsy results. Clinicians who agreed on definitions of criteria tested themselves with pictures before study (data on interobserver variability not provided). Data are provided for both melanomas and atypical nevi.

LIMITATIONS All examiners were dermatologists. Data are not provided on the numbers of patients who had pigmented lesions who did not undergo biopsy.

The researchers in this study modified the previously suggested “E” criteria from “elevated” to “enlarged.” By definition, enlargement is self-reported by the patient and is the only historical item among the criteria. As single variables, the most important are size greater than or equal to 6 mm (D) and enlargement (E).

The analysis of “number” of positive findings is informative. Clinicians can decide the level that they are willing to accept to justify a biopsy. A strategy to biopsy all patients with even 1 positive finding will lead to a diagnosis in 97% of patients, but there will be numerous biopsies for nonmalignant melanocytic tumors—in many clinical settings, this is not an acceptable strategy. The data are also interesting in that they allow inferences about the independence of find-

ings. Multiplying the positive LRs for each finding gives an LR of 179 compared to the actual value of 98 for 5 positive findings. These LRs produce similar effects on the posterior probability and suggest that the presence of findings confers independent information. On the other hand, serially multiplying the negative LRs gives an LR of 0.006; that value on a logarithmic scale is much lower than the actual LR of 0.07 when all 5 findings are absent. Nonetheless, an LR of 0.07 is low and will rule out melanoma for many patients.

Many primary care clinicians would use a more pragmatic reference standard than melanoma alone because atypical (dysplastic) nevi are markers for melanoma risk. When the data are compared with a reference standard that considers either melanoma or atypical nevi as “positive,” the operating characteristics of almost every finding improve (the positive LR increases and the negative LR decreases). Given what appears to be independence for the presence of the ABCDE criteria, this further justifies using the presence of only 1 criterion as an indication for biopsy or referral to a dermatologist. Because the criteria are not efficient at distinguishing melanoma from atypical nevi, this may lead to an increase in the removal of nonmalignant atypical nevi.

Reviewed by David L. Simel, MD, MHS, and James M. Grichnik, MD, PhD

CHAPTER 30

Does This Adult Patient Have Acute Meningitis?

John Attia, MD, PhD

Rose Hatala, MD, MSc

Deborah J. Cook, MD, MSc

Jeffrey G. Wong, MD

CLINICAL SCENARIOS

CASE 1 A 30-year-old man presents to the emergency department with a 24-hour history of chills and a stiff neck. On clinical examination, he is afebrile and has normal mental status. He can fully flex his neck, although he complains of pain over his cervical spine when doing so. Kernig and Brudzinski signs are absent.

CASE 2 A previously healthy 70-year-old woman presents to the emergency department with a 3-day history of fever, confusion, and lethargy. She is unable to cooperate with a full physical examination, but she has neck stiffness on neck flexion. The findings from a chest radiograph and urinalysis are normal.

WHY IS CLINICAL EXAMINATION IMPORTANT?

If, in a fever, the neck be turned awry on a sudden, so that the sick can hardly swallow, and yet no tumour appear, it is mortal.

—Hippocrates, Aphorism XXXV

As early as the fifth century BC, clinicians recognized the seriousness of infectious meningitis.¹ In the 20th century, the annual incidence of bacterial meningitis ranges from approximately 3 per 100 000 population in the United States² to 45.8 per 100 000 in Brazil³ to 500 per 100 000 in the “meningitis belt” of Africa.⁴ In one county in Minnesota, there was an incidence rate of viral meningitis of 10.9 per 100 000 person-years from 1950 to 1981, with most cases occurring in the summer months.⁵

Despite the availability of antimicrobial therapy, meningitis-related case fatality rates remain high, with a 17% all-cause mortality rate between 1980 and 1988 reported for community-acquired and nosocomial bacterial meningitis among patients aged 16 years and older.⁶ Among previously healthy patients who survive pneumococcal meningitis, up to 18% may experience long-term sequelae, including dizziness, excessive fatigue, and gait ataxia.⁷ Clinical signs and symptoms at presentation may predict prognosis.⁸ Thus, early clinical recognition of meningitis is imperative to allow clinicians to efficiently complete further investigations and initiate appropriate therapy, with a goal of minimizing these adverse outcomes.

The purpose of this systematic review is to provide clinicians with an understanding of the literature from which the current clinical approach to meningitis is derived. Optimal use of the clinical examination aids physicians in identifying patients at sufficient risk for meningitis to require further definitive diagnostic testing with a lumbar puncture (LP). Patients in whom meningitis is suspected require this invasive procedure to effectively establish or refute the diagnosis. In addition, evaluation of the cerebrospinal fluid (CSF) may

help direct antimicrobial therapy.⁹ To avoid unnecessary invasive procedures, identifying clinical features that could distinguish patients at high and low risk of meningitis would be useful. Clinical findings with a high specificity will assist clinicians in the decision to proceed to LP. Conversely, clinical findings with a high sensitivity will aid clinicians in deciding against invasive investigation, particularly for patients for whom the clinical suspicion of meningitis is relatively low.

This systematic review will focus on the features of history taking and physical examination that clinicians use to identify adult, immunocompetent patients at risk for acute meningitis for whom further diagnostic testing is indicated. We use the term *meningitis* to refer to acute infections of the meninges of either bacterial or viral origin.

Pathophysiology of Meningitis

The brain is protected from infection by the skull; the pia, arachnoid, and dural meninges covering its surface; and the blood-brain barrier. When any of these defenses are breached by a pathogen, infection of the meninges and subarachnoid space results in meningitis. Predisposing factors for the development of community-acquired meningitis include preexisting diabetes mellitus, otitis media, pneumonia, sinusitis, and alcohol abuse.⁶

The clinical features of meningitis are a reflection of the underlying pathophysiologic processes (Table 30-1). Systemic infection generates nonspecific findings such as fever, myalgia, and rash. Once the blood-brain barrier is breached, an inflammatory response within the CSF occurs. The resultant meningeal inflammation and irritation elicit a protective reflex to prevent stretching of the inflamed and hypersensitive nerve roots, which is detectable clinically as neck stiffness or Kernig or Brudzinski signs.^{10,11} The meningeal inflammation may also cause headache and cranial nerve palsies. Elevated intracranial pressure, altered mental status, vomiting, and seizures may ensue.

Examination for the Signs and Symptoms of Meningitis

The classic clinical presentation of acute meningitis is the triad of fever, neck stiffness, and an altered mental state. However, less than two-thirds of patients present with all 3 clinical findings.⁶ While taking the patient's history, clinicians suspecting meningitis will examine for general symptoms of infection (such as fever, chills, and myalgias), as well as

symptoms suggesting central nervous system infection (photophobia, headache, nausea and vomiting, focal neurologic symptoms, or changes in mental status).

The physical examination must include checking the vital signs and a brief mental status examination. General inspection may reveal a rash. In patients with severe meningeal irritation, the patient may spontaneously assume the tripod position (also called Amoss sign or Hoyne sign), sitting on the edge of the bed with the knees and hips flexed, the back arched lordotically, the neck extended, and the arms brought back to support the thorax.¹²

Physical examination specifically for meningitis includes assessing neck stiffness, testing for Kernig and Brudzinski signs, and assessing jolt accentuation of headache. Neck stiffness is assessed by examining the neck for rigidity by gentle forward flexion, with the patient in the supine position.

Like neck stiffness, Kernig and Brudzinski signs also indicate meningeal irritation. Vladimir Kernig, a Russian physician, first published the description of the sign that bears his name in 1884,^{10,13} although the sign had been previously described by Lazarevic in 1880 and by Forst in 1881.¹² In Kernig's original description, when patients sat on the edge of a bed with their legs dangling, an attempt to extend the knee joint more than 135 degrees, or in severe cases more than 90 degrees, elicited spasm of the extremity that disappeared when the patients lay supine or stood up. Today, the maneuver is most commonly performed with the patient lying supine and the hip flexed at 90 degrees. A positive sign is present when extension of the knee from this position elicits resistance or pain in the lower back or posterior thigh.

In 1909, Josef Brudzinski, a Polish physician, described many meningeal signs in children.^{10,14} His best-known "nape of the neck" sign (Brudzinski sign) is present when passive neck flexion in a supine patient results in flexion of the knees and hips. A separate sign, the contralateral reflex, is present if passive flexion of the hip and knee causes flexion of the contralateral leg.

An additional maneuver in assessing for meningitis is to elicit jolt accentuation of the patient's headache by asking the patient to turn his or her head horizontally at a frequency of 2 to 3 rotations per second. Worsening of a baseline headache represents a positive sign.¹⁵

A complete neurologic examination follows these more specific tests for meningitis, including examination of the cranial nerves, the motor and sensory systems, and reflexes and testing for Babinski reflex. A general examination follows, with an emphasis on the ears, sinuses, and respiratory system.

Table 30-1 Pathophysiology of Clinical Findings in Meningitis

Pathophysiology	Clinical Features
Systemic infection	Fever, myalgia, rash
Meningeal inflammation	Neck stiffness, Kernig sign, Brudzinski sign, jolt accentuation of headache, cranial nerve palsies
Cerebral vasculitis secondary to meningeal inflammation	Focal neurologic abnormalities, seizures
Elevated intracranial pressure secondary to meningeal inflammation and cerebral edema	Change in mental status, headache, cranial nerve palsies, seizures

METHODS

Literature Search and Selection

We searched MEDLINE for articles published from 1966 to July 1997, using a structured search strategy (available from the authors on request) to retrieve English- and French-language articles describing the precision and accuracy of the clinical examination in the diagnosis of meningitis. This

search strategy yielded 139 abstracts, which were reviewed by one of us (J.A.) for relevance. Full-text articles were retrieved for abstracts that potentially met the inclusion criteria. Additional references were identified by searching the reference lists of pertinent articles.

Explicit inclusion and exclusion criteria were applied to the retrieved articles. We included articles that were original studies describing the accuracy or precision of the clinical examination in the diagnosis of meningitis in which most patients had objectively confirmed bacterial or viral meningitis. We excluded studies that enrolled only children or immunocompromised adults, described mixed patient populations from which adult data could not be extracted, or focused only on metastatic meningitis or meningitis of a sin-

gle specific microbial origin (ie, *Listeria meningitidis* or *Mycobacterium tuberculosis*). Tuberculous meningitis was also excluded on the grounds that this infection is more prevalent in patients with human immunodeficiency virus infection¹⁶ and in children, neither of which represents our target population. However, we retained in our analyses 2 studies^{15,17} in which there were insufficient data to separate the patients with tuberculous meningitis (Table 30-2).

Study Characteristics

This systematic review differs from previous Rational Clinical Examination articles in that all but 1 article¹⁷ of the 9 articles^{6,17-24} that met our inclusion criteria were retrospective

Table 30-2 Studies Assessing Clinical Presentation of Patients

Source, y	Clinical Setting, Years	No. of Patients	Age, y, Mean (Range)	Type of Meningitis ^a	Patient Identification	Clinical Findings Defined
Sigurdardottir et al, ¹⁸ 1997	All hospitals in Iceland, 1975-1994	119	44% >45	Bacterial	All patients with bacterial isolates from cerebrospinal fluid or meningococcmia, processed at national central laboratory, complete hospital records for 119 of 132 patient episodes	No
Durand et al, ⁶ 1993	University hospital, 1962-1988	259	56% >50 (16-88) ^b	Bacterial	Hospital diagnosis of acute bacterial meningitis, including transferred patients	No
Uchihara and Tsukagoshi, ¹⁵ 1991 ^c	General hospital, dates not specified	34	38.6 (15-71)	Aseptic (n = 28), bacterial/tuberculous (n = 1), other ^d	Patients presenting to outpatient or emergency department with headache and fever	Yes
Genton and Berger, ¹⁹ 1988	University hospital, 1977-1982	112	Women, 41; men, 40 (16-89)	Bacterial	Patients admitted and discharged with a diagnosis of meningitis	No
Gorse et al, ²⁰ 1984 ^e	University and Veterans Affairs hospitals, 1970-1982	54	64 (50-95)	Bacterial	Patients with a discharge diagnosis of meningitis	No
Gorse et al, ²⁰ 1984 ^e	University hospital, 1970-1982	32	(15-49) ^f	Bacterial	Patients with a discharge diagnosis of meningitis	No
Massanari, ²¹ 1977	University hospital, 1965-1975	17	>65 ^g	Bacterial	Patients with a chart diagnosis of meningitis	No
Magnussen, ²² 1980	Community hospital, 1969-1978	59	39 ^h	Aseptic (n = 34), bacterial	Patients with a discharge diagnosis of acute meningitis	No
Domingo et al, ²³ 1990	Hospital, 1974-1988	59	71 (65-87)	Bacterial	Not indicated	No
Behrman et al, ²⁴ 1989	University hospital, 1970-1985	31	72 (65-89)	Aseptic (n = 4), bacterial	Patients with a discharge diagnosis of meningitis, subdural empyema, brain abscess, or epidural abscess	Yes
Rasmussen et al, ¹⁷ 1992	Community hospitals, 1976-1988	48	69 ⁱ (60-88) ^j	Tuberculous (n = 6), bacterial	Computer search of hospital records for patients with a diagnosis of acute bacterial meningitis	No

^aInfections included in calculations of sensitivities for clinical findings.

^bCommunity-acquired meningitis.

^cProspective study design, assessing clinical findings compared with cerebrospinal fluid pleocytosis in patients presenting with headache and fever.

^dPredominantly aseptic meningitis (28/54 patients). Other includes subarachnoid hemorrhage (n = 2), acute monocytic leukemia (n = 1), Sjögren syndrome (n = 1), upper respiratory tract infection (n = 11), infectious diarrhea (n = 3), edentulous (n = 2), glaucoma (n = 1), and not specified (n = 3).

^eTwo patient groups were included in this study: 54 patients older than age 50 years and 32 patients aged between 15 and 49 years. Each age group is reported separately.

^fMean age not reported.

^gMean age and range not reported.

^hMean age calculated from data in study; range not reported.

ⁱMedian age and range.

chart reviews. These studies assessed the clinical presentation of a total of 845 patient episodes (824 patients), in patients aged 16 to 95 years, with meningitis confirmed by LP or autopsy (Table 30-2).

Because no quality grading system for chart reviews has been widely established, we assessed the validity of these studies by critically appraising several components of the study design (Table 30-2). These components included an assessment of the reference standard used to diagnose meningitis (LP or autopsy), the completeness of patient ascertainment, and whether the clinical examination was described in sufficient detail to be reproducible. The major limitation common to all these studies was the lack of a control population, which means that only sensitivities were available for most of the clinical findings. In addition, the reported sensitivities may overestimate the true sensitivities (as could be established in a prospective study) because the clinical examinations recorded in the charts could have been performed with knowledge of the LP results.

The single prospective study included 54 inpatients and outpatients presenting with fever and headache to a Japanese center (Table 30-2).¹⁵ A standardized clinical examination was performed by an examiner before LP was undertaken, and clinical findings were compared with those of CSF pleocytosis.

Data Analysis

Clinical examination findings that differ between viral and bacterial causes are explicitly indicated. Sensitivities for the various signs and symptoms of meningitis were calculated from the data in each study. Pooled sensitivities were calculated for each feature of the clinical examination, using a random-effects model.²⁵

Because control groups of patients without meningitis were not included in the 9 retrospective studies, specificities for many features of the clinical examination were unavailable.

For the findings assessed in the prospective study, specificities and likelihood ratios (LRs) were calculated and included.¹⁵

RESULTS

Precision of Symptoms and Signs of Meningitis

Data on the precision of the clinical examination for meningitis were not available from the retrospective studies. In the prospective study, a single clinician completed all clinical examinations.¹⁵

Accuracy of the Clinical History in the Diagnosis of Meningitis

The individual components of the clinical history have low sensitivity for the diagnosis of meningitis, as indicated in Table 30-3. In addition to symptoms of headache and nausea and vomiting, neck pain was reported to have a sensitivity of 28% among patients with meningitis.²⁰ Data from the prospective trial suggest that the clinical history also lacks specificity for the diagnosis of meningitis, with reported specificities of 15% for a nonpulsatile headache, 50% for a generalized headache, and 60% for nausea and vomiting.¹⁵ Thus, clinical history alone is not useful in establishing a diagnosis of meningitis. The inaccuracy of the clinical history may relate to the frequently impaired mental status of patients with meningitis (pooled sensitivity, 67%; 95% confidence interval [CI], 52%-82%) (Table 30-4), who are relatively incapable of providing an accurate clinical history.^{21,22}

Accuracy of the Physical Examination in the Diagnosis of Meningitis

In contrast to the clinical history, elements of the physical examination have sensitivities that are clinically useful. The frequency with which patients presented with the classic clinical triad of fever, neck stiffness, and a change in mental sta-

Table 30-3 Sensitivity of Clinical History in the Diagnosis of Meningitis

Source, y	No. of Patient Episodes	Headache, %	Nausea and Vomiting, %	Neck Pain, %
Uchihara and Tsukagoshi, ¹⁵ 1991 ^a	34	27	32	NA
Gorse et al, ²⁰ 1984 ^b	54	43	30	28
Massanari, ²¹ 1977	17	41	NA	NA
Magnussen, ²² 1980	59	78	NA	NA
Domingo et al, ²³ 1990	59	81	NA	NA
Behrman et al, ²⁴ 1989	32 ^c	31	NA	NA
Rasmussen et al, ¹⁷ 1992	48	46	29	NA
Pooled sensitivity (95% confidence interval)	50 (32-68) [n = 303] ^d	30 (22-38) [n = 136] ^d	NA	

Abbreviation: NA, not assessed.

^aOnly study patients with pleocytosis were included in the calculation of sensitivity.

^bData reported only for patients older than 50 years.

^cThirty-one patients with 32 patient episodes.

^dNumber in brackets is patients included in calculation of sensitivity.

Table 30-4 Sensitivity of the Physical Examination in the Diagnosis of Meningitis^a

Source, y	No. of Patient Episodes	Fever	Neck Stiffness	Altered Mental Status	Fever, Neck Stiffness, and Altered Mental Status	Focal Neurologic Findings ^b	Rash	Kernig Sign	Jolt Accentuation of Headache
Sigurdardottir et al, ¹⁸ 1997	119	97	82	66	51	10	52	NA	NA
Durand et al, ⁶ 1993	279 ^c	95	88	78	66	29	11	NA	NA
Uchihara and Tsukagoshi, ¹⁵ 1991 ^d	34	71	15	NA	NA	NA	NA	9 ^e	97 ^f
Genton and Berger, ¹⁹ 1988	112	NA	NA	32	NA	10	NA	NA	NA
Gorse et al, ²⁰ 1984 ^g	54	91	81	89	NA	39	NA	NA	NA
Gorse et al, ²⁰ 1984 ^g	32	75	66	53	NA	22	NA	NA	NA
Massanari, ²¹ 1977	17	88	76	88	NA	NA	NA	NA	NA
Magnussen, ²² 1980	59	42	81	20 ^h	NA	10	NA	NA	NA
Domingo et al, ²³ 1990	59	95	92 ⁱ	88	NA	37	NA	NA	NA
Behrman et al, ²⁴ 1989	32 ^j	94	59	88	18 ^k	38	NA	NA	NA
Rasmussen et al, ¹⁷ 1992	48	79	54	69	NA	21	4	NA	NA
Pooled sensitivity (95% confidence interval)	85 (78-91) (n = 733)	70 (58-82) (n = 733)	67 (52-82) (n = 811)	46 (22-69) (n = 426)	23 (15-31) (n = 794)	22 (1-43) (n = 446)			

Abbreviation: NA, not assessed.

^aAll data are presented as percentage unless otherwise noted.

^bFocal neurologic findings include bilateral Babinski reflexes, pupillary abnormalities, hemiparesis, cranial nerve abnormalities, nystagmus, convulsion or seizure, and tremor.

^cThere were 279 patient episodes in 259 patients.

^dOnly study patients with pleocytosis were included in the calculation of sensitivity.

^eSpecificity of 100%; Brudzinski sign was not assessed.

^fSpecificity of 60%.

^gTwo patient groups were included in this study: 54 patients older than 50 years and 32 patients aged between 15 and 49 years. Sensitivities were calculated separately for each age group.

^hModerate or severe alteration in mental status.

ⁱAuthors refer to this clinical finding as *meningeal signs*.

^jThirty-two patient episodes in 31 patients.

^kFor this triad, assessed only in patients (n = 28) with bacterial meningitis. The authors of this study described the triad of symptoms as fever, neck stiffness, and headache.

tus (or headache²⁴) was assessed in 3 studies. Although the pooled sensitivity for the presence of all 3 symptoms was low (Table 30-4), 95% of patients had 2 or more symptoms,²⁴ and 2 studies reported that between 99% and 100% of patients had at least 1 of these clinical findings.^{6,18} Thus, the diagnosis of meningitis may be effectively eliminated in adult patients presenting without any of the symptoms of fever, neck stiffness, or a change in mental status.

As indicated in Table 30-4, documentation of fever has a pooled sensitivity of 85% (95% CI, 78%-91%) for the diagnosis of meningitis. As would be expected of a single physical finding common to many disorders, fever has a low specificity of 45%.¹⁵ Normal body temperature may significantly decrease the likelihood that a patient has meningitis, although the presence of a fever does not definitively establish the disease. The relationship between body temperature and meningitis may be U-shaped because hypothermic patients with sepsis are more likely to be severely ill than normothermic patients.²⁶

Neck stiffness is also a relatively useful clinical finding, with a pooled sensitivity of 70% (95% CI, 58%-82%) (Table 30-4). Other signs of meningeal irritation, namely, Kernig and Brudzinski signs, have not been well studied, although in Brudzinski's original description of 42 cases of meningitis (including 21 cases of tuberculous meningitis), Kernig sign had a sensitivity of 57%, whereas Brudzinski's nape of the neck sign had a sensitivity of 97% and the contralateral reflex sign had a sensitivity of 66%.¹⁰ Brudzinski himself claimed to confirm the specificity of his nape of the neck sign by attempting (and failing) to elicit it in children with other neurologic conditions.¹⁰ The Uchihara and Tsukagoshi¹⁵ prospective study of younger adult patients (mean age, 39 years) reported a sensitivity of 9% and a specificity of 100% for the Kernig sign, whereas neck stiffness had a sensitivity of 15% and a specificity of 100%. Because this study enrolled patients presenting with fever and headache and excluded those with mental status abnormalities or focal neurologic findings, the low reported sensitivities may result from

excluding patients with the highest likelihood of having meningeal signs.

Considering that these signs of meningeal irritation have been in use for almost a century, assessment of their accuracy has been limited. Indirect evidence of poor specificity comes from a case series of 74 acute-care and 287 geriatric patients (hospitalized patients in the acute-care or rehabilitation geriatric wards) aged 17 to 92 years.²⁷ Puxty et al²⁷ found that 13% of the acute-care patients and 35% of the geriatric patients had nuchal rigidity despite the absence of meningitis. Kernig sign was present in 1.5% of the acute-care and 12% of the geriatric populations. The low specificity of the meningeal signs may be caused by the frequent presence of cervical arthritis and spondylosis among older patients. Clearly, a well-designed prospective study in which patients suspected of having meningitis are observed prospectively is necessary to definitively establish the accuracy of meningeal signs.

Alterations in mental status, ranging from confusion to coma, have a pooled sensitivity of 67% (95% CI, 52%-82%) (Table 30-4), indicating that normal mental status may be helpful in ruling out meningitis in low-risk patients. One study directly comparing aseptic with bacterial meningitis reported that moderate to severe mental status abnormalities were more common in patients with bacterial meningitis than with aseptic meningitis (44% vs 3%, respectively).²² Similarly, a second study reported that all patients with bacterial meningitis had a change in mental status, whereas none of the aseptic meningitis patients did.²⁴

One of the most sensitive maneuvers in the diagnosis of meningitis is jolt accentuation of headache, as described by Uchihara and Tsukagoshi.¹⁵ Of 34 patients with pleocytosis in this study, 30 had meningitis and 4 had other conditions. Jolt accentuation of headache was present in 33 of these patients compared with 8 of 20 patients without pleocytosis, yielding a sensitivity of 97% and a specificity of 60%. The associated positive likelihood ratio (LR+) was 2.4, and the negative likelihood ratio (LR-) was 0.05. If we calculate the LRs specifically for those patients with meningitis, we obtain a sensitivity of 100%, a specificity of 54%, an LR+ of 2.2, and an LR- of 0. In patients presenting with fever and headache, a lack of jolt accentuation of headache on physical examination may essentially exclude meningitis. The main limitation to widespread application of these results is the small sample of patients assessed in this study.

Rashes occurred most frequently in the presentation of meningitis due to *Neisseria meningitidis*, with prevalences of 63%⁶ and 80%.¹⁷ A petechial rash occurred in 73% of patients with meningococcemia, whereas purpura was described in only 20% of these patients.⁶ Petechial, purpuric, and ecchymotic rashes also occurred, with lower frequency, in infections caused by *Haemophilus influenzae*, *Streptococcus pneumoniae*, and *L moncytogenes*. Because the overall incidence of *N meningitidis* among patients with community-acquired bacterial meningitis was low (14% in 1 series⁶), the pooled sensitivity of a rash for the diagnosis of meningitis was poor (Table 30-4).

One or more focal neurologic abnormalities were described in many of the case series, including bilateral

Babinski reflexes, pupillary abnormalities, hemiparesis, cranial nerve abnormalities, nystagmus, convulsion or seizure, and tremor. As summarized in Table 30-4, the pooled sensitivity for these signs is low, and they are not clinically useful in ruling out meningitis.

CLINICAL SCENARIOS—RESOLUTIONS

The first scenario described a 30-year-old man with chills, who complained of a stiff neck but had no fever or meningeal signs on examination. We would ask the patient about a headache, and, if present, assess for jolt accentuation. His lack of fever, normal mental status, and lack of jolt accentuation would be sufficient to assure us that this patient does not have meningitis.

In the second scenario, a 70-year-old woman presented with fever, confusion, and neck stiffness. Although we do not know the specificity of these findings, their presence causes us to suspect that she may have meningitis. To establish or refute the diagnosis in this scenario, we would proceed to definitive testing by LP.

THE BOTTOM LINE

Assessment of the accuracy of the clinical examination in the diagnosis of meningitis is severely limited by the paucity of prospective data on this topic. Despite classic descriptions of meningeal signs and sweeping statements about clinical presentations in generations of textbooks, the signs and symptoms of meningitis have been inadequately studied, and the conclusions of this systematic review are that more prospective research is required. According to the limited studies included in this systematic review, we suggest the following to make optimal use of the clinical examination.

1. The absence of all 3 signs of the classic triad of fever, neck stiffness, and an altered mental status virtually eliminates a diagnosis of meningitis.
2. Fever is the most sensitive of the classic triad of signs of meningitis and occurs in a majority of patients, with neck stiffness the next most sensitive sign. Alterations in mental status also have a relatively high sensitivity, indicating that normal mental status helps to exclude meningitis in low-risk patients. Changes in mental status are more common in bacterial than viral meningitis.
3. Among the signs of meningeal irritation, Kernig and Brudzinski signs appear to have low sensitivity but high specificity.
4. Jolt accentuation of headache may be a useful adjunctive maneuver for patients with fever and headache. In patients at sufficient risk of meningitis, a positive test result may aid in the decision to proceed to LP, whereas a negative test result essentially excludes meningitis.

Author Affiliations at the Time of the Original Publication

Departments of Medicine (Drs Attia, Hatala, and Cook) and Clinical Epidemiology and Biostatistics (Dr Cook), McMaster University, Hamilton, Ontario, Canada; and Division of

General Medical Sciences, Washington University School of Medicine, St Louis, Missouri (Dr Wong). Dr Wong is now with the Yale University School of Medicine, New Haven, Connecticut.

Acknowledgments

We thank Lauren Griffith, MSc, for assistance with the statistical analysis. We also thank Vance Fowler, MD, and Peter Margolis, MD, for their helpful comments on an earlier version of the manuscript.

REFERENCES

1. Sprengell C. *The Aphorisms of Hippocrates, and the Sentences of Celsus*. 2nd ed. London, England: R Wilkin; 1735.
2. Tunkel AR, Scheld WM. Pathogenesis and pathophysiology of bacterial meningitis. *Clin Microbiol Rev*. 1993;6(2):118-136.
3. Bryan JP, de Silva HR, Tavares A, Rocha H, Scheld WM. Etiology and mortality of bacterial meningitis in northeastern Brazil. *Rev Infect Dis*. 1990;12(1):128-135.
4. Scheld WM. Meningococcal diseases. In: Warren KS, Mahmoud AAF, eds. *Tropical and Geographical Medicine*. 2nd ed. New York, NY: McGraw-Hill; 1990:798-814.
5. Nicolosi A, Hauser WA, Beghi E, Kurland LT. Epidemiology of central nervous system infections in Olmsted County, Minnesota, 1950-1981. *J Infect Dis*. 1986;154(3):399-408.
6. Durand ML, Calderwood SB, Weber DJ, et al. Acute bacterial meningitis in adults: a review of 493 episodes. *N Engl J Med*. 1993;328(1):21-28.
7. Bohr V, Paulson OB, Rasmussen N. Pneumococcal meningitis: late neurologic sequelae and features of prognostic impact. *Arch Neurol*. 1984;41(10):1045-1049.
8. Aronin SI, Peduzzi P, Quagliarello VJ. Community-acquired bacterial meningitis: risk stratification for adverse clinical outcome and effect of antibiotic timing. *Ann Intern Med*. 1998;129(11):862-869.
9. Spanos A, Harrell FE, Durack DT. Differential diagnosis of acute meningitis: an analysis of the predictive value of initial observations. *JAMA*. 1989;262(19):2700-2707.
10. Brody IA, Wilkins RH. The signs of Kernig and Brudzinski. *Arch Neurol*. 1969;21(2):215-218.
11. O'Connell JEA. The clinical signs of meningeal irritation. *Brain*. 1946;69:9-21.
12. Sapiro JD. *The Art and Science of Bedside Diagnosis*. Baltimore, MD: Urban & Schwarzenberg Inc; 1990:469-470.
13. Kernig W. Ueber ein Wenig bemerktes Meningitis-Symptom. *Berl Klin Wochenschr*. 1884;21:829-832.
14. Brudzinski J. Un signe nouveau sur les membres inférieurs dans les meninges chez les enfants (signe de la nuque). *Arch Med Enfants*. 1909;12:745-752.
15. Uchihara T, Tsukagoshi H. Jolt accentuation of headache: the most sensitive sign of CSF pleocytosis. *Headache*. 1991;31(3):167-171.
16. Berger JR. Tuberculous meningitis. *Curr Opin Neurol*. 1994;7(3):191-200.
17. Rasmussen HH, Sorensen HT, Moller-Petersen J, Mortensen FV, Nielsen B. Bacterial meningitis in elderly patients: clinical picture and course. *Age Ageing*. 1992;21(3):216-220.
18. Sigurdardottir B, Bjornsson OM, Jonsdottir KE, Erlendsdottir H, Gudmundsson S. Acute bacterial meningitis in adults: a 20-year overview. *Arch Intern Med*. 1997;157(4):425-430.
19. Genton B, Berger JP. Les méningites de l'adulte au centre hospitalier universitaire vaudois: diagnostic, prise en charge et évolution de 112 cas. *Rev Med Suisse Romande*. 1988;108(9):795-803.
20. Gorse GJ, Thrupp LD, Nudleman KL, Wyle FA, Hawkins B, Cesario TC. Bacterial meningitis in the elderly. *Arch Intern Med*. 1984;144(8):1603-1607.
21. Massanari RM. Purulent meningitis in the elderly: when to suspect an unusual pathogen. *Geriatrics*. 1977;32(3):55-59.
22. Magnussen CR. Meningitis in adults: ten-year retrospective analysis at a community hospital. *N Y State J Med*. 1980;80(6):901-906.
23. Domingo P, Mancebo J, Blanch L, Coll P, Net A, Nolla J. Acute bacterial meningitis in the elderly. *Arch Intern Med*. 1990;150(7):1546-1547.
24. Behrman RE, Meyers BR, Mendelson MR, Sacks HS, Hirschman SZ. Central nervous system infections in the elderly. *Arch Intern Med*. 1989;149(7):1596-1599.
25. Fleiss JL. The statistical basis of meta-analysis. *Stat Methods Med Res*. 1993;2(2):121-145.
26. Lewin S, Brettman LR, Holzman RS. Infection in hypothermic patients. *Arch Intern Med*. 1981;141(7):920-925.
27. Puxty JAH, Fox RA, Horan MA. The frequency of physical signs usually attributed to meningeal irritation in elderly patients. *J Am Geriatr Soc*. 1983;31(10):590-592.

This page intentionally left blank

Prepared by Rose Hatala, MD, MSc,
John Attia, MD, PhD, and Jeffrey G. Wong, MD

Reviewed by Stephen Bent, MD

CLINICAL SCENARIO

A previously healthy 41-year-old woman presents to the emergency department with fever and a headache. She has had symptoms of an upper respiratory tract infection for the last week. During the previous 24 hours, she has developed a fever and a frontal headache. In the emergency department, her temperature is 38.5°C (101.3°F). On examination, she has neck stiffness and jolt accentuation of her headache but no neurologic abnormalities.

UPDATED SUMMARY ON MENINGITIS

Original Review

Attia J, Hatala R, Cook DJ, Wong JG. Does this adult patient have acute meningitis? *JAMA*. 1999;282(2):175-181.

UPDATED LITERATURE SEARCH

We replicated the original search strategy to identify articles on the diagnosis of meningitis. We searched MEDLINE for articles from 1996 to November 2004, written in English or French, that described the precision and accuracy of the clinical examination in the diagnosis of meningitis. Search terms included “meningitis” combined with “physical examination,” “medical history taking,” or “professional competence,” in addition to combining “meningitis” with “sensitivity and specificity” or “reproducibility of results.” Additional references were identified by searching the reference lists of pertinent articles.

NEW FINDINGS

- Additional studies, both prospective and retrospective, confirm that no single classic item of medical history or physical examination is sufficiently accurate to rule in or rule out meningitis.
- The absence of all items in the classic triad of fever, neck stiffness, and altered mental status is not sufficiently sensitive to rule out meningitis.

- Additional prospective studies are necessary to establish the accuracy of history and physical examination, including jolt accentuation of headache, in patients with suspected meningitis. Assessment of combinations of clinical findings may be more helpful than any individual item. However, more retrospective research will be of minimal value because such studies contain no specificity data.
- Patients with suspected meningitis may safely undergo lumbar puncture (LP) without previous CT head scan unless they have a decreased level of consciousness or focal neurologic findings, recent seizures or a history of central nervous system (CNS) disease, immunocompromised status, or age greater than 60 years.

Details of the Update

There continues to be a paucity of high-quality studies assessing the accuracy of clinical examination for the diagnosis of meningitis. One new higher-quality prospective article¹ was identified from 235 potentially relevant articles regarding the clinical examination for meningitis. Five additional retrospective cohort study articles²⁻⁶ were identified with similar design flaws to the previous literature on this topic. These 5 articles have not been individually summarized, but their data have been included in the updated summary tables. One new article⁷ was also identified from 41 potentially relevant articles regarding the safety of LP before computed tomography (CT) head scan in patients suspected of having meningitis. In updating the earlier data, which included 9 retrospective studies and 1 prospective study, we removed the one previous prospective study⁸ and separately combined its results in a discussion with the newer prospective study.

IMPROVEMENTS IN THE DATA PRESENTED IN THE ORIGINAL PUBLICATION

In updating the review, we were able to reconfigure the data so that the retrospective studies (sensitivity only, [Table 30-5](#)) are displayed separately from the prospectively collected data ([Table 30-6](#)). One new prospective study¹ provides

Table 30-5 Sensitivity of Findings for Meningitis in Adults, Retrospective Studies

Finding (No. of Combined Studies)	Sensitivity (95% CI)
Medical history	
Headache (11)	0.68 (0.55-0.79)
Nausea and vomiting (5)	0.52 (0.34-0.71)
Physical examination	
Fever (14)	0.87 (0.79-0.92)
Neck stiffness (13)	0.80 (0.74-0.85)
Altered mental status (15)	0.69 (0.57-0.79)
Classic triad (fever, neck stiffness, headache) (4)	0.46 (0.28-0.64)
Focal neurologic findings (12)	0.21 (0.15-0.29)
Rash (6)	0.13 (0.04-0.27)

Abbreviation: CI, confidence interval.

Table 30-6 Likelihood Ratios for Findings for Meningitis in Adults, Prospective Studies^{1,8}

Finding	Sensitivity (95% CI)	LR+ (95% CI)	LR- (95% CI)
Historical Findings			
Headache	0.92 (0.84-0.96)	1.1 (1.0-1.3)	0.43 (0.19-0.96)
Nausea/vomiting			
Thomas et al ¹	0.70 (0.59-0.79)	1.3 (1.1-1.6)	0.64 (0.44-0.92)
Uchihara and Tsukagoshi ⁸	0.32 (0.18-0.48)	0.81 (0.39-1.7)	1.1 (0.74-1.7)
Neck stiffness ¹		1.1 (0.82-1.4)	0.95 (0.74-1.2)
Physical Examination			
Fever ¹	0.43 (0.32-0.53)	0.82 (0.62-1.1)	1.2 (0.94-1.5)
Kernig sign			
Thomas et al ¹	0.05 (0.02-0.13)	0.97 (0.27-3.6)	1.0 (0.94-1.1)
Uchihara and Tsukagoshi ⁸	0.09 (0.02-0.21)	4.2 (0.23-77)	0.92 (0.81-1.0)
Brudzinski sign ¹	0.05 (0.02-0.13)	0.97 (0.26-3.5)	1.0 (0.94-1.1)
Neck stiffness			
Thomas et al ¹	0.30 (0.21-0.41)	0.94 (0.64-1.4)	1.0 (0.87-1.2)
Uchihara and Tsukagoshi ⁸	0.15 (0.06-0.28)	6.6 (0.38-113)	0.83 (0.74-1.0)
Jolt accentuation ⁸	0.97 (0.83-0.99)	2.4 (1.4-4.2)	0.05 (0.01-0.35)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

specificity estimates that allowed us to calculate likelihood ratios (LRs). Five additional retrospective studies²⁻⁶ narrowed the confidence intervals (CIs) around the pooled sensitivity estimates. These data confirmed that no single item of medical history or physical examination has sufficient sensitivity to rule out a diagnosis of meningitis. The new data, both retrospective and prospective, changed our previous conclusion regarding the classic triad of fever, neck stiffness, and headache to clarify that this triad is not sufficiently sensitive to rule out meningitis.

CHANGES IN THE REFERENCE STANDARD

The reference standard remains microbiologic culture.

RESULTS OF LITERATURE REVIEW

Retrospective Studies

Although the previous pooled sensitivities have been updated to include the 5 additional retrospective studies, most of the updated sensitivities did not have clinically important changes. As in our previous review, the major limitation common to these studies is the lack of a control population, such that only sensitivities are available.¹ There is significant heterogeneity among the individual study results that may be, in part, caused by varying definitions of meningitis (viral vs bacterial, positive cerebrospinal fluid [CSF] culture result vs absolute CSF white blood cell [WBC] count). Overall, these studies confirm that no single finding is of adequate sensitivity that its absence rules out meningitis. The addition of the newer studies to the previous pooled estimates has also clarified that the absence of the classic triad of fever, neck stiffness, and headache is not sufficiently sensitive to rule out meningitis, a conclusion that is different from that of our previous review. With the relatively narrow CIs around these pooled estimates, additional studies of retrospectively collected data on patients with meningitis are unlikely to change these conclusions.

Prospective Studies

Because the 2 prospective studies are the most rigorous to date, we believe the estimates from these studies are the most accurate. Uchihara and Tsukagoshi⁸ enrolled 54 patients (inpatients and outpatients) with fever and headache who were examined by 1 investigator before LP. Because fever and headache were inclusion criteria, they are not summarized in the table. In addition, neck stiffness and Kernig sign had a sensitivity of 100% ($n = 20$ patients), a finding that was not replicated in the larger study ($n = 297$ patients) by Thomas et al.¹

The study by Thomas et al¹ included patients presenting to the emergency department with “clinically suspected meningitis.” Unfortunately, the physical examination technique of the examining physicians was not standardized, a design flaw common in our previous review, so not all patients underwent all aspects of the clinical examination.

There are significant differences in the sensitivities calculated for the pooled retrospective studies compared with the prospective data. This largely reflects the inherent difficulty with the retrospective design wherein the clinician’s assessment or recording of the patient’s clinical findings may have occurred after receiving the LP results. However, the essential conclusion for this update remains: no single classic item of medical history or physical examination is sufficiently accurate to rule in or rule out the diagnosis of meningitis. Whereas previously the triad of fever, neck stiffness, and altered mental status appeared helpful in ruling out meningitis in low-risk patients, the LRs from the prospective studies associated with the absence of fever and neck stiffness on

physical examination all approach 1 and suggest that the triad will not be helpful in ruling out meningitis. The patient's symptoms might be more important than the signs. The absence of headache or nausea/vomiting has summary LRs that decrease a patient's pretest probability of meningitis but would not definitively rule it out. Although they report weak positive LRs for combinations of positive physical findings, Thomas et al¹ did not evaluate whether the combined absence of headache and nausea/vomiting provides more information than the individual findings. Jolt accentuation of headache (positive LR = 2.4; negative LR = 0.05), previously found to be helpful in the diagnosis of meningitis, was not assessed in the study by Thomas et al.¹

Future research evaluating the diagnostic value of clinical features suggesting meningitis should require prospective collection of data on consecutive patients suspected of having meningitis, with an adequate gold standard in all patients. Assessment of combinations of clinical findings, rather than individual historical and physical examination features, is more likely to lead to useful results. Because meningitis is a disease with potentially serious clinical consequences if missed, it is especially important to identify clinical examination findings (either alone or in combination) with near-perfect sensitivity and very low negative LRs that clinically rule out the disease. To date, studies have not identified any single finding or combination of clinical findings that fulfills this criterion.

Patients suspected of having meningitis require LP. A new study, using the same patient population as the prospective clinical examination study, assessed the necessity of a CT head scan before LP.⁷ The study demonstrated that for patients lacking specific baseline characteristics, it appears that LP can be safely performed without a CT head scan. The baseline characteristics associated with any abnormality on

CT head scan included age greater than 60 years, immunocompromised state, history of CNS disease, seizure within 1 week of presentation, and neurologic abnormality. The neurologic abnormalities were a decreased level of consciousness, inability to answer questions and follow commands, gaze palsy, abnormal visual fields, facial palsy, abnormal motor function, and abnormal language.

EVIDENCE FROM GUIDELINES

No federal guidelines address the diagnostic approach to meningitis for immunocompetent adults. The Centers for Disease Control and Prevention recommends routine meningococcal vaccination beginning in the pre-high school years (<http://www.cdc.gov/vaccines/vpd-vac/mening/vac-mening-fs.htm>; accessed June 3, 2008). Patients with meningitis symptoms should be asked whether they have been vaccinated. However, because the efficacy of the vaccine is less than 100% (although high) and does not cover all meningococcal strains, patients with symptoms should still be appropriately evaluated for meningitis.

CLINICAL SCENARIO—RESOLUTION

This patient has fever, headache, and neck stiffness. Her symptoms alone raise the possibility of meningitis, and her probability of meningitis is increased by the positive jolt accentuation of her headache. You decide to proceed to LP. She has none of the baseline characteristics associated with an abnormal CT head scan result, so you undertake her LP directly without obtaining a CT scan.

MENINGITIS—MAKE THE DIAGNOSIS

PRIOR PROBABILITY

Meningitis can occur sporadically or in outbreaks. It is impossible to come up with a single prior probability estimate for all patients with symptoms compatible with meningitis. Among patients presenting to the emergency department at a single US hospital with a clinical suspicion of meningitis who underwent LP, the prevalence of meningitis (CSF WBC $\geq 6/\text{mL}$) was 27%.¹ Among the patients in this study,¹ the prevalence of bacterial meningitis as defined by a positive CSF culture result was 1%. The rates of meningococcal meningitis are low (approximately 1 case/100000 persons each year; <http://www.cdc.gov/meningitis/tech-clinical.htm>; accessed June 3, 2008).

POPULATION FOR WHOM MENINGITIS SHOULD BE CONSIDERED

Among immunocompetent patients, meningitis should be considered for patients presenting with combinations of findings that include fever, headache, altered mental status, neck stiffness, or photophobia.

DETECTING THE LIKELIHOOD OF MENINGITIS

The most common symptoms associated with meningitis are not particularly useful when interpreted in isolation ([Table 30-7](#)).

Table 30-7 Likelihood Ratios of Headache and Nausea/Vomiting Are Not Highly Useful

	LR+ (95% CI)	LR- (95%CI)
Headache	1.1 (1.0-1.3) ¹	0.43 (0.19-0.96) ¹
Nausea/vomiting	1.3 (1.1-1.6) ¹	0.64 (0.44-0.92) ¹

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio, LR-, negative likelihood ratio.

Once meningitis is considered, clinicians should determine whether a patient requires cranial imaging before LP. Items from the medical history and physical examination that should be used include age greater than 60 years, immunocompromised state, history of CNS disease, seizure within 1 week of presentation, and neurologic abnormality (decreased level of consciousness, inability to answer questions and follow commands, gaze palsy, abnormal visual fields, facial palsy, abnormal motor function, and abnormal language).

Prospective studies have failed to identify individual findings that are accurate enough to diagnosis meningitis. Jolt accentuation of the headache might be useful but requires validation in more studies. The absence of the classic triad of fever, neck stiffness, and headache does not rule out meningitis.

REFERENCE STANDARD TEST

Microbiologic culture.

REFERENCES FOR THE UPDATE

- Thomas KE, Hasburn R, Jekel J, Quagliarello VJ. The diagnostic accuracy of Kernig's sign, Brudzinski's sign, and nuchal rigidity in adults with suspected meningitis. *Clin Infect Dis*. 2002;35(1):46-62.
- van de Beek D, De Gans J, Spanjaard L, Weisfelt M, Reitsma JB, Vermeulen M. Clinical features and prognostic factors in adults with bacterial meningitis. *N Engl J Med*. 2004;351(18):1849-1859.
- Chotmongkol V, Techoruangwiwat C. Community acquired bacterial meningitis in adults. *Southeast Asian J Trop Med Public Health*. 2000;31(3):506-508.
- Hussein AS, Sharfran SD. Acute bacterial meningitis in adults: a 12-year review. *Medicine*. 2000;79(6):360-368.
- Aronin SI, Peduzzi P, Quagliarello VJ. Community-acquired bacterial meningitis: risk stratification for adverse clinical outcome and effect of antibiotic timing. *Ann Intern Med*. 1998;129(11):862-869.
- Hosoglu S, Ayaz C, Geyik MF, Kokoglu OF, Ozen A. Acute bacterial meningitis in adults: analysis of 218 episodes. *Ir J Med Sci*. 1997;166(4):231-234.
- Hasbun R, Abrahams J, Jekel J, Quagliarello VJ. Computed tomography of the head before lumbar puncture in adults with suspected meningitis. *N Engl J Med*. 2001;345(24):1727-1733.^a
- Uchihara T, Tsukagoshi H. Jolt accentuation of headache: the most sensitive sign of CSF pleocytosis. *Headache*. 1991;31(3):167-171.

^aFor the Evidence to Support the Update for this topic, see <http://www.JAMAevidence.com>.

EVIDENCE TO SUPPORT THE UPDATE:

30

Meningitis, Adult

TITLE Computed Tomography of the Head Before Lumbar Puncture in Adults With Suspected Meningitis.

AUTHORS Hasbun R, Abrahams J, Jekel J, Quagliarello VJ.

CITATION *N Engl J Med.* 2001;345(24):1727-1733.

QUESTION Can the absence of certain clinical features at baseline be used to identify adults with suspected meningitis who are unlikely to have abnormal findings on computed tomography (CT) head scan, particularly mass effect?

DESIGN Prospective cohort study.

SETTING Emergency department of Yale–New Haven Hospital, New Haven, Connecticut.

PATIENTS Of 511 adults (>16 years) with clinically suspected meningitis potentially eligible between July 1995 and June 1999, 301 were enrolled in the study. The remainder were excluded mainly because they were identified too late, that is, after the CT or after discharge. The average patient was young (median age, 40 years), white (52%), and immunocompetent (75%).

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

A clinician or study investigator collected standardized baseline clinical characteristics before the lumbar puncture (LP) and CT. CT scans were interpreted blindly by 2 independent neuroradiologists; disagreements were resolved by a third neuroradiologist. Scans were categorized as normal, focal, or nonfocal abnormality and with or without mass effect.

MAIN RESULTS

Of the 301 patients, 235 underwent CT before LP. Fifty-six patients (24%) had a CT abnormality, of which only 11 (5%) had evidence of mass effect. The baseline characteristics associated with any abnormality on CT head scan included being older than 60 years, immunocompromised state, history of central nervous system disease, seizure within 1 week of pre-

Table 30-8 Accuracy of Baseline Characteristic to Detect Any Abnormality on Computed Tomographic Head Scan

	Sensitivity, % (95% CI)	Specificity, % (95% CI)	LR+ (95% CI)	LR− (95% CI)
Any baseline characteristic	95 (85-98)	52 (45-59)	2.0 (1.7-2.3)	0.10 (0.03-0.31)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR−, negative likelihood ratio.

sensation, and neurologic abnormality. The neurologic abnormalities were a decreased level of consciousness, inability to answer questions and follow commands, gaze palsy, abnormal visual fields, facial palsy, abnormal motor function, and abnormal language.

The accuracy of any of the above baseline characteristics for detecting any abnormality on CT is shown in Table 30-8. Table 30-9 presents the accuracy of any of the significant baseline characteristics to detect mass effect on CT head scan.

CONCLUSION

LEVEL OF EVIDENCE Level 2.

STRENGTHS Prospective data collection with appropriate blinding on consecutive patients.

LIMITATIONS Most (78%), but not all, patients had CT before LP.

For patients with suspected meningitis, it is common practice in some centers to order a CT before proceeding to LP to detect any mass effect and avoid causing transtentorial herniation with an LP. A previous review¹ suggested that

Table 30-9 Accuracy of Baseline Characteristic to Detect Mass Effect on Computed Tomographic Head Scan

	Sensitivity, % (95% CI)	Specificity, % (95% CI)	LR+ (95% CI)	LR− (95% CI)
Any baseline characteristic	91 (62-96)	42 (36-49)	1.6 (1.2-2.0)	0.21 (0.03-1.4)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR−, negative likelihood ratio.

1. the risk of herniation in the setting of increased intracranial pressure without obstruction to cerebrospinal fluid flow had been overstated; and
2. CT was probably not necessary before proceeding to LP in patients without neurologic abnormalities or atypical features (such as being immunocompromised).

First, the results indicate that the absence of any significant baseline characteristic (detailed above) is a reasonably strong indicator of a lack of mass effect on CT head scan. Only 1 of 11 (9%) patients with mass effect was missed with these criteria, although the confidence interval (CI) indicates that the true value may be as high as 38%. Given a pretest probability of mass effect in this study population of 5%, the absence of these characteristics reduces the posttest probability to 1.0% (95% CI, 0.16%-6.9%).

Second, the risk of herniation after LP, even in the presence of mass effect on CT, is low. Of the 11 patients with mass effect, 7 went on to have LP anyway (including the 1 patient missed with the baseline criteria) and none had herniation at clinical follow-up 1 week later.

Overall, the evidence suggests that the absence of specific findings on clinical history and neurologic examination can reasonably safely identify those who do not need CT before LP.

REFERENCE FOR THE EVIDENCE

1. Archer BD. Computed tomography before lumbar puncture in acute meningitis: a review of the risks and benefits. *CMAJ*. 1993;148(6):961-985.

Reviewed by John Attia, MD

TITLE The Diagnostic Accuracy of Kernig's Sign, Brudzinski's Sign, and Nuchal Rigidity in Adults With Suspected Meningitis.

AUTHORS Thomas KE, Hasbun R, Jekel J, Quagliarello VJ.

CITATION *Clin Infect Dis*. 2002;35(1):46-62.

QUESTION What is the accuracy of Kernig sign, Brudzinski sign, and nuchal rigidity in adults with suspected meningitis?

DESIGN Prospective cohort study.

SETTING Emergency department of Yale-New Haven Hospital

PATIENTS Two hundred ninety-seven patients presenting to the emergency department between July 1995 and June 1999 with suspected meningitis (clinical symptoms compatible with meningitis) who underwent lumbar puncture. This study population was also used to evaluate the safety of lumbar puncture without computed tomographic (CT) head scan.¹ Of 301 patients who were enrolled, 4 were excluded because of mass effect on CT head scan.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

An emergency department physician recorded the clinical history and physical examination results before lumbar puncture. A patient was considered to have meningitis if the cerebrospinal fluid white blood cell count was greater than or equal to 6 cells/mL.

MAIN RESULTS

Eighty patients had meningitis and 217 did not. Seventeen percent of the entire cohort were immunocompromised.

None of the history items were helpful in ruling in meningitis, although the absence of a headache or nausea and vomiting would decrease the probability of meningitis (Table 30-10). Neither a fever nor any of the maneuvers were accurate (Kernig or Brudzinski signs or nuchal rigidity) (Table 30-11).

CONCLUSION

LEVEL OF EVIDENCE Level 2.

STRENGTHS Prospective, consecutive patients for whom the clinicians had a suspicion of meningitis. The physical examination was always done blinded to the results of the lumbar puncture.

Table 30-10 History Items

Finding (No.)	Sensitivity, % (95% CI)	Specificity, % (95% CI)	LR+ (95% CI)	LR- (95% CI)
Headache (282)	92 (84-96)	19 (14-25)	1.1 (1.0-1.3)	0.43 (0.19-0.96)
Nausea and vomiting (290)	70 (59-79)	47 (40-54)	1.3 (1.1-1.6)	0.64 (0.44-0.92)
Neck stiffness (296)	48 (37-59)	55 (49-62)	1.1 (0.82-1.4)	0.95 (0.74-1.2)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

Table 30-11 Physical Examination

Finding (No.)	Sensitivity, % (95% CI)	Specificity, % (95% CI)	LR+ (95% CI)	LR- (95% CI)
Fever (297)	43 (32-53)	48 (41-55)	0.82 (0.62-1.1)	1.2 (0.94-1.5)
Kernig sign (237)	5 (1.6-13)	95 (91-98)	0.97 (0.27-3.6)	1.0 (0.94-1.1)
Brudzinski sign (236)	5 (1.6-13)	95 (91-98)	0.97 (0.26-3.5)	1.0 (0.94-1.1)
Neck stiffness (297)	30 (21-41)	68 (62-74)	0.94 (0.64-1.4)	1.0 (0.87-1.2)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

LIMITATIONS The manner in which the emergency department physicians performed the physical examination was not standardized, and not all patients were assessed for each physical examination maneuver.

This study is of better quality than most others on the physical examination in meningitis. Unfortunately, the authors did not standardize the techniques for performing the physical examination maneuvers. As a result, almost 20% of the patients were not examined for Kernig or Brudzinski signs. The small proportion of patients who were immunocompromised may have contributed to the lower accuracies of these findings because some of these patients may have been unable to mount an inflammatory response to central nervous system infection.²

Although the classic triad of fever, neck stiffness, and altered mental status was not directly addressed in this study,

the very weak negative likelihood ratio associated with the 3 findings individually (1.2, 1.0, and 0.97, respectively) casts doubt on our previous assertion in the original Rational Clinical Examination article that the absence of this triad virtually eliminates a diagnosis of meningitis.

Reviewed by Rose Hatala, MD

REFERENCES FOR THE EVIDENCE

1. Hasbun R, Abrahams J, Jekel J, Quagliarello VJ. Computed tomography of the head before lumbar puncture in adults with suspected meningitis. *N Engl J Med.* 2001;345:1727-1733.
2. Chuck SL, Sande M. Infection with *Cryptococcus neoformans* in the acquired immunodeficiency syndrome. *N Engl J Med.* 1989;321(12):794-799.

This page intentionally left blank

CHAPTER 31

Is This Woman Perimenopausal?

Lori A. Bastian, MD, MPH

Crystal M. Smith, MD

Kavita Nanda, MD, MHS

CLINICAL SCENARIOS

Are These Women Perimenopausal?

For each of the following cases, the clinician may need to determine the probability that the patient is perimenopausal.

CASE 1 A 45-year-old woman who had a hysterectomy at age 42 years for uterine fibroids reports that she has hot flashes and has felt irritable for the past month.

CASE 2 A 41-year-old woman tells her physician that she thinks she is starting menopause. She smokes 1 pack of cigarettes a day, as she has for the past 20 years.

CASE 3 A 47-year-old woman who has been taking oral contraceptives for the past 25 years requests information about her menopausal status. She is sexually active and wants to know whether she needs to continue taking birth control medication.

WHY IS THE DIAGNOSIS IMPORTANT?

The question, “Is this woman perimenopausal?” is important for clinicians because patients ask and want to know whether they are undergoing a physical and emotional change and whether they are experiencing the menopausal transition. Physicians need information to identify perimenopausal women, to be able to reply to women’s questions about the changes they may be experiencing, and to offer counseling on symptom relief, contraception, and disease prevention. As women begin the perimenopausal years, clinicians should counsel them on strategies to prevent osteoporosis, as well as on evidence-based treatment options for climacteric symptoms such as hot flashes and night sweats. Clinicians commonly identify perimenopausal women by their ages, by inquiring about their menstrual histories and symptoms, and by ordering laboratory tests to examine hormone levels, such as follicle-stimulating hormone (FSH) and estradiol levels, to confirm their clinical suspicions. It would be useful to know how age, self-assessment, family and medical history, symptoms, physical signs, and laboratory tests affect the probability that the woman is perimenopausal.

In this article, we intend to answer the following questions: What is the value of asking a woman whether she thinks she is starting menopause? How accurate are symptoms and signs in detecting perimenopause? Is there any value in asking about family and medical history in determining menopausal status? Are laboratory tests more useful than clinical examination in diagnosing perimenopause?

PHYSIOLOGY AND DEFINITIONS

Climacteric is a general term referring to the entire transition from the reproductive to the postreproductive interval in a woman's life.¹ Thus, it includes immediate premenopausal, perimenopausal, and postmenopausal women. All women do not go through the same transition of regular menses to irregular menses to amenorrhea as they approach menopause. In 2001, a panel of experts (from the Stages of Reproductive Aging Workshop) met to discuss a staging system to classify reproductive aging.² This proposed new classification of the transition from reproductive to postmenopausal includes 7 stages, which are based on menstrual cycles and plasma FSH levels. The experts of this system observed that this is a work in progress, and it has not been validated in research settings.

The World Health Organization³ defines *natural menopause* as "the permanent cessation of menstruation, determined retrospectively after 12 consecutive months of amenorrhea without any other pathological or physiological cause."³ Menstruation ceases as ovarian follicle stores are depleted and ovarian function is diminished, leading to eventual decreased production of estrogen by the ovary and decreased stimulation of the endometrial lining.⁴ Analysis of longitudinal data of women at all ages shows a probability of less than 2% for spontaneous menstruation after 12 months of amenorrhea.⁵ The accurate diagnosis of perimenopause allows patients and physicians to predict the onset of menopause.

Perimenopause refers to the year before the final menstrual period through the first year after the final menstrual period.^{3,6} During perimenopause, ovulation occurs irregularly because of fluctuations in the hormones of the hypothalamic-pituitary-ovarian axis.⁶ For example, in early perimenopause, inhibin B levels decline, resulting in an increase in FSH levels, with no significant change in inhibin A or estradiol levels. FSH levels may increase during some cycles but return to premenopausal levels in subsequent cycles. Further complicating the determination of FSH concentration is the pulsatile pattern of secretion. Similarly, concentrations of estradiol also may decrease or even increase during perimenopause.⁵ This hormonal variability creates difficulties in interpreting a single laboratory test value.

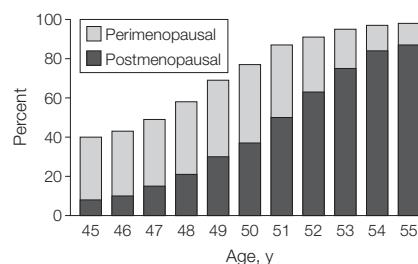


Figure 31-1 Prevalence of Perimenopause and Postmenopause by Age
Median age at perimenopause is 47.5 years, and median age at postmenopause is 51.3 years. Adapted with permission from McKinlay et al.⁹ Massachusetts Women's Health Study ($n = 5547$).

According to longitudinal data of women's menstrual cycles, Brambilla et al⁷ and Dudley et al⁸ further refined the definition of perimenopause by considering a woman perimenopausal if she has not had a period within the previous 3 to 11 months or if she has experienced changes in menstrual regularity (either shortening or lengthening of time between menses) during the past 12 months. In a 5-year population-based study, Brambilla et al⁷ found that 3 to 11 months of amenorrhea or irregular periods among women aged 45 to 55 years were most predictive of menopause within the following 3 years (sensitivity, 72%; specificity, 76%). Dudley et al⁸ validated this definition, finding that these 2 characteristics are the best predictors of menopause 4 years after baseline (sensitivity, 32%; specificity, 99%). The perimenopausal definition by Brambilla et al⁷ was used as our reference standard for this systematic review.

ESTIMATING THE PRETEST PROBABILITY OF PERIMENOPAUSE

To determine a woman's likelihood of perimenopause, clinicians must first estimate the pretest probability of perimenopause. This estimate should be based primarily on the patient's age, although certain aspects of the medical and family history also may be useful.

In a 30-year study that enrolled college women and followed them throughout their lifetime until menopause, Treloar et al⁵ reported the mean age of onset of perimenopause as 45.5 years, with a mean duration of 6.2 years. According to 5-year follow-up data from a population-based study of 5547 women aged 45 to 55 years, McKinlay et al,⁹ in the Massachusetts Women's Health Study (1992), reported the median age of onset of perimenopause as 47.5 years, with a mean duration of 3.8 years. Figure 31-1 shows the prevalence of perimenopause and postmenopause according to age from McKinlay et al⁹ data. Unfortunately, estimating the time of onset of perimenopause is difficult, and data were not available from the literature on the prevalence of perimenopause among women younger than 45 years. McKinlay et al⁹ reported that by age 45 years, 40% of all women have started or completed the menopause transition (32% are perimenopausal and 8% are postmenopausal). By age 50 years, 75% of women have started or completed the transition (38% perimenopausal and 37% postmenopausal). By age 55 years, only 2% of women are premenopausal.

EVALUATION OF PERIMENOPAUSE

This evaluation can be divided into 5 basic categories: self-assessment, symptoms, family and medical history, physical signs, and laboratory tests.

Self-Assessment

Clinicians can ask a woman whether she thinks she is starting menopause. Women may base their perceptions of their menopausal status on awareness of the subtle changes taking place in their bodies.^{10,11} In a cross-sectional study by Garamszegi et al,¹⁰

self-reported menopausal status was more correlated with symptoms than menstrual cycle characteristics.

Symptoms

Climacteric symptoms typically include vasomotor complaints, such as hot flashes and night sweats. Other symptoms associated with perimenopause in cross-sectional studies are thought to be associated with fluctuating levels of estrogen and progesterone. These include vaginal dryness, variable sexual interest, urinary incontinence, depressed mood, nervous tension and irritability, and sleep disturbances.¹

Hot Flashes

Hot flashes are sudden sensations of heat, sweating, and flushing that most often occur in the face, head, neck, and chest. Chills, clamminess, and anxiety also may accompany hot flashes. They generally last 1 to 5 minutes, though 6% of women experience hot flashes lasting longer than 6 minutes.¹² Most North American, European, and Australian women report that they experience hot flashes (50%-85%)^{9,12-14} and that they occur periodically during a span of 1 to 5 years.^{15,16} There appear to be cultural differences in the reporting or experiencing of hot flashes. For example, only 10% to 20% of Indonesian women¹⁷ and 10% to 25% of Chinese women¹⁸ report experiencing them. The mechanism triggering these episodes is thought to be a combination of fluctuating estradiol levels and a narrowing of the thermoneutral zone.¹⁹

Night Sweats

Night sweats are hot flashes that occur at night, usually while the woman is sleeping. Often, she will awake drenched in sweat. If night sweats interfere with sleeping patterns, this may explain reports of insomnia, fatigue, and irritability among climacteric women.

Vaginal Dryness

Vaginal dryness is sometimes experienced as a result of decreasing estrogen production during perimenopause. This can lead to urogenital atrophy and changes in the quantity or composition of vaginal secretions. Estimates of the prevalence of vaginal dryness among late perimenopausal women range from 18%²⁰ to 21%.²¹

Variable Sexual Interest

Dennerstein et al²² report in a study of Australian women that although most indicated no change in sexual interest during menopause, 31% experienced a decrease and 7% reported an increase in sexual interest. Only 6% of those reporting a decrease indicated menopause as a reason for the decline in interest.²² This decrease may be caused by physiologic factors making sexual relations more difficult (eg, vaginal dryness, hot flashes, urinary incontinence) or social and environmental factors. Several studies have found that menopausal symptoms are but one of many factors affecting sexual interest among women in midlife and later.^{23,24}

Urinary Incontinence

Urinary incontinence affects between 26%²⁵ and 55%²⁶ of middle-aged women from western countries and may be caused or exacerbated by declining estrogen levels. Lower

estrogen levels can lead to atrophy of the urethral mucosa and the trigone, the muscle controlling urination, resulting in less urinary control.⁶ Some studies have found an association between increased prevalence of urinary incontinence and menopause,²⁵ whereas others have not.^{27,28}

Depressed Mood

Avis et al²⁹ classified 10% of 45- to 55-year-old women participating in a population-based longitudinal study of women from Massachusetts as experiencing clinical depression. Many studies do not find an association of menopause with depression or find that it can be explained by other menopausal symptoms.²⁹⁻³⁴ Evidence from North American²⁹ and British³⁵ cohorts found high rates of depression among perimenopausal women with a history of depression, supporting the theory that women with previous affective disorders may be at an increased risk for recurrent depression. Conclusions from other reports have suggested that depression could be increased because of declines in estrogen levels,³⁶ changes in social circumstances,³⁷ and changes in self-concept as women lose reproductive function.³⁸

Nervous Tension and Irritability

Many symptom checklists for menopause symptoms used in epidemiologic studies include nervous tension and irritability.^{21,39-42} Although the relevance of these symptoms is unclear, they could be caused by lack of sleep because of menopausal symptoms, illness, or stressful life events. Some authors suggest that they could result from changes in hormone levels, which also occurs during the 10- to 14-day luteal phase of the menstrual cycle.⁴³

Family and Medical History

Age of Mother's Menopause

Genetic factors seem to predispose women to menopause at an earlier age.^{44,45} Torgerson et al⁴⁴ reported that women with premature (<40 years) and early (<45 years) menopause report significantly younger maternal menopausal ages than did women with normal menopausal ages. In a case-control study of women from the greater Boston area, Cramer et al⁴⁵ found that women with a family history (eg, mother, sister, aunt, grandmother) of menopause before age 46 years had a higher risk of early menopause (odds ratio, 6.1; 95% confidence interval [CI], 3.9-9.4).

Cigarette Use

Approximately 23% of US adult women smoke cigarettes regularly.⁴⁶ Evidence indicates that women who smoke experience menopause 1 to 2 years earlier than do nonsmokers.^{2,47-54} Cigarette smoking reduces bioavailable estrogen by increasing hepatic metabolism of estrogen,^{55,56} decreasing production of estrogen,^{57,58} or increasing circulation of androgens.⁵⁹ Several studies support the assertion that quitting smoking can significantly delay menopause.^{48,49} Other evidence suggests that the median age of menopause is not statistically different between women who have never smoked and ex-smokers.^{60,61} Nevertheless, a majority of research on cigarette smoking and menopause does indicate a dose-response relationship between number of cigarettes currently smoked and age at menopause.^{48,49,53,62} Furthermore, Gold et al^{20,54} observed that “past smoking and current smoking were positively associated with prevalence of vasomotor symptoms,” in agreement with most previous data.⁵⁴

Hysterectomy Status

It is often assumed that women who have had a hysterectomy with conservation of the ovaries should not experience menopausal symptoms earlier or more severely because of their hysterectomy. Nonetheless, evidence shows that women with ovarian conservation after hysterectomy report more vasomotor complaints, vaginal dryness, and other complaints than do women of similar age who did not have a hysterectomy.^{63,64} In developed countries, hysterectomy is one of the most frequent operations in adult women⁶³; one-third of US women will have had a hysterectomy by age 65 years.⁶⁵ Hysterectomy may inhibit blood circulation in ovaries, decreasing ovarian function⁶⁴ and causing more frequent or severe menopausal symptoms.

Physical Signs

Maturation Index

One proposed assessment of vaginal estrogen deficiency is an evaluation of the vaginal epithelium maturation index. This procedure involves obtaining cells from the junction of the upper and middle third of the lateral vaginal wall with a brush. These cells are prepared on a slide with the Papanicolaou technique, and the percentages of parabasal, intermediate, and superficial cells are counted.⁶⁶ Although the maturation index changes significantly after estrogen replacement therapy, diagnostic studies have not compared the maturation index with menstrual cycle characteristics.

Vaginal pH

Some investigators suggest that an increased vaginal pH (6.0–7.5) in the absence of potentially pathogenic bacteria may be a reasonable marker of decreased estradiol serum levels.⁶⁷ This test is performed by directly applying pH paper to the lateral vaginal wall at the outer third of the vagina. Changes in pH can alter the composition of vaginal secretions that accompany atrophy.

Skin Thickness

Estrogen stimulates the epidermal growth rate and promotes the formation of collagen and hyaluronic acid, which increase the turgor and vascularization of the skin.⁶⁸ During climacteric, declining estrogen levels result in the thinning and atrophy of the epidermis.⁶⁸ Investigators have proposed measuring skin thickness with ultrasonography at the greater trochanter area to estimate menopausal status, but this procedure has not been supported by research to date.⁶⁸

Laboratory Tests

Follicle-Stimulating Hormone

Measurement of FSH plasma levels has been used to try to identify perimenopausal and postmenopausal women. High FSH levels indicate that menopausal changes are occurring in the ovary. As the ovary becomes less responsive to stimulation by FSH from the pituitary gland (and produces less estrogen), the pituitary gland increases production of FSH to try to stimulate the ovary to produce

more estrogen (Figure 31-1). However, some clinicians and researchers doubt the clinical value of FSH measurements in perimenopausal women because FSH levels fluctuate considerably each month, depending on whether ovulation has occurred.^{2,69,70}

Estradiol

Recent longitudinal studies have reported that early perimenopausal (change in cycle frequency) women maintained premenopausal estradiol levels, whereas late perimenopausal (no menses in previous 3–11 months) and postmenopausal women experienced significant declines in estradiol levels.⁷¹ Estradiol can be measured using plasma, urine, and saliva. Like FSH, estradiol levels are highly variable during perimenopause.¹

Inhibins

Inhibin A and inhibin B are secreted by the ovaries and, like estradiol, exert negative feedback on the pituitary gland, reducing FSH and luteinizing hormone secretion (Figure 31-1). Loss of inhibin contributes to the increase in FSH that occurs with ovarian senescence. A recent longitudinal study of hormone levels throughout the menopause transition reported that inhibin B levels decline as women progress through perimenopause, whereas inhibin A levels remain unchanged. Inhibin A levels did decrease at approximately the final menstrual period.⁷¹ Inhibin levels are usually measured in plasma. The ovaries produce less inhibin B as fewer follicles proceed to maturation, and the number of follicles declines with age.⁷²

METHODS

Search Strategy and Quality Review

We searched the MEDLINE database for English-language articles concerning the diagnosis of menopause that were published between 1966 and 2001. The key words used included “menopause, perimenopause, premenopause, climacteric, sensitivity” and “specificity, diagnosis, prospective/cross-sectional studies, health status,” and “hormones of the hypothalamic-pituitary-ovarian axis.” We included articles that used the diagnosis of perimenopause based on menstrual irregularity or 3 to 11 months of amenorrhea, included a premenopausal control group, and presented data that could be extracted to calculate both sensitivity and specificity rates. We included articles on laboratory tests that are available to clinicians for 2 reasons. First, women may ask for laboratory tests to assess their menopausal status. Second, the results of the tests must be coupled closely with the clinical examination for proper interpretation. We excluded reviews and articles that included men, hormone replacement therapy (HRT), cancer, or osteoporosis as major foci of the papers. We developed the search strategy with a medical librarian, and this is available from the authors on request. Two authors (L.A.B. and C.M.S.) systematically reviewed and identified titles and abstracts for content and quality. Articles using a definition of perimenopause different from 3 to 11 months of amenorrhea or irreg-

ular periods, those lacking a control group (a remote premenopausal group), and studies for which data could not be classified into contingency tables were excluded. Articles using a young control group (ie, 20-year-old women) or an older postmenopausal group (ie, 60- to 70-year-old women) or including women receiving HRT also were excluded. Two authors (L.A.B. and C.M.S.) abstracted the articles with a standardized abstraction form. Each publication was given a grade of A, B, or C according to the study design and level of

evidence (see Table 1-7 for a summary of Evidence Grades and Levels).⁷³ Discrepancies about quality were resolved by a third author (K.N.).

The MEDLINE search identified 1221 articles, and from the references cited in these and other publications known to us, another 25 articles were added to the review pool. Sixteen articles^{10,11,20-22,26,28,39,42,54,68,74-78} met all the inclusion criteria described above and were included in the final analysis (Table 31-1).

Table 31-1 Studies Included in the Analysis

Source, y	Study Population	Setting	Study Design	Age Range, y	Premenopause, No.	Perimenopause, No. (%)	Symptoms and Signs Studied	Study Quality Score ^a
Chompootweep et al, ⁷⁴ 1993	Thai women	Health centers	Cross-sectional	45-59	735	292 (28)	Hot flash, mood, insomnia	A
Dennerstein et al, ⁷⁵ 1993	Australian women	Population database	Cross-sectional	45-55	316	549 (63)	Hot flash, mood, insomnia, nervous tension	A
Dennerstein et al, ²² 1994	Australian women	Population database	Cross-sectional	45-55	290	504 (63)	Sexual interest	A
Punyahotra et al, ⁷⁶ 1997	Companions of outpatients in Thailand	Outpatient clinic	Cross-sectional	40-59	127	22 (15)	Hot flash, night sweat, mood, nervous tension	B
Burger et al, ⁷¹ 1998	Australian women	Population database	Prospective	45-55	28	59 (68)	Inhibins	B
Garamszegi et al, ¹⁰ 1998	Australian women	Population database	Prospective	45-55	91	182 (67)	Night sweat, self-rating	B
Stellato et al, ⁷⁷ 1998	Massachusetts women	Population database	Cross-sectional	50-60	99	179 (64)	FSH	B
Ho et al, ³⁹ 1999	Chinese women	Population database	Cross-sectional	44-55	1258	92 (7)	Hot flash, mood, insomnia	B
Kuh et al, ²⁶ 1999	British women born in 1946	Population database	Prospective	48	480	319 (40)	Incontinence	A
Dennerstein et al, ²¹ 2000	Australian women	Population database	Prospective	45-55	172	254 (60)	Vaginal dryness	B
Gold et al, ²⁰ 2000	US ethnic communities	Community (SWAN)	Cross-sectional	40-55	4497	4158 (48)	Vaginal dryness, insomnia, incontinence	B
Harlow et al, ¹¹ 2000	US ethnic communities	Community (SWAN)	Cross-sectional	40-55	4234	3928 (48)	Self-rating	B
Bromberger et al, ⁴² 2001	US ethnic communities	Community (SWAN)	Cross-sectional	40-55	4483	4143 (48)	Psychological distress ^b	B
Gold et al, ⁵⁴ 2001	US ethnic communities	Community (SWAN)	Cross-sectional	40-55	4514	4173 (48)	Cigarette smoking	B
Maartens et al, ⁷⁸ 2001	Dutch women	Population database	Cross-sectional	47-54	526	1250 (70)	Hot flash, night sweat, mood, insomnia, nervous tension, vaginal dryness, incontinence	B
Sherburn et al, ²⁸ 2001	Australian women	Population database	Prospective	45-55	471	393 (45)	Urinary incontinence	A

Abbreviations: FSH, follicle-stimulating hormone; SWAN, Study of Women's Health Across the Nation.

^aSee Table 1-7 for a summary of Evidence Grades and Levels.

^bPsychological distress is defined as depression, irritability, or nervous tension in the past 2 weeks.

Table 31-2 History, Symptoms, and Hormone Levels in the Prediction of Perimenopause

Symptoms and Signs	No. of Participants	Sensitivity Range	Specificity Range	LR+ (95% CI) or Range ^a	LR- (95% CI) or Range ^a
Hot flashes ^{39,74-76,78}	5167	0.22-0.55	0.83-0.91	2.2-4.1	0.54-0.87
Night sweats ^{10,76,78}	2198	0.20-0.50	0.74-0.87	1.9 (1.6-2.2) ^b	0.67-0.92
Vaginal dryness ^{20,21,78}	10857	0.11-0.29	0.80-0.97	1.5-3.8	0.92 (0.91-0.94) ^b
Incontinence ^{20,26,28,78}	12094	0.16-0.39	0.64-0.91	1.1-1.7	0.91 (0.89-0.93) ^b
Depressed mood ^{39,74-76,78}	5167	0.09-0.47	0.64-0.97	1.3-3.1	0.82-0.94
Insomnia ^{20,39,74,75,78}	13673	0.21-0.53	0.63-0.83	0.98-2.1	0.79-1.0
Nervous tension or irritability ^{75,76,78}	2790	0.41-0.59	0.51-0.68	1.2 (1.1-1.3) ^b	0.83 (0.77-0.90) ^b
Psychological distress ⁴²	8626	0.28	0.79	1.3 (1.2-1.4)	0.91 (0.89-0.93)
Sexual interest ²²	799	0.25	0.84	1.6 (1.2-2.1)	0.89 (0.83-0.96)
Self-rating ^{10,11}	8435	0.77-0.94	0.39-0.64	1.5-2.1	0.18-0.36
Current smoking ⁵⁴	8185	0.24	0.82	1.3 (1.2-1.4)	0.93 (0.91-0.95)
FSH ⁷⁷ (≥ 24 mIU/L)	278	0.65	0.79	3.1 (2.1-4.5)	0.45 (0.36-0.56)
Inhibin A ⁷¹ (<1.28 U/L)	87	0.61	0.54	1.3 (0.84-2.0)	0.73 (0.46-1.2)
Inhibin B ⁷¹ (<30 ng/L)	87	0.46	0.78	2.0 (0.96-4.4)	0.70 (0.51-0.96)
IR-INH ⁷¹ (<30 ng/L)	87	0.07	0.96	1.9 (0.22-16)	0.97 (0.88-1.1)

Abbreviations: CI, confidence interval; FSH, follicle-stimulating hormone; IR-INH, immunoreactive inhibin; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

^aLR+ is a measure of how well a positive result rules in perimenopause and an LR- measures how well a negative test result rules out perimenopause. Where one of these operating characteristics was homogeneous ($P > .05$ for the χ^2 test), the summary value and a 95% CI are given. Where they are heterogeneous, only the range is given.

^bFor LRs, a summary measure is reported only when more than 2 studies were identified and found to be homogeneous; otherwise, a range was reported.

Statistical Methods

We calculated values and CIs for sensitivity, specificity, positive likelihood ratios (LRs+), and negative likelihood ratios (LRs-), using statistical software (SAS version 8.0; SAS Institute Inc, Cary, North Carolina). Perimenopause is the target condition, and the reference standard is based on the definition by Brambilla et al.⁷

The LR+ (sensitivity/[1 – specificity]) is a measure of how well a positive test result rules in perimenopause, whereas the LR- ([1 – sensitivity]/specificity) is a measure of how well a negative test result rules out perimenopause. An LR close to 1 does not appreciably predict the likelihood of perimenopause. An LR greater than 1 increases the likelihood of perimenopause, whereas an LR less than 1 decreases the likelihood of perimenopause. We assessed sensitivity, specificity, LR+, and LR- for homogeneity. When the χ^2 statistic suggested homogeneity ($P > .05$), we combined the data to produce a random-effects estimate.⁷⁹ For heterogeneous data, variables are given as ranges.

RESULTS

Findings that were similar across studies (Table 31-2), that is, those that had the greatest LR+ and were therefore best at ruling in perimenopausal status were hot flashes (LR+, 2.2-4.1), night sweats (LR+, 1.9; 95% CI, 1.6-2.2), and vaginal dryness (LR+, 1.5-3.8). The absence of findings was not efficient at ruling out perimenopausal status; self-rating (LR-, 0.18-0.36) and hot flashes (LR-, 0.54-0.87) had the smallest

LR-. Only 1 study each reported enough data to calculate sensitivity, specificity, and LRs for FSH and the inhibins^{71,77}; no study reported enough data to calculate these values for estradiol. High FSH levels (≥ 24 mIU/L) and low inhibin B levels (≤ 30 ng/L) provided weak evidence to rule in perimenopause (LR+, 3.1; 95% CI, 2.1-4.5; and LR+, 2.0; 95% CI, 0.96-4.4, respectively). However, neither normal FSH levels nor normal inhibin B levels could rule out perimenopause (LR-, 0.45; 95% CI, 0.36-0.56; and LR-, 0.70; 95% CI, 0.51-0.96, respectively).

CLINICAL SCENARIOS—RESOLUTIONS

Case 1 describes a 45-year-old woman with a moderately high pretest probability of being perimenopausal or postmenopausal (40%) according to her age (Figure 31-1) and probably even higher because she has had a hysterectomy and is experiencing climacteric symptoms. Because she has reported hot flashes (LR +, 2.2-4.1) and irritability (LR +, 1.2; 95% CI, 1.1-1.3), the calculated posttest probability of her being perimenopausal ranges from 40% to 100%. Our recommendation would be to not order FSH or other laboratory tests but to tell her that she is perimenopausal, to counsel her on increasing her calcium intake, and advise her to increase exercise to prevent osteoporosis.

In case 2, a 41-year-old woman might have a pretest probability of being perimenopausal or postmenopausal (estimate, 10%) according to her age. Because she is currently smoking cigarettes (LR+, 1.3; 95% CI, 1.2-1.4), this

aspect of her medical history raises her probability of being perimenopausal to 12% to 14%. Because she thinks she is starting menopause (LR+, 1.5-2.1), her calculated probability of being perimenopausal might be further increased to 18% to 30%. We would inquire about her menstrual patterns and presence of climacteric symptoms, tell her she may be close to perimenopause, and discuss contraception and smoking cessation with her.

Finally, the 47-year-old woman in case 3 has a different type of question. Her pretest probability of being perimenopausal or postmenopausal is high (50%) according to her age (Figure 31-1). In this patient, we cannot assess for menstrual patterns or menopausal symptoms because she is using an oral contraceptive. Although she is likely perimenopausal, ovulation may still be possible.⁸⁰ If she desires to continue with her oral contraceptive, the American College of Obstetricians and Gynecologists recommends discontinuing this therapy between the ages of 50 and 55 years.⁸¹ As suggested in case 1, we would counsel her on increasing calcium intake and increasing exercise for osteoporosis prevention.

CONCLUSION

No single element of the medical history or clinical examination is powerful enough to confirm the probability of being perimenopausal. Besides menstrual history, the most powerful predictor of menopausal status is a woman's age. The median age at perimenopause is 47.5 years,⁹ and 87% of women are perimenopausal or postmenopausal by the age of 51 years. The clinical question of perimenopausal status is more difficult in patients in their early to middle 40s. Many clinicians rely on the measurement of hormone levels, such as FSH, to confirm the diagnosis. In the clinical scenarios we evaluated, FSH measurement did not help the clinician make a diagnosis. Further research needs to be conducted to document the additional benefit of these hormone level tests in making a diagnosis of perimenopause.

Author Affiliations at the Time of the Original Publication

Departments of Internal Medicine (Dr Bastian) and Obstetrics and Gynecology (Drs Bastian and Nanda), Duke University, Durham, North Carolina; Center for Health Services Research in Primary Care, Veterans Affairs Medical Center, Durham, North Carolina (Drs Bastian and Smith); and Family Health International, Research Triangle Park, North Carolina (Dr Nanda).

Acknowledgments

Dr Bastian's work was supported by a Veterans Affairs Career Development Award from Health Services Research and Development.

We thank Joanne Piscitelli, MD, Ann Brown, MD, and Peter F. Blomgren, MD, for their thoughtful comments on an

earlier version of this work, and Lesa Hall-Young (Medical Media, Durham VA Medical Center, North Carolina) for her technical assistance with the figure.

REFERENCES

1. Reame NE. Neuroendocrine regulation of the perimenopause transition. In: Lobo RA, Kelsey J, Marcus R, eds. *Menopause: Biology and Pathobiology*. San Diego, CA: Academic Press; 2000:95-96.
2. Soules MR, Sherman S, Parrott E, et al. Executive summary: Stages of Reproductive Aging Workshop (STRAW). *Fertil Steril*. 2001;76(5):874-878.
3. World Health Organization. *Report of a WHO Scientific Group: Research on the Menopause in the 1990's*. Geneva, Switzerland: World Health Organization; 1996. WHO Technical Report Series 866.
4. O'Connor KA, Holman DJ, Wood JW. Menstrual cycle variability and the perimenopause. *Am J Hum Biol*. 2001;13(4):465-478.
5. Treloar AE, Boynton RE, Behn BG, Brown BW. Variation of the human menstrual cycle through reproductive life. *Int J Fertil*. 1967;12(1 pt 2):77-126.
6. Greendale GA, Lee NP, Arriola ER. The menopause. *Lancet*. 1999;353 (9152):571-580.
7. Brambilla DJ, McKinlay SM, Johannes CB. Defining the perimenopause for application in epidemiologic investigations. *Am J Epidemiol*. 1994;140(12):1091-1095.
8. Dudley EC, Hopper JL, Taffe J, Guthrie JR, Burger HG, Dennerstein L. Using longitudinal data to define the perimenopause by menstrual cycle characteristics. *Climacteric*. 1998;1(1):18-25.
9. McKinlay SM, Brambilla DJ, Posner JG. The normal menopause transition. *Maturitas*. 1992;14(2):103-115.
10. Garamszegi C, Dennerstein L, Dudley E, Guthrie JR, Ryan M, Burger H. Menopausal status: subjectively and objectively defined. *J Psychosom Obstet Gynaecol*. 1998;19(3):165-173.
11. Harlow SD, Crawford SL, Sommer B, Greendale GA. Self-defined menopausal status in a multi-ethnic sample of midlife women. *Maturitas*. 2000;36(2):93-112.
12. Kronenberg F, Downey JA. Thermoregulatory physiology of menopausal hot flashes: a review. *Can J Physiol Pharmacol*. 1987;65(6):1312-1324.
13. McKinlay SM, Jefferys M. The menopausal syndrome. *Br J Prev Soc Med*. 1974;28(2):108-115.
14. Jaszmann L, Van Lith ND, Zaai JC. The age of menopause in the Netherlands: the statistical analysis of a survey. *Int J Fertil*. 1969;14(2):106-117.
15. Feldman BM, Voda A, Gronseth E. The prevalence of hot flash and associated variables among perimenopausal women. *Res Nurs Health*. 1985;8(3):261-268.
16. Kronenberg F. Hot flashes: epidemiology and physiology. *Ann NY Acad Sci*. 1990;592:52-86.
17. Flint M, Samil RS. Cultural and subcultural meanings of the menopause. *Ann NY Acad Sci*. 1990;592:134-148.
18. Tang GW. The climacteric of Chinese factory workers. *Maturitas*. 1994;19(3):177-182.
19. Freedman RR. Physiology of hot flashes. *Am J Hum Biol*. 2001;13(4): 453-464.
20. Gold EB, Sternfeld B, Kelsey JL, et al. Relation of demographic and life-style factors to symptoms in a multi-racial/ethnic population of women 40-55 years of age. *Am J Epidemiol*. 2000;152(5):463-476.
21. Dennerstein L, Dudley EC, Hopper JL, Guthrie JR, Burger HG. A prospective population-based study of menopausal symptoms. *Obstet Gynecol*. 2000;96(3):351-358.
22. Dennerstein L, Smith AM, Morse CA, Burger HG. Sexuality and the menopause. *J Psychosom Obstet Gynaecol*. 1994;15(1):59-66.
23. Avis NE, Stellato R, Crawford S, Johannes C, Longcope C. Is there an association between menopause status and sexual functioning? *Menopause*. 2000;7(5):297-309.
24. Koster A, Garde K. Sexual desire and menopausal development: a prospective study of Danish women born in 1936. *Maturitas*. 1993;16(1):49-60.

25. Rekers H, Drogendijk AC, Valkenburg HA, Riphagen F. The menopause, urinary incontinence and other symptoms of the genito-urinary tract. *Maturitas*. 1992;15(2):101-111.
26. Kuh D, Cardozo L, Hardy R. Urinary incontinence in middle aged women: childhood enuresis and other lifetime risk factors in a British prospective cohort. *J Epidemiol Community Health*. 1999;53(8):453-458.
27. Milsom I, Ekelund P, Molander U, Arvidsson L, Areskoug B. The influence of age, parity, oral contraception, hysterectomy and menopause on the prevalence of urinary incontinence in women. *J Urol*. 1993;149(6):1459-1462.
28. Sherburn M, Guthrie JR, Dudley EC, O'Connell HE, Dennerstein L. Is incontinence associated with menopause? *Obstet Gynecol*. 2001;98(4):628-633.
29. Avis NE, Brambilla D, McKinlay SM, Vass K. A longitudinal analysis of the association between menopause and depression: results from the Massachusetts Women's Health Study. *Ann Epidemiol*. 1994;4(3):214-220.
30. Dennerstein L, Lehter P, Burger H, Dudley E. Mood and the menopausal transition. *J Nerv Ment Dis*. 1999;187(11):685-691.
31. Matthews KA, Wing RR, Kuller LH, Meilahn EN, Plantinga P. Influence of the perimenopause on cardiovascular risk factors and symptoms of middle-aged healthy women. *Arch Intern Med*. 1994;154(20):2349-2355.
32. Pearlstein T, Rosen K, Stone AB. Mood disorders and menopause. *Endocrinol Metab Clin North Am*. 1997;26(2):279-294.
33. Kaufert PA, Gilbert P, Tate R. The Manitoba Project: a re-examination of the link between menopause and depression. *Maturitas*. 1992;14(2):143-155.
34. Bosworth HB, Bastian LA, Kuchibhatla MN, et al. Depressive symptoms, menopausal status, and climacteric symptoms in women at midlife. *Psychosom Med*. 2001;63(4):603-608.
35. Hunter MS. Psychological and somatic experience of the menopause: a prospective study [corrected]. *Psychosom Med*. 1990;52(3):357-367.
36. Dennerstein L. Depression in the menopause. *Obstet Gynecol Clin North Am*. 1987;14(1):33-48.
37. Winokur G. Depression in the menopause. *Am J Psychiatry*. 1973;130(1):92-93.
38. Deutsh H. *The Psychology of Women: A Psychoanalytic Interpretation*. New York, NY: Grune & Stratton; 1945.
39. Ho SC, Chan SG, Yip YB, Cheng A, Yi Q, Chan C. Menopausal symptoms and symptom clustering in Chinese women. *Maturitas*. 1999;33(3):219-227.
40. Berg JA, Taylor DL. Symptom experience of Filipino American midlife women. *Menopause*. 1999;6(2):105-114.
41. Leidy LE. Menopausal symptoms and everyday complaints. *Menopause*. 1997;4(3):154-160.
42. Bromberger JT, Meyer PM, Kravitz HM, et al. Psychologic distress and natural menopause: a multiethnic community study. *Am J Public Health*. 2001;91(9):1435-1442.
43. Johnson SR. The epidemiology and social impact of premenstrual symptoms. *Clin Obstet Gynecol*. 1987;30(2):367-376.
44. Torgerson DJ, Thomas RE, Reid DM. Mothers and daughters menopausal ages: is there a link? *Eur J Obstet Gynecol Reprod Biol*. 1997;74(1):63-66.
45. Cramer DW, Harlow BL, Xu H, Fraer C, Barbieri R. Cross-sectional and case-controlled analyses of the association between smoking and early menopause. *Maturitas*. 1995;22(2):79-87.
46. Bergen AW, Caporaso N. Cigarette smoking. *J Natl Cancer Inst*. 1999;91(16):1365-1375.
47. McKinlay SM. The normal menopause transition: an overview. *Maturitas*. 1996;23(2):137-145.
48. Willett W, Stampfer MJ, Bain C, et al. Cigarette smoking, relative weight, and menopause. *Am J Epidemiol*. 1983;117(6):651-658.
49. Kaufman DW, Slone D, Rosenberg L, Miettinen OS, Shapiro S. Cigarette smoking and age at natural menopause. *Am J Public Health*. 1980;70(4):420-422.
50. Brambilla DJ, McKinlay SM. A prospective study of factors affecting age at menopause. *J Clin Epidemiol*. 1989;42(11):1031-1039.
51. Bromberger JT, Matthews KA, Kuller LH, Wing RR, Meilahn EN, Plantinga P. Prospective study of the determinants of age at menopause. *Am J Epidemiol*. 1997;145(2):124-133.
52. Hiatt RA, Fireman BH. Smoking, menopause, and breast cancer. *J Natl Cancer Inst*. 1986;76(5):833-838.
53. Adena MA, Gallagher HG. Cigarette smoking and the age at menopause. *Ann Hum Biol*. 1982;9(2):121-130.
54. Gold EB, Bromberger J, Crawford S, et al. Factors associated with age at natural menopause in a multiethnic sample of midlife women. *Am J Epidemiol*. 2001;153(9):865-874.
55. Jensen J, Christiansen C, Rodbro P. Cigarette smoking, serum estrogens, and bone loss during hormone-replacement therapy early after menopause. *N Engl J Med*. 1985;313(16):973-975.
56. Michnovicz JJ, Hershcopf RJ, Naganuma H, Bradlow HL, Fishman J. Increased 2-hydroxylation of estradiol as a possible mechanism for the anti-estrogenic effect of cigarette smoking. *N Engl J Med*. 1986;315(21):1305-1309.
57. MacMahon B, Trichopoulos D, Cole P, Brown J. Cigarette smoking and urinary estrogens. *N Engl J Med*. 1982;307(17):1062-1065.
58. Barbieri RL, McShane PM, Ryan KJ. Constituents of cigarette smoke inhibit human granulosa cell aromatase. *Fertil Steril*. 1986;46(2):232-236.
59. Friedman AJ, Ravnikar VA, Barbieri RL. Serum steroid hormone profiles in postmenopausal smokers and nonsmokers. *Fertil Steril*. 1987;47(3):398-401.
60. Hardy R, Kuh D, Wadsworth M. Smoking, body mass index, socioeconomic status and the menopausal transition in a British national cohort. *Int J Epidemiol*. 2000;29(5):845-851.
61. Kato I, Toniolo P, Akhmedkhanov A, Koenig KL, Shore R, Zeleniuch-Jacquotte A. Prospective study of factors influencing the onset of natural menopause. *J Clin Epidemiol*. 1998;51(12):1271-1276.
62. Torgerson DJ, Avenell A, Russell IT, Reid DM. Factors associated with onset of menopause in women aged 45-49. *Maturitas*. 1994;19(2):83-92.
63. Oldenhove A, Jaszmann LJ, Everaerd WT, Haspels AA. Hysterectomized women with ovarian conservation report more severe climacteric complaints than do normal climacteric women of similar age. *Am J Obstet Gynecol*. 1993;168(3 pt 1):765-771.
64. Hartmann BW, Kirchengast S, Albrecht A, Metka M, Huber JC. Hysterectomy increases the symptomatology of postmenopausal syndrome. *Gynecol Endocrinol*. 1995;9(3):247-252.
65. Pokras R. Hysterectomy: past, present and future. *Stat Bull Metrop Insur Co*. 1989;70(4):12-21.
66. Nilsson K, Risberg B, Heimer G. The vaginal epithelium in the postmenopause—cytology, histology and pH as methods of assessment. *Maturitas*. 1995;21(1):51-56.
67. Caillouette JC, Sharp CF Jr, Zimmerman GJ, Roy S. Vaginal pH as a marker for bacterial pathogens and menopausal status. *Am J Obstet Gynecol*. 1997;176(6):1270-1275; discussion 1275-1277.
68. Panyakhamlerd K, Chotnopparatpattara P, Taechakraichana N, Kukulprasong A, Chaikittilap S, Limpaphayom K. Skin thickness in different menopausal status. *J Med Assoc Thai*. 1999;82(4):352-356.
69. Hee J, MacNaughton J, Bangah M, Burger HG. Perimenopausal patterns of gonadotrophins, immunoreactive inhibin, oestradiol and progesterone. *Maturitas*. 1993;18(1):9-20.
70. Sherman BM, Korenman SG. Hormonal characteristics of the human menstrual cycle throughout reproductive life. *J Clin Invest*. 1975;55(4):699-706.
71. Burger HG, Cahir N, Robertson DM, et al. Serum inhibins A and B fall differentially as FSH rises in perimenopausal women. *Clin Endocrinol (Oxf)*. 1998;48(6):809-813.
72. Illingworth PJ, Reddi K, Smith KB, Baird DT. The source of inhibin secretion during the human menstrual cycle. *J Clin Endocrinol Metab*. 1991;133(3):667-673.
73. Holleman DR Jr, Simel DL. Does the clinical examination predict airflow limitation? *JAMA*. 1995;273(4):313-319.
74. Chompootweep S, Tankeyoon M, Yamarat K, Poomsuwan P, Dusitsin N. The menopausal age and climacteric complaints in Thai women in Bangkok. *Maturitas*. 1993;17(1):63-71.
75. Dennerstein L, Smith AM, Morse C, et al. Menopausal symptoms in Australian women. *Med J Aust*. 1993;159(4):232-236.
76. Punyahotra S, Dennerstein L, Lehter P. Menopausal experiences of Thai women, part 1: symptoms and their correlates. *Maturitas*. 1997;26(1):1-7.
77. Stellato R, Crawford SL, McKinlay SM, Longcope C. Can follicle-stimulating hormone be used to define menopausal status? *Endocr Pract*. 1998;4(3):137-141.
78. Maartens LW, Leusink GL, Knotterus JA, Smeets CG, Pop VJ. Climacteric complaints in the community. *Fam Pract*. 2001;18(2):189-194.
79. Eddy D, Hasselblad V. *Fast*Pro Version, 1.8: Software for Meta-analysis by the Confidence Profile Method*. San Diego, CA: Academic Press; 1992.
80. Klein J, Sauer MV. Assessing fertility in women of advanced reproductive age. *Am J Obstet Gynecol*. 2001;185(3):758-770.
81. American College of Obstetricians and Gynecologists. *The Use of Hormonal Contraception in Women With Coexisting Medical Conditions*. Washington, DC: ACOG; 2000. ACOG Practice Bulletin No. 18.

UPDATE : Menopause 31

Prepared by Lori A. Bastian, MD, PhD

Reviewed by Joanne T. Piscitelli, MD

CLINICAL SCENARIO

A 42-year-old female patient presents to your clinic with concern about whether she is starting menopause. She does observe that her periods are lasting longer (approximately 8-9 days), but they continue to occur at regular intervals, every 28 days. Her mother started menopause at age 48 years. The patient has noticed no symptoms of menopause such as hot flashes or night sweats. She does not smoke. She ordered a home-testing menopause kit via the Internet, and the results suggested that she is “starting menopause.” She wants to know about the accuracy of these kits and what type of changes she should expect during the next year.

UPDATED SUMMARY ON PERIMENOPAUSE

Original Review

Bastian LA, Smith C, Nanda K. Is this woman perimenopausal? *JAMA*. 2003;289(7):895-902.

UPDATED LITERATURE SEARCH

Details of the Update

Our literature search used the parent search strategy for The Rational Clinical Examination series, combined with the subject “perimenopause,” published in English from 2002 to September 2004. The results yielded 499 titles, for which we reviewed the titles and abstracts; 36 articles were selected for additional review. These articles were reviewed to identify studies that assessed the sensitivity and specificity of the medical history or physical examination features of perimenopause, defined as greater than 3 (but < 12) months of amenorrhea or menstrual irregularity. Only 2 articles on the perimenopause were retained.^{1,2} The remaining articles did not measure perimenopause or presented mean values that could not be used in 2 × 2 tables.

Many women use home-testing kits to assess their menopausal status, making the results of home testing part of the clinical history. A Google search revealed 13 100 sites for “menopause diagnostic kits,” yet there are no reports of

the effectiveness of these kits. We summarized the results of this search strategy after exploring the Web sites of available kits.

NEW FINDINGS

The effectiveness of menopause home diagnostic kits (based on urine tests of follicle-stimulating hormone [FSH]) has not been published.

IMPROVEMENTS IN THE DATA PRESENTED IN THE ORIGINAL PUBLICATION

Age is an important factor for perimenopause, and the Wise et al¹ article measured incidence of perimenopause among women aged 36 to 45 years.

CHANGES IN THE REFERENCE STANDARD

- The reference standard, which is based on menstrual history, remains the same: 3 to 11 months of amenorrhea or irregular periods.
- As noted in the original review, a panel of experts (from the Stages of Reproductive Aging Workshop) proposed a new system to classify reproductive aging that uses age, menstrual history, and FSH and estradiol levels.³ The system is of uncertain validity because large categories of women, such as cigarette smokers and obese women, are excluded. More recently, the Women’s Ischemia Syndrome Evaluation (WISE) study developed a new algorithm for classifying menopausal status.⁴ The apparent advantage of the new staging system is the ability to diagnose perimenopause in women who have had a hysterectomy. Using an expert consensus panel as the reference standard, WISE’s hormonal algorithm had a sensitivity of 88% and specificity of 97% for diagnosing perimenopause.

RESULTS OF LITERATURE REVIEW

Among women aged 45 to 55 years, self-rating of any decline in personal health has no predictive value for identifying

Table 31-3 The Incidence of Perimenopause Related to Age

Age, y	Incidence Rate of Perimenopause (%)
36-37	19/90 (21)
38-39	34/162 (21)
40-41	45/137 (33)
42-43	53/153 (35)
44-45	26/61 (43)

Table 31-4 Likelihood Ratios of Features for Perimenopause

Finding	Development of Perimenopause Within 36 Months	
	LR+ (95% CI)	LR- (95% CI)
Nonwhite	2.5 (1.3-4.8)	0.9 (0.9-1.0)
Family history of early menopause	2.0 (1.1-3.5)	0.9 (0.9-1.0)
BMI \geq 30	1.8 (1.1-2.8)	0.9 (0.9-1.0)
Age, y		
42-45	1.4 (1.1-1.5)	0.8 (0.7-0.9)
40-45	1.3 (1.2-1.5)	0.6 (0.5-0.8)
>5 y passive smoke exposure	1.4 (1.1-1.8)	0.8 (0.7-1.0)
Current smoker	1.4 (0.9-2.3)	0.9 (0.9-1.0)
Depression (<i>DSM-III</i> defined) ⁵	1.2 (1.0-1.6)	0.9 (0.8-1.0)
Decline in self-rated health	1.2 (0.6-2.7)	0.9 (0.8-1.1)

Abbreviations: BMI, body mass index; CI, confidence interval; *DSM-III*, *Diagnostic and Statistical Manual of Mental Disorders* (Third Edition); LR+, positive likelihood ratio; LR-, negative likelihood ratio.

perimenopause (likelihood ratio [LR] approaches 1).² The incidence of perimenopause among women 35 to 40 years of age is approximately 20% (Table 31-3); a family history of early menopause in the mother has an LR of 2.0 for identifying younger women (age 36-45 years) who might become perimenopausal during the ensuing 36 months (Table 31-4).¹

Menopause home diagnostic kits are popular in the lay health literature. These tests measure FSH levels, and results are considered “positive” when the FSH is elevated and in a menopausal range. The accuracy of these tests compares the test result to laboratory-based FSH measures, and they are reviewed by the Food and Drug Administration (FDA). The home testing kits approved by the FDA have over 90% accuracy for the home test result compared to a test result obtained in a laboratory (see <http://www.fda.gov/cdrh/orid/homeuse-menopause.html>; accessed June 3, 2008).

EVIDENCE FROM GUIDELINES

The US Preventive Services Task Force recommends counseling women approaching the menopausal transition. The evidence report did not address the diagnosis of the menopausal transition.

CLINICAL SCENARIO—RESOLUTION

This scenario describes a 42-year-old woman with no symptoms but with minor changes in her menstrual flow. According to her age, the pretest probability of being either perimenopausal or postmenopausal is 35%. Her menopause home diagnostic kit result was positive, which may suggest an FSH level greater than or equal to 25 mIU/mL (corresponding positive LR, 3.1). Although clinicians should rely on the medical history and demographic features to assess menopause without routine FSH testing, this patient has provided a home test result. With the information from her home diagnostic kit, her calculated posttest probability of perimenopause is 62%. She might experience more irregularity of her periods during the next 1 to 2 years, with resulting amenorrhea.

PERIMENOPAUSE—MAKE THE DIAGNOSIS

PRIOR PROBABILITY

The probability of menopause is estimated best from the patient's age (see Figure 31-1).

At age 36 to 39 years, the incidence is approximately 20%; 40 to 43 years, approximately 34%; and 44 to 45 years, 43%.

POPULATION FOR WHOM PERIMENOPAUSE SHOULD BE CONSIDERED

- Women with irregular menses or amenorrhea for more than 3 months
- Women with hot flashes or night sweats
- Those who have had hysterectomy
- Those undergoing chemotherapy

FINDINGS FOR PERIMENOPAUSE

Most findings other than age have low accuracy for identifying women in perimenopause. The presence of a family history of early menopause or hot flashes and the results of a home test for FSH may be the best findings (Table 31-5).

Table 31-5 Likelihood Ratios of Findings for Perimenopause

Finding	Development of Perimenopause Within 36 Months	
	LR+ (95% CI) or Range	LR- (95% CI) or Range
FSH (≥ 24 mIU/mL) ^a	3.1 (2.1-4.5)	0.45 (0.36-0.56)
Family history of early menopause	2.0 (1.1-3.5)	0.9 (0.9-1.0)
Hot flashes	2.1-4.1	0.54-0.87

Abbreviations: CI, confidence interval; FSH, follicle-stimulating hormone; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

^aThe role of routine hormonal testing for diagnosing perimenopausal status has not been established. The FSH may prove most useful for women after a hysterectomy because they cannot report menstrual symptoms. Home testing kits approved by the Food and Drug Administration have over 90% accuracy for the home test result compared to a test result obtained in a laboratory.

REFERENCE STANDARD TEST

Menstrual history.

REFERENCES FOR THE UPDATE

1. Wise LA, Krieger N, Zierler S, Harlow BL. Lifetime socioeconomic position in relation to onset of perimenopause. *J Epidemiol Community Health*. 2002;56(11):851-860.^a
2. Dennerstein L, Dudley EC, Guthrie JR. Predictors of declining self-rated health during the transition to menopause. *J Psychosom Res*. 2003;54(2):147-153.^a
3. Soules MR, Sherman S, Parrott E, et al. Executive summary: Stages of Reproductive Aging Workshop (STRAW). *Fertil Steril*. 2001;76(5):874-878.
4. Johnson BD, Bairey-Merz CN, Braunstein GD, et al. Determination of menopausal status in women: the NHLBI-sponsored Women's Ischemia Syndrome Evaluation (WISE) Study. *J Womens Health*. 2004;13(8):872-887.
5. Spitzer RL, Williams JB, Gibbon M, et al. The structured clinical interview for DSM-III-R (SCID), I: history, rationale, and description. *Arch Gen Psychiatry*. 1992;49(8):624-629.

^aFor the Evidence to Support the Update for this topic, see <http://www.JAMAevidence.com>.

This page intentionally left blank

EVIDENCE TO SUPPORT THE UPDATE:

31

Menopause

TITLE Predictors of Declining Self-rated Health During the Transition to Menopause.

AUTHORS Dennerstein L, Dudley EC, Guthrie JR.

CITATION *J Psychosom Res.* 2003;54(2):147-153.

QUESTION What is the role of declining self-rated health in the diagnosis of perimenopause?

DESIGN All data were collected prospectively in an 8-year cohort study called the Melbourne Women's Midlife Health Project. Self-rated health was measured annually.

SETTING Population-based cohort of middle-aged (aged 45-55 years at baseline) Australian-born women.

PATIENTS Two hundred sixty-two women completed the year 8 self-rated health assessment article; 136 women were perimenopausal (3-11 months of amenorrhea) and 44 were in the premenopausal control group. Exclusions were incomplete data, surgical menopause, and hormone therapy use.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

On a mailed questionnaire, women were asked to rate their present health compared with that of other women about the same age as follows: worse than most, about the same as others, or better than most.

MAIN OUTCOME MEASURES

Sensitivity, specificity, and likelihood ratios.

MAIN RESULTS

Women's perception of a decline in their health does not indicate they are perimenopausal ([Table 31-6](#)).

Table 31-6 Likelihood Ratio for Self-Rated Decline in Health as a Predictor for Perimenopausal State

Test	Sensitivity	Specificity	LR+ (95% CI)	LR- (95% CI)
Decline in self-rated health	0.20	0.84	1.2 (0.6-2.7)	0.9 (0.8-1.1)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

CONCLUSIONS

LEVEL OF EVIDENCE Level 1.

STRENGTHS Well-respected cohort study with measures of self-rated health completed prospectively. Perimenopause was determined independently.

LIMITATIONS None.

There is a low sensitivity for decline in self-rated health as a predictor of change to perimenopause. A decline in self-rated health does not identify women who are perimenopausal.

Reviewed by Lori A. Bastian, MD

TITLE Lifetime Socioeconomic Position in Relation to Onset of Perimenopause.

AUTHORS Wise LA, Krieger N, Zierler S, Harlow BL.

CITATION *J Epidemiol Community Health.* 2002;56(11): 851-860.

QUESTION What is the association between demographic, behavioral, and reproductive factors and onset of perimenopause?

DESIGN All data were collected prospectively in a cohort study designed to assess the association between major depression and ovarian function among women of late reproductive age.

SETTING A mailed questionnaire to a random sample of 6228 women aged 36 to 45 years, residing in 7 Boston-area communities from 1996 to 1997 (94% white).

PATIENTS Seven hundred thirty-three women (81% response rate) completed the follow-up survey. Women were excluded if they were pregnant, had a hysterectomy or surgical menopause, had menopausal irregularity at the baseline survey, had medical menopause, underwent fertility therapy, or started hormone therapy.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

Baseline demographic and reproductive characteristics were measured, such as age, race/ethnicity, family history of early menopause (defined as mother or sister with natural menopause before age 46 years), smoking history, body mass index, passive smoke exposure, and depression defined by the Structured Clinical Interview applied to *Diagnostic and Statistical Manual of Mental Disorders* (Fourth Edition) criteria.¹

Perimenopausal status was measured during a 36-month follow-up by subjective report of menstrual irregularity or an absolute change of 7 days or greater in menstrual cycle length compared with baseline, a change in menstrual flow amount, or periods of amenorrhea lasting 3 to 6 months.

MAIN OUTCOME MEASURES

Sensitivity, specificity, likelihood ratios (LRs), and incidence of perimenopause by age categories.

MAIN RESULTS

Of 603 women, 177 (29%) developed perimenopause during the 36-month follow-up period. Twenty percent of women ages 36 to 39 years were perimenopausal (Table 31-7). Baseline demographic, family history, smoking history, and the presence of depression were not particularly useful for identifying perimenopause (Table 31-8).

CONCLUSIONS

LEVEL OF EVIDENCE Level 1.

STRENGTHS Large cohort study with all variables entered prospectively.

Table 31-7 Incidence of Perimenopause Related to Age

Age, y	Incidence Rate of Perimenopause (%)
36-37	19/90 (21)
38-39	34/162 (21)
40-41	45/137 (33)
42-43	53/153 (35)
44-45	26/61 (43)

Table 31-8 Likelihood Ratio of Features for Perimenopause

Finding	Sensitivity	Specificity	Development of Perimenopause Within 36 Months	
			LR+ (95% CI)	LR- (95% CI)
Nonwhite	0.10	0.96	2.5 (1.3-4.8)	0.9 (0.9-1.0)
Family history early menopause	0.12	0.94	2.0 (1.1-3.5)	0.9 (0.9-1.0)
BMI ≥30	0.16	0.91	1.8 (1.1-2.8)	0.9 (0.9-1.0)
Age, y				
42-45	0.45	0.68	1.4 (1.1-1.7)	0.8 (0.7-0.9)
40-45	0.70	0.47	1.3 (1.2-1.5)	0.6 (0.5-0.8)
>5 y passive smoke exposure	0.37	0.74	1.4 (1.1-1.8)	0.8 (0.7-1.0)
Current smoker	0.13	0.91	1.4 (0.9-2.3)	0.9 (0.9-1.0)
Depression (DSM-IV defined) ¹	0.37	0.70	1.2 (1.0-1.6)	0.9 (0.8-1.0)

Abbreviations: BMI, body mass index; CI, confidence interval; *DSM-IV*, *Diagnostic and Statistical Manual of Mental Disorders* (Fourth Edition); LR+, positive likelihood ratio; LR-, negative likelihood ratio.

LIMITATIONS Perimenopause definition required only 3 to 6 months of amenorrhea. Estimates of perimenopause are more liberal than previous studies requiring 3 to 11 months of amenorrhea.

The most important finding of this study is that the incidence rates for perimenopause among women 36 to 45 years of age may be higher than is appreciated by many generalist physicians and their patients.

Women with menstrual irregularity at baseline were excluded from this cohort study. Therefore, the results should not be used to estimate the probability that a woman with ongoing menstrual irregularity is actually perimenopausal. Although it may seem awkward to use LRs to describe the utility of these demographic features and historical items, the values can be applied to women aged 36 to 45 years and express the increased likelihood of *developing* perimenopause.

Reviewed by Lori A. Bastian, MD

REFERENCE FOR THE EVIDENCE

- Spitzer RL, Williams JB, Gibbon M, et al. The structured clinical interview for DSM-III-R (SCID), I: history, rational, and description. *Arch Gen Psychiatry*. 1992;49(8):624-629.

CHAPTER 32

Does This Patient Have Aortic Regurgitation?

Niteesh K. Choudhry, MD

Edward E. Etchells, MD, MSc

CLINICAL SCENARIO

You are asked to see a 59-year-old woman with liver cirrhosis and esophageal varices. When she was checked into the clinic, she had a pulse pressure of 70 mm Hg. Because of the wide pulse pressure, you wonder if she has aortic regurgitation (AR). You conduct a complete physical examination and hear no early-diastolic murmur in the third or fourth intercostal spaces at the left sternal border. You suspect that the wide pulse pressure is a peripheral hemodynamic consequence of cirrhosis, not AR. Do you need an echocardiogram to confirm your clinical impression that she does not have AR?

WHY IS THE CLINICAL EXAMINATION IMPORTANT IN EVALUATING FOR AORTIC REGURGITATION?

Aortic regurgitation is a potentially serious cardiac abnormality that may be caused by important underlying disorders. Patients with AR require careful clinical monitoring to identify the optimal time for surgical intervention. Asymptomatic patients with severe AR may benefit from vasodilator therapy.¹

The use of noninvasive cardiac testing, such as echocardiography, has increased in recent years. It is estimated that 2% of the general population undergo noninvasive cardiac diagnostic evaluation annually.² If a careful clinical examination can exclude the presence of AR, then there would be no need to proceed with further cardiac evaluation.

Anatomic and Physiologic Origins of Diastolic Murmurs

The cardinal manifestation of AR is a diastolic murmur. Diastolic murmurs are important indicators of structural cardiac abnormalities or pathologic states of increased flow ([Table 32-1](#)). As discussed in a previous article in this series,³ heart murmurs are produced when turbulent blood flow causes prolonged auditory vibrations of cardiac structures. The intensity of the murmur depends on many factors, including blood viscosity, blood flow velocity and turbulence, the distance between the vibrations and the stethoscope, the angle at which the vibrations meet the stethoscope, the transmission qualities of the tissue between the vibration and the stethoscope, and the auditory skills of the examiner.⁴

How to Examine for Aortic Regurgitation

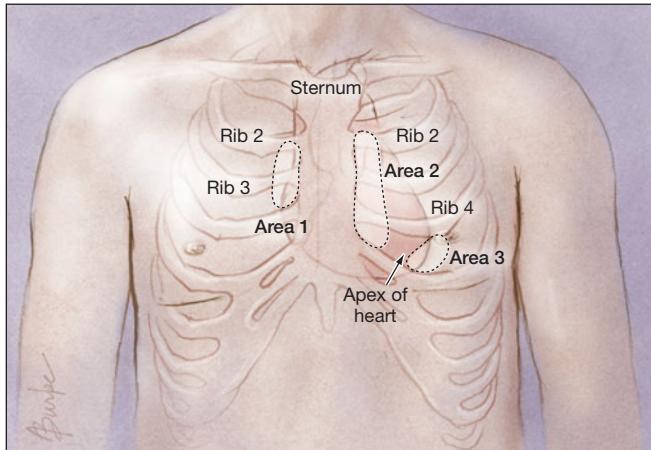
A complete clinical history and physical examination are essential in the evaluation of a patient with a diastolic murmur. A diastolic murmur in a patient with renal failure and volume overload will have different significance than a diastolic murmur in a patient with a history of rheumatic fever and atrial fibrillation.

The examiner's ability to detect a diastolic murmur can be undermined by environmental factors such as noisy rooms,

Table 32-1 Selected Causes of Diastolic Murmurs

Abnormal cardiac structure
Aortic regurgitation
Mitral stenosis
Pulmonic regurgitation
Tricuspid stenosis
Atrial myxoma
Ventricular septal defect ^a
Atrial septal defect ^a
Mitral regurgitation ^a
Normal cardiac structure, increased flow
Renal failure with volume overload
Thyrotoxicosis
Anemia
Sepsis

^aDiastolic murmurs are caused by abnormally increased diastolic flow across the mitral or tricuspid valves.

**Figure 32-1** Typical Location of Abnormal Diastolic Murmurs

There are 3 important areas to auscultate for diastolic murmurs. Area 1 is the second and third intercostal spaces at the right sternal border. Area 2 is the second and fourth intercostal spaces at the left sternal border. Aortic regurgitation murmur may be heard in both areas 1 and 2. If the murmur is loudest in area 1, then the underlying cause of aortic regurgitation may be an ascending aortic aneurysm or aortic dissection. Pulmonic regurgitation murmurs are loudest in the superior part of area 2 and may radiate downward. The murmur of mitral stenosis and the Flint murmur of aortic regurgitation are best heard at the apex (area 3).

examiner factors such as fatigue or haste, and patient factors such as dyspnea or tachycardia.⁵ If examining conditions are not optimal, the examination should be repeated when conditions improve.

The precision and accuracy of many components of the examination for AR, including all of the cardiac history and most of the physical examination, have not been adequately evaluated. This article will focus on aspects of the cardiac physical examination that have been sufficiently assessed for precision or accuracy.

Cardiac Auscultation

During routine auscultation, the examiner attempts to detect a diastolic murmur. Diastole is the period that begins with the closure of the aortic and pulmonic valves (second heart sound [S_2]) and ends with the closure of the mitral and tricuspid valves (first heart sound [S_1]). A common maneuver used to identify diastole is to palpate the carotid artery pulse during auscultation; S_1 is synchronous with the carotid artery pulsation, whereas S_2 follows the pulse. A diastolic murmur is a diastolic sound longer than a heart sound. Examiners should describe the grade, location of maximal intensity (Figure 32-1), timing (Figure 32-2), duration, pitch, and radiation of the murmur.

The Levine grading system,⁶ with slight modifications,⁷ was developed for systolic murmurs but may also be used to describe diastolic murmurs. A grade 1 murmur is not heard immediately on auscultation but is heard after the examiner focuses for a few seconds. Grade 2 murmurs are heard immediately on auscultation but are softer than the loud grade 3. Grade 4 murmurs are associated with a palpable precordial vibration called a *thrill*. Grades 5 and 6 murmurs are also associated with a thrill. A grade 5 murmur is audible when only one edge of the stethoscope is on the chest, and a grade 6 murmur is audible with the entire stethoscope lifted off the chest.

The typical AR murmur is an early-diastolic, decrescendo blowing sound (Figure 32-2) that may be accentuated with the patient sitting upright and leaning forward.⁸ In some cases, S_2 can be obscured by the murmur. Most AR murmurs are high pitched and are best heard with the diaphragm of the stethoscope placed firmly on the chest wall. Some AR murmurs are low pitched and are better heard with the bell of the stethoscope placed lightly on the chest wall. For example, the AR murmur associated with endocarditis and a fenestrated aortic valve can be low pitched.

The examiner should apply the stethoscope to the chest wall in the third or fourth intercostal space at the left sternal border and listen between normal breaths at the end of expiration. The patient should not voluntarily breath-hold because it may inadvertently create a Valsalva maneuver. If the murmur is louder at the second to third right intercostal space, the underlying cause of AR may be an ascending aortic aneurysm or aortic dissection.⁹

Aortic regurgitation also may be associated with a systolic murmur,¹⁰ created by the flow of an abnormally large volume of blood through a nonstenotic aortic valve or a bicuspid aortic valve. The murmur is an early-peaking, crescendo-decrescendo systolic sound that is best heard with the diaphragm of the stethoscope applied to the second right intercostal space.

The Flint murmur is a low-pitched late-diastolic apical murmur, which is associated with AR. The murmur is likely produced when the regurgitant jet of blood collides with the left ventricular endocardium.¹¹ The murmur may have a mid-diastolic component, but the original description by Flint¹² referred only to “presystolic blubbling.”¹² It is best heard with the patient in the left-lateral decubitus position, using the bell of the stethoscope. Differentiating the Flint murmur from the

murmur of mitral stenosis can be difficult. The murmur of mitral stenosis is primarily mid-diastolic (possibly with a late-diastolic component) and may be associated with an opening snap (OS) and a loud S₁ (Figure 32-2).¹³

The typical murmur of pulmonic regurgitation (PR) is an early-diastolic decrescendo murmur heard best in the second-left intercostal space at the sternal border. The murmur may radiate to the third and fourth left intercostal spaces and may increase during quiet inspiration. If there is splitting of S₂, the astute examiner may note that the murmur begins after the pulmonic valve component (P₂) of S₂ rather than the aortic component. The murmur of PR may be lower pitched than the murmur of AR, unless pulmonary hypertension is present. A right-sided Flint murmur can be heard, particularly in patients with pulmonary hypertension.

Mitral stenosis is associated with a mid-diastolic, decrescendo, low-frequency rumble, which, if the patient is in sinus rhythm, may be followed by late-diastolic (presystolic) crescendo that ends with the mitral component of S₁ (Figure 32-2). It is best heard using the bell of the stethoscope placed at the apex soon after moving the patient into the left lateral decubitus position. Rolling the patient onto the left side brings the left ventricle closer to the chest wall and serves as a form of exercise, increasing blood flow across the mitral valve and increasing the murmur's intensity.⁹ The murmur of mitral stenosis may be inaudible in patients with low cardiac output.

The S₁ may be increased in intensity in mitral stenosis.¹³ A normal S₁ is best appreciated near the apex, where it should be louder than S₂. The S₁ is normally softer than S₂ in the second right and second left intercostal spaces adjacent to the sternum. If S₁ is as loud as or louder than the S₂ in these areas, then the S₁ is increased in intensity.

An OS is a high-frequency, early-diastolic sound that is associated with the opening of a stenotic mitral valve. It occurs 50 to 100 ms after the aortic valve component (A₂) of S₂ and is best heard in the area from the left sternal border to the apex. Much like the murmur of mitral stenosis, it may be accentuated by auscultating while the patient is in the left lateral decubitus position shortly after the patient has performed exercise. The A₂-OS interval shortens with increasing severity of mitral stenosis. The OS may be absent in the case of a heavily calcified immobile mitral valve. It is often difficult to differentiate an OS from the P₂ of S₂. The OS usually decreases in intensity with inspiration and S₂-OS interval widens on standing. Conversely, P₂ becomes louder with inspiration, and the A₂-P₂ interval remains the same or narrows with standing.¹³ In addition, P₂ is not expected to be heard at the apex unless the patient has pulmonary hypertension.

Maneuvers

Selective use of maneuvers can enhance the detection and interpretation of diastolic murmurs. There is no point in doing maneuvers if a loud AR murmur has been detected during routine auscultation. However, if the clinician is unsure about the presence of a faint diastolic murmur, then a maneuver that increases murmur intensity may clarify the

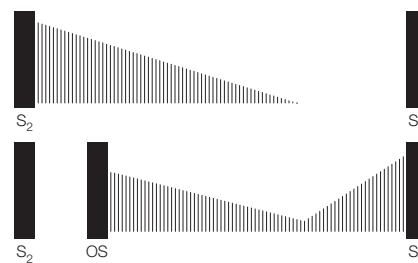


Figure 32-2 Selected Features of Diastolic Murmurs

Diastolic murmurs are classified according to the time of onset of the murmur.¹⁴ An early diastolic murmur begins with the second heart sound (S₂). Top, Early diastolic murmurs typically decrease in intensity (decrescendo) and disappear before the first heart sound (S₁). In some cases, an early diastolic murmur can continue through diastole. Bottom, A mid-diastolic murmur begins clearly after S₂ (in mitral stenosis, classically after an opening snap [OS]). A late-diastolic (or presystolic) murmur begins in the interval immediately before S₁. In mitral stenosis, the mid-diastolic murmur may merge with the late-diastolic (presystolic) murmur.

situation. If the clinician has a heightened suspicion for AR (eg, after hearing an aortic ejection sound), or if examining conditions are not optimal, then a maneuver to augment murmur intensity might bring out an otherwise inaudible murmur. Finally, the maneuvers may help distinguish PR from AR. In this latter situation, the clinician should listen where the murmur is just barely audible, so that it is easy to detect a decrease or increase in murmur intensity during the maneuver.

Quiet inspiration increases venous return and augments right-sided heart murmurs such as PR. To determine the effect of inspiration on the intensity of the murmur, the examiner should listen during quiet inspiration, rather than asking the patient to breathe deeply, because the murmur may be obscured by breath sounds.

Transient arterial occlusion primarily increases systemic arterial resistance that intensifies left-sided regurgitant lesions such as AR and may help distinguish the murmur from PR. To perform this maneuver, sphygmomanometers are placed around both of the patient's arms and are inflated to 20 to 40 mm Hg above the previously recorded systolic blood pressure. Any changes in murmur intensity are noted 20 seconds after cuff inflation.¹⁵

Peripheral Hemodynamic Signs

There are a variety of peripheral hemodynamic signs traditionally associated with AR. Some of these signs have been adequately evaluated, including de Musset head-bobbing sign,¹⁶ a wide pulse pressure,¹⁷ the brachial-popliteal pulse gradient (Hill sign¹⁸), Duroziez femoral murmur,¹⁶ the femoral pistol shot murmur,¹³ and Corrigan water hammer pulse.¹⁹ The de Musset head-bobbing sign consists of a forward shaking of the head with every heartbeat; it is best observed in patients who are sitting.¹⁶

Pulse pressure refers to the difference between systolic and diastolic blood pressures. A widened pulse pressure may be defined as greater than 50 mm Hg.²⁰ Other definitions include a pulse pressure greater than 50% of the systolic

pressure.¹⁷ Determination of the blood pressure has been described in another article in this series.²¹

The brachial-popliteal pulse gradient (Hill sign) can be defined as a systolic blood pressure in the lower extremities that is at least 20 mm Hg higher than that in the arms.²⁰ To determine a popliteal blood pressure, an appropriately sized blood pressure cuff should be placed on the patient's thigh¹³ with the artery marker over the popliteal artery. The cuff should be inflated and the systolic pressure can then be determined in the popliteal fossa either by palpation, as judged by the point where the pulse reappears as the cuff is deflated, or by auscultation, listening for Korotkoff sounds to appear. Both the brachial and popliteal blood pressures should be measured while the patient is supine. The average of repeated readings should be used, especially in patients with irregular heart rates, such as atrial fibrillation.

Duroziez double intermittent femoral bruit is elicited by first gently compressing the femoral artery with the diaphragm of the stethoscope. This will yield a systolic bruit in all patients. As increasing pressure is applied to the diaphragm, an early-diastolic bruit will become audible in patients with AR. While listening to the diastolic bruit, the clinician should tilt the stethoscope so that the distal rim (closest to the patient's feet) is compressing the femoral artery. If the bruit becomes louder with this maneuver, then the diastolic bruit is due to the retrograde flow of blood toward the heart in AR. The stethoscope should then be tilted such that the proximal rim (closest to the patient's head) is compressing the femoral artery. If the diastolic bruit becomes softer, this can be taken as supportive evidence of the presence of retrograde blood flow. If, however, the bruit becomes louder with proximal pressure (and softer with distal pressure), then this sign should not be used as evidence of AR but may indicate the presence of a high-flow state such as renal failure with volume overload.²²

Femoral pistol shot sounds are elicited by auscultating with the diaphragm of the stethoscope over the femoral arteries. A high-pitched pistol shot sound may be heard in AR. Corrigan water hammer pulse refers to an increased volume and rate of increase of the radial pulse when the wrist is elevated perpendicular to the body of a supine patient. The radial pulse should first be assessed while the patient is lying supine with his or her arms resting at the sides. Sufficient pressure should be applied to obliterate the pulse. While this pressure is maintained, the patient's arm should be elevated so that it is perpendicular to the plane of the body. In AR, the pulse will become palpable despite applying an equivalent amount of pressure as when the arm was at the patient's side.

Other peripheral hemodynamic signs, such as Mayne sign (a decrease in diastolic blood pressure of 15 mm Hg when the arm is held above the head compared with when the arm is held at the level of the heart),²³ Quincke capillary pulsation, Muller pulsatile uvula, and Rosenbach liver pulsation, have not been adequately evaluated for precision or accuracy.

METHODS

To identify articles pertaining to the precision and accuracy of the physical examination for AR, we used standard meth-

ods for conducting research overviews.²⁴ Our data collection strategy involved 3 steps and was deliberately broad to reduce the possibility of overlooking important articles. First, we searched MEDLINE for English-language articles published from 1966 through July 1997, using a structured search strategy (available on request from the authors). Second, we manually reviewed potentially relevant articles and their reference lists. Third, we contacted the authors of relevant studies for additional information. Studies were excluded if they were review articles, involved patients younger than 18 years, were small (ie, <20 participants), involved prosthetic heart valves, had no clinical examination performed or reported, or had no acceptable reference standard (Doppler echocardiography or cardiac catheterization).

Studies were independently reviewed for methodologic quality by the 2 authors, and disagreements were resolved by consensus. Quality grades were assigned using published guidelines (see Table 1-7 for a summary of Evidence Grades and Levels).²⁵ Grade A studies involve the independent comparison of a sign or symptom with a reference standard of diagnosis among a large number of consecutive patients suspected of having the target condition. Grade B studies meet the criteria for grade A studies but have a small number of patients. Grade C studies involve nonconsecutive patients, patients who are known to have the target condition and healthy individuals, nonindependent comparisons between the sign or symptom and the reference standard, or nonindependent comparisons with a reference standard of uncertain validity. Grade C studies tend to overestimate the accuracy of the sign or symptom.

We created contingency tables for all studies and determined the likelihood ratios (LRs) for aortic regurgitation.^{26,27} We also sought information on the examination for other causes of diastolic murmurs, such as mitral stenosis or PR. Unfortunately, we found few studies of sufficient methodologic quality for these conditions. This relative lack of information implies that methodologically sound studies are needed but does not imply that the clinical examination for these conditions is imprecise, inaccurate, or unimportant.

Precision of the Examination Related to Diastolic Murmurs

Precision refers to agreement regarding a particular clinical finding between different physicians (interobserver) or between multiple assessments by the same physician (intraobserver). The precision of the clinical examination for diastolic murmurs has been evaluated in usual clinical situations by auscultating patients^{28,29} or in controlled nonclinical circumstances by listening to recorded audiotapes (Table 32-2).³⁰

There have been 4 studies that address the interobserver precision of cardiac auscultation to detect diastolic murmurs (Table 32-2). Although simple agreement is high in these studies, the one study for which it was possible to calculate agreement adjusted for chance (κ) showed only moderate agreement. The experience of observers likely affects precision. The one study²⁸ that compared cardiologists with noncardiologists found a higher simple agreement for cardiologists.

The interobserver agreement between examiners on the intensity of heart sounds is excellent (92%).²⁹ In this study, examiners progressively inserted 0.5-mm-thick paper disks between the patient's chest and the stethoscope. The total thickness of the disks was used as a measure of heart sound intensity. Murmur intensity was also assessed with this technique (Table 32-2).

The Bottom Line for Precision

The interobserver precision of cardiologists examining for any diastolic murmur is moderate with audiotapes ($\kappa = 0.51$) and good in the clinical setting (simple agreement, 94%). Noncardiologists may be less precise than cardiologists. The precision of examining for the intensity of murmurs and heart sounds with a standardized series of paper disks to assess intensity is good (simple agreement, 92%-96%).

Accuracy of the Examination for Aortic Regurgitation

We consider Doppler echocardiography and cardiac catheterization to be acceptable reference standards for AR (Table 32-3). In one study,³⁷ the reference standard was open-heart surgery.

Cardiologists conducted the clinical examinations in most studies. Too few studies, using few patients, allow for reasonable estimates of the accuracy of noncardiologists, although noncardiologists are likely less adept at detecting the diastolic murmur of AR. Approximately 20% of residents and medical students correctly identified the murmur of AR on high-fidelity digitized audiotapes,³¹ whereas 46% of internal medicine residents correctly identified an AR murmur on a patient simulator.³²

The best-studied physical finding is the typical early-diastolic murmur of AR.³³⁻⁴⁶ If an examiner does not hear a typical AR murmur, then the likelihood that the patient has moderate or greater AR is significantly reduced (negative likelihood ratio [LR-], 0.1 for grade A studies); the likelihood of mild or greater AR is also significantly reduced (LR-, 0.2-0.3 for grade A studies). If an examiner hears the typical AR murmur, the likelihood that the patient has moderate or greater AR is increased (positive likelihood ratio [LR+], 4.0-8.3 for grade A studies); the likelihood of mild or greater AR is also significantly increased (LR+, 8.8-32 for grade A studies).^{33,34}

The intensity of the murmur correlates with the severity of echocardiographic AR. Desjardins et al⁴⁷ studied 40 patients with echocardiographic AR, including 17 with severe AR. A grade 3 diastolic murmur had an LR of 4.5 (95% CI, 1.6-14) for distinguishing severe AR from less severe AR, whereas a grade 2 murmur had an LR of 1.1 (95% CI, 0.5-2.4), a grade 1 murmur had an LR of 0.0 (95% CI, 0.0-0.9), and absence of a diastolic murmur had an LR of 0.0 (95% CI, 0.0-1.1).⁴⁷

Two grade C studies of the Flint murmur and some peripheral hemodynamic findings are reported in Table 32-3. Grade C studies tend to overestimate diagnostic test accuracy. Despite this tendency, one study found that absence of a Flint murmur did not rule out AR (LR-, 0.5-0.8).⁴⁸ Another study of patients

Table 32-2 Interobserver Reliability (Precision) for Detecting Diastolic Murmurs

Finding	Type of Examiner	No. of Examiners	No. of Patients	κ^a	Simple Agreement, %
Murmur absent vs present	Cardiologists (tapes) ³⁰	5	100	0.51	79
	Cardiologists ²⁸	2	32	...	94
	Noncardiologists ²⁸	3	32	...	78
Intensity of murmur	Not stated ^{29,b}	5	25	...	92

^aEllipses indicate data not available.

^bExaminers used paper disks, 0.5 mm in thickness, that were progressively inserted between the chest wall and the stethoscope until the murmur became inaudible. The total thickness of the disks used was used as the measure of intensity. For example, if a murmur was inaudible after insertion of 3 disks, then this was a 1.5-mm murmur.

with mild to severe AR found only that a wide pulse pressure or peripheral hemodynamic sign (Duroziez bruit, femoral pistol shots, and Corrigan pulses) was not helpful for distinguishing mild AR from moderate or severe AR.²⁰ The de Musset head-bobbing sign was seen in only 1 of 20 patients (sensitivity, 5%), while Duroziez femoral bruit was observed in 8 of 12 patients (sensitivity, 67%),¹⁶ making them interesting but not particularly useful findings.

THE BOTTOM LINE FOR AORTIC REGURGITATION

When a cardiologist hears the typical murmur of AR, the likelihood of mild or greater AR is increased significantly (2 grade A studies). The absence of a typical diastolic murmur significantly reduces the likelihood of AR (2 grade A studies). Noncardiologists may be less proficient than cardiologists at detecting the murmur of AR.

Mitral Stenosis and Pulmonic Regurgitation

In one grade A study of 529 unselected nursing home residents (31 with mitral stenosis), a cardiologist detected a mid-diastolic murmur in all cases of mitral stenosis, with no false-positive or -negative examinations.⁴⁹ Only 1 patient had an audible OS.

Noncardiologists may be less proficient at detecting the physical findings of mitral stenosis. Less than 10% of residents and medical students correctly identified a mid-diastolic murmur of mitral stenosis on a high-fidelity digitized audiotape,³¹ whereas 43% of medical residents identified a mid-diastolic murmur of mitral stenosis with a patient simulator. In the latter study, only 21% identified the OS of mitral stenosis.³²

The only evaluated element of the clinical examination for PR is the presence of a typical diastolic decrescendo murmur best audible in the second intercostal space at the left-upper sternal border, which may increase in intensity with quiet inspiration. All studies used cardiologists as examiners and were of poor methodologic quality (grade C).

Table 32-3 Accuracy of the Physical Examination for Detecting Aortic Regurgitation

Study, y	Patient Population	Reference Standard	No. of Patients With AR	LR+ (95% CI)	LR- (95% CI)	Quality Grade ^a
Typical Murmur With Severity of AR Specified						
Aronow and Kronzon ³³ (1989)	Elderly patients	Echocardiography (n = 450)				A
		Mild or greater AR	131	32 (16-63)	0.2 (0.1-0.3)	
		Moderate or greater AR	74	8.3 (6.2-11)	0.1 (0.0-0.2)	
Grayburn et al ³⁴ (1986)	Referred for catheterization	Catheterization (n = 106)				A
		Mild or greater AR	82	8.8 (2.8-32)	0.3 (0.2-0.4)	
		Moderate or greater AR	57	4.0 (2.5-6.9)	0.1 (0.1-0.3)	
Roldan et al ³⁵ (1996)	Asymptomatic connective tissue disease and controls	Echocardiography (n = 143)				C
		Mild or greater AR	10	80 (14-470)	0.4 (0.2-0.7)	
		Moderate or greater AR	5	69 (18-270)	0.0 (0.0-0.6)	
Rahko ³⁶ (1989)	Referred for echocardiogram	Echocardiography (n = 403)				C
		Mild or greater AR	134	27 (13-60)	0.4 (0.3-0.5)	
		Moderate or greater AR	82	12 (8.1-19)	0.2 (0.1-0.3)	
Cohn et al ³⁷ (1967)	Mitral valve repair	Open-heart surgery (n = 156)				C
		Mild or greater AR	50	5.2 (3.3-8.4)	0.3 (0.2-0.4)	
		Moderate or greater AR	37	3.9 (2.6-5.7)	0.2 (0.1-0.4)	
Meyers et al ³⁸ (1982)	Referred for aortography	Catheterization (n = 75)				C
		Mild or greater AR	66	3.3 (1.3-12)	0.4 (0.2-0.7)	
		Moderate or greater AR	39	1.6 (1.2-2.4)	0.4 (0.2-0.7)	
Dittmann et al ³⁹ (1987)	Valvular heart disease	Catheterization (n = 55)				C ^b
		Mild or greater AR	42	16 (2.1-155)	0.4 (0.3-0.6)	
		Severe AR	19	3.6 (2.1-6.6)	0.1 (0.0-0.4)	
Meyers et al ⁴⁰ (1985)	Valvular heart disease	Catheterization (n = 20)				C
		Mild or greater AR	11	9.8 (1.3-96)	0.5 (0.2-0.9)	
		Moderate or greater AR	3	5.7 (1.4-14)	0.0 (0.0-0.9)	
Linhart ⁴¹ (1971)	Mitral stenosis	Catheterization (n = 28)				C
		Mild or greater AR	11	6.2 (1.9-23)	0.3 (0.1-0.7)	
		Moderate or greater AR	7	7.0 (2.5-20)	0.0 (0.0-1.3)	
Typical Murmur Without AR Severity Specified (May Include Trivial AR)						
Come et al ⁴² (1986)	Mitral valve prolapse, plus patients with systolic flow murmurs	Echocardiography (n = 165)	7	90 (8-982)	0.7 (0.4-0.9)	C
Nienaber et al ⁴³ (1993)	Clinically suspected aortic dissection	Echocardiography (n = 110)	32	33 (9.4-120)	0.2 (0.1-0.3)	C ^c
Ward et al ⁴⁴ (1977)	Clinically suspected aortic dissection	Catheterization (n = 65)	49	13 (2.9-75)	0.2 (0.1-0.3)	C ^c
Esper ⁴⁵ (1982)	AR and other heart disease	Echocardiography (n = 43)	24	12 (2.4-67)	0.4 (0.3-0.7)	C
Saal et al ⁴⁶ (1985)	Mitral stenosis	Catheterization (n = 45)	35	8.0 (1.9-45)	0.2 (0.1-0.4)	C
Maneuver						
With transient arterial occlusion murmur increases in intensity ¹⁵	Patients with AR, mitral stenosis, and pulmonic regurgitation	Catheterization or echocardiography (n = 16)	10	8.4 (1.3-81)	0.3 (0.1-0.8)	C
Associated Physical Findings						
Flint murmur ⁴⁸	Isolated AR and controls	Echocardiography (n = 36)				C
		Mild or greater AR	28	4 (0.5-40)	0.8 (0.6-1.3)	
		Moderate or greater AR	13	25 (2.8-243)	0.5 (0.2-0.7)	

(continued)

Table 32-3 Accuracy of the Physical Examination for Detecting Aortic Regurgitation (*Continued*)

Study, y	Patient Population	Reference Standard	No. of Patients With AR	LR+ (95% CI)	LR- (95% CI)	Quality Grade ^a
Any systolic murmur ⁴⁸	Isolated AR and controls	Echocardiography (n = 36)				C
		Mild or greater AR	28	1.3 (0.9-2.7)	0.5 (0.2-1.6)	
		Moderate or greater AR	13	1.5 (1.0-2.1)	0.0 (0.0-1.0)	
Popliteal-brachial gradient > 20 mm Hg ²⁰	Mild to severe AR	Catheterization (n = 33)				C
		Moderate or greater AR	28	8.2 (1.5-78)	0.2 (0.1-0.5)	
Peripheral hemodynamic signs ^{20d}	Mild to severe AR	Catheterization (n = 34)				C
		Moderate or greater AR	28	2.1 (0.3-22)	0.8 (0.7-1.7)	
Pulse pressure > 50 mm Hg ²⁰	Mild to severe AR	Catheterization (n = 33)				C
		Moderate or greater AR	28	1.0 (0.7-2.2)	0.9 (0.2-5.5)	

Abbreviations: AR, aortic regurgitation; CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

^aSee Table 1-7 for a summary of Evidence Grades and Levels.

^bGrade B study, except catheterization results were not interpreted independently of clinical findings.

^cGrade B study, except echocardiograms were not interpreted independently of clinical findings.

^dIncluded Duroziez bruit, femoral pistol shots, and Corrigan pulses.

When a cardiologist hears the murmur of PR, the likelihood of PR increases (LR+, 17 in both studies), but the absence of a PR murmur was not helpful for ruling out PR (LR, 0.9 in both studies).^{36,42}

The Bottom Line for Mitral Stenosis and Pulmonic Regurgitation

The presence of a mid-diastolic murmur significantly increases the likelihood of mitral stenosis, whereas the absence of a mid-diastolic murmur significantly reduces the likelihood of mitral stenosis (1 grade A study). When a cardiologist hears a typical PR murmur, the likelihood of PR increases significantly. The absence of a typical murmur does not alter the likelihood of PR (2 grade C studies). Noncardiologists may be less proficient at detecting the mid-diastolic murmur of mitral stenosis.

Diastolic Murmurs in Patients With Renal Failure

Diastolic murmurs caused by abnormal flow states, rather than abnormal cardiac structure, may be associated with a variety of conditions. Renal failure with volume overload is the only abnormal flow state associated with diastolic murmurs that has been evaluated.

Up to 9% of patients with end-stage renal disease have diastolic murmurs, particularly when these patients also have volume overload, anemia, and hypertension.⁵⁰ These murmurs typically disappear after the treatment of volume overload, as was demonstrated in 2 small studies (grade C).^{50,51} These murmurs are probably due to transient pulmonary hypertension and dilatation of the pulmonary artery root, leading to PR.⁵¹

THE BOTTOM LINE FOR DIASTOLIC MURMURS IN PATIENTS WITH RENAL FAILURE

Although there is an insufficient amount of data on which to make rigorous recommendations, if an early-diastolic mur-

mur is heard in a dialysis patient with volume overload, the patient should be reexamined after treatment because the murmur may disappear.

When to Examine for Aortic Regurgitation

There are no evaluative data on which to base a recommendation regarding when to examine for AR. Undetected AR may be common in elderly persons: 13% (n = 552) of asymptomatic elderly Finnish persons had moderate or severe echocardiographic AR.⁵² Unfortunately, that study does not indicate how many of these patients had audible diastolic murmurs. Audible diastolic murmurs may be relatively uncommon findings in asymptomatic persons. In one study, only 1% (n = 103) of elderly asymptomatic nursing home residents had an audible diastolic murmur.⁵³

Despite the lack of evaluative data, we think that a prudent clinician will examine for AR in most clinical settings. AR is a serious cardiac abnormality, which may be caused by underlying disorders and may be asymptomatic. The clinician's suspicion for AR may be heightened by evidence of systemic disease, such as ankylosing spondylitis, a peripheral hemodynamic finding (although these are by no means indicative of underlying AR), or an abnormality detected during routine auscultation (such as an aortic ejection sound). Other findings may suggest different cardiac abnormalities associated with diastolic murmurs, such as evidence of pulmonary hypertension (for PR), a wide-fixed split S₂ (for atrial-septal defect), or a holosystolic apical murmur (for mitral regurgitation).

Recommendations for Further Research

Most studies used cardiologists to conduct clinical examinations. There are some data that suggest that noncardiologists may be less accurate than cardiologists, so studies evaluating techniques to improve the skills of noncardiologists are needed. There are also no studies defining the optimal examination technique for detecting the AR murmur.

CLINICAL SCENARIO—RESOLUTION

Your patient has a wide pulse pressure but no typical early-diastolic murmur. The likelihood of mild or moderate AR is significantly reduced by the absence of a typical early-diastolic murmur (LR₋, 0.1-0.3; 2 grade A studies). You perform transient arterial occlusion, and no diastolic murmur appears, which enhances your confidence (LR₋, 0.3). You are confident in your assessment because it was conducted in a quiet room with a comfortable and cooperative patient. Therefore, AR is unlikely and echocardiography is not necessary.

Author Affiliations at the Time of the Original Publication

Division of General Internal Medicine and Clinical Epidemiology, Department of Medicine, University of Toronto and the University Health Network, Toronto, Ontario, Canada.

Acknowledgment

This work was funded in part by grant 98-01R from Physicians Services Incorporated Foundation.

The views expressed in this article are those of the authors and do not reflect those of any sponsoring or supporting agency.

We thank Eugene Oddone, MD, MHS, for his helpful comments on earlier drafts of this article, and Wilbert S. Aronow, MD, and A. Nienaber, MD, for providing additional data and methodologic information about their studies.

REFERENCES

- Scognamiglio R, Rahimtoola SH, Fasoli G, Nistri S, Volta SD. Nifedipine in asymptomatic patients with severe aortic regurgitation and normal left ventricular function. *N Engl J Med.* 1994;331(11):689-694.
- Ballard DJ, Khandheira BK, Tajik AJ, Seward JB, Weber VP, Melton LJ. Population-based study of echocardiography. *Int J Technol Assess Health Care.* 1989;5(2):249-261.
- Etchells E, Bell C, Robb K. Does this patient have an abnormal systolic murmur? *JAMA.* 1997;277(7):564-571.
- Rushmer RF, Sparkman DR, Polley RFL, et al. Variability in detection and interpretation of heart murmurs. *Am J Dis Child.* 1952;83(6):740-754.
- Feinstein AR, Di Massa R. The unheard diastolic murmur in acute rheumatic fever. *N Engl J Med.* 1959;260(26):1331-1333.
- Freeman AR, Levine SA. The clinical significance of the systolic murmur: a study of 1000 consecutive cases. *Ann Intern Med.* 1933;6:1371-1385.
- Constant J, Lippischutz EJ. Diagramming and grading heart sounds and murmurs. *Am Heart J.* 1965;70:322-326.
- Constant J. *Essentials of Bedside Cardiology.* 2nd ed. Totowa, NJ: Human Press; 2003:243-263.
- Harvey WP. Cardiac pearls. *Dis Mon.* 1994;40(2):43-113.
- Shaver JA. Cardiac auscultation. *Curr Probl Cardiol.* 1995;20(7):445-530.
- Landzberg JS, Pflugfelder PW, Cassidy MM, Schiller NB, Higgins CB, Cheitlin MD. Etiology of the Austin Flint murmur. *J Am Coll Cardiol.* 1992;20:408-413.
- Flint A. On cardiac murmurs. *Am J Med Sci.* 1862;44:29-54.
- Sapira JD. *The Art and Science of Bedside Diagnosis.* Baltimore, MD: Urban & Schwarzenberg; 1990:283-330.
- Soffer A, Feinstein A, Rosener S, et al. Glossary of cardiologic terms related to physical diagnosis and history. *Am J Cardiol.* 1967;20:285-286.
- Lembo NJ, Dell'Italia LJ, Crawford MH, O'Rourke RA. Diagnosis of left-sided regurgitant murmurs by transient arterial occlusion: a new maneuver using blood pressure cuffs. *Ann Intern Med.* 1986;105(3):368-370.
- Sapira JD. Quincke, de Musset, Duroziez, and Hill: some aortic regurgitations. *South Med J.* 1981;74(4):459-467.
- Hegglin R, Scheu H, Rothlin M. Aortic insufficiency. *Circulation.* 1968;37(suppl V):V77-V92.
- Hill L. The measurement of systolic blood pressure in man. *Heart.* 1909;1:73-82.
- Warnes CA, Harris PC, Fritts HW. Effect of elevating the wrist on the radial pulse in aortic regurgitation. *Am J Cardiol.* 1983;51(9):1551-1553.
- Frank MJ, Casanegra P, Migliori AJ, Levinson GE. The clinical evaluation of aortic regurgitation. *Arch Intern Med.* 1965;116:357-365.
- Reeves RA. Does this patient have hypertension? *JAMA.* 1995;273(15):1211-1218.
- Blumgart HL, Ernstene AC. Two mechanisms in the production of Duroziez's sign. *JAMA.* 1933;100:173-177.
- Abbas F, Sapira JD. Mayne's sign is not pathognomonic of aortic insufficiency. *South Med J.* 1987;80(8):1051-1052.
- L'Abbe KA, Detksky AS, O'Rourke K. Meta-analysis in clinical research. *Ann Intern Med.* 1987;107(2):224-233.
- Holleman DR, Simel DL. Does the clinical examination predict airflow limitation? *JAMA.* 1995;273(4):313-319.
- Koopman PAR. Confidence intervals for the ratio of two binomial proportions. *Biometrics.* 1984;40:513-517.
- Centor R, Keightley J. *The Two by Two Analyzer.* Birmingham, AL: CenterSoft Statistical Programs for Medical Decision Making Research; 1992.
- Raftery EB, Holland WW. Examination of the heart. *Am J Epidemiol.* 1967;85(3):438-444.
- Sharma RS. Measurement of intensity of heart sounds and murmurs. *Lancet.* 1981;2(8247):612.
- Dubrow RJ, Calatayud JB, Abraham S, Caceres CA. A study of physician variation in heart-sound interpretation. *Med Ann Dist Columbia.* 1964;33:305-308, 355.
- Mangione S, Nieman LZ. Cardiac auscultatory skills of internal medicine and family practice trainees. *JAMA.* 1997;278(9):717-722.
- St Clair EW, Oddone EZ, Waugh RA, Corey GR, Feussner JR. Assessing housestaff diagnostic skills using a cardiology patient simulator. *Ann Intern Med.* 1992;117(9):751-756.
- Aronow WS, Kronzon I. Correlation of prevalence and severity of aortic regurgitation detected by pulsed Doppler echocardiography with the murmur of aortic regurgitation in elderly patients in a long-term health care facility. *Am J Cardiol.* 1989;63:128-129.
- Grayburn PA, Smith MD, Handshoe R, Friedman BJ, DeMarie AN. Detection of aortic insufficiency by standard echocardiography, pulsed Doppler echocardiography, and auscultation. *Ann Intern Med.* 1986;104:599-605.
- Roldan CA, Shively BK, Crawford MH. Value of the cardiovascular physical examination for detecting valvular heart disease in asymptomatic subjects. *Am J Cardiol.* 1996;77:1327-1331.
- Rahko PS. Prevalence of regurgitant murmurs in patients with valvular regurgitation detected by Doppler echocardiography. *Ann Intern Med.* 1989;111:466-472.
- Cohn LH, Mason DT, Ross J, Morrow AG, Braunwald E. Preoperative assessment of aortic regurgitation in patients with mitral valve disease. *Am J Cardiol.* 1967;19:177-182.
- Meyers DG, Sagar KB, Ingram RF, Paulsen WJH, Romhilt DW. Diagnosis of aortic insufficiency. *South Med J.* 1982;75:1192-1194.
- Dittmann H, Karsch KR, Seipel L. Diagnosis and quantification of aortic regurgitation by pulsed Doppler echocardiography in patients with mitral valve disease. *Eur Heart J.* 1987;8(suppl C):53-57.
- Meyers DG, Olson TS, Hansen DA. Auscultation, M-mode echocardiography and pulsed Doppler echocardiography compared with angiography for diagnosis of chronic aortic regurgitation. *Am J Cardiol.* 1985;56:811-812.
- Linhart JW. Aortic regurgitation. *Ann Thorac Surg.* 1971;11:27-37.
- Come PC, Riley MF, Carl JV, Nakao S. Pulsed Doppler echocardiographic evaluation of valvular regurgitation in patients with mitral valve prolapse. *J Am Coll Cardiol.* 1986;8:1355-1364.
- Nienaber CA, von Kodolitsch Y, Nicolas V, et al. The diagnosis of thoracic aortic dissection by noninvasive imaging procedures. *N Engl J Med.* 1993;328:1-9.
- Ward JM, Baker CW, Rubenstein SA, Johnston SL. Detection of aortic insufficiency by pulse Doppler echocardiography. *J Clin Ultrasound.* 1977;5:5-10.

45. Esper RJ. Detection of mild aortic regurgitation by range-gated pulsed Doppler echocardiography. *Am J Cardiol.* 1982;50:1037-1043.
46. Saal AK, Gross BW, Franklin DW, Pearlman AS. Noninvasive detection of aortic insufficiency in patients with mitral stenosis by pulsed Doppler echocardiography. *J Am Coll Cardiol.* 1985;5:176-181.
47. Desjardins VA, Enriquez-Sarano M, Tajiik AJ, Bailey KR, Seward JB. Intensity of murmurs correlates with severity of valvular regurgitation. *Am J Med.* 1996;100(2):149-156.
48. Lee D, Chen CH, Hsu TL, Chiang CE, Wang SP, Chang MS. Reappraisal of cardiac murmurs related to aortic regurgitation. *Chin Med J (Engl).* 1995;56:152.
49. Aronow WS, Schwartz KS, Koenigsberg M. Correlation of murmurs of mitral stenosis and mitral regurgitation with presence or absence of mitral annular calcium in person older than 62 years in a long-term health care facility. *Am J Cardiol.* 1987;59(1):181-182.
50. Alexander WD, Polak A. Early diastolic murmurs in end-stage renal failure. *Br Heart J.* 1977;39(8):900.
51. Perez JE, Smith CA, Meltzer VN. Pulmonic valve insufficiency. *Ann Intern Med.* 1985;103(4):497-502.
52. Lindroos M, Kupari M, Heikkila J, Tilvis R. Prevalence of aortic valve abnormalities in the elderly. *J Am Coll Cardiol.* 1993;21(5):1220-1225.
53. Bethel CS. Heart sounds in the aged. *Am J Cardiol.* 1963;11:763-767.

This page intentionally left blank

Prepared by David Cescon, MD, and Edward Etchells, MD

Reviewed by Eugene Oddone, MD

CLINICAL SCENARIO

A 58-year-old man presents for a routine physical examination, not having visited a physician in many years. He denies any cardiovascular symptoms. On auscultation, you are surprised to hear a loud (grade 3) early diastolic murmur. There is an audible S_3 and a collapsing radial pulse (Corrigan sign). After explaining to the patient that you heard a murmur, he asks, "How bad do you think it is?"

ORIGINAL REVIEW

Choudhry NK, Etchells EE. Does this patient have aortic regurgitation? *JAMA*. 1999;281(23):2231-2238.

UPDATED LITERATURE SEARCH

Our literature search combined the parent search strategy for The Rational Clinical Examination with the terms "diastolic and murmur, aortic valve insufficiency, mitral valve stenosis," and "pulmonary valve insufficiency," limited to English-language publications in the Ovid MEDLINE database from 1997 to July 2004. The titles and abstracts of the search results were screened, case reports were excluded, and 8 potentially relevant articles were retrieved and reviewed. We manually reviewed the reference list of each article for additional studies. Articles were retained if they were studies of adult participants, included sensitivity and specificity data of physical findings, and had a quality score of level 3 or greater. This yielded 1 new study, and we found 1 other study during the updated literature search for systolic murmurs.

NEW FINDINGS

The presence of an S_3 in patients with isolated aortic regurgitation (AR) predicts severity.

Details of the Update

Were There Changes in the Original Publication?

In the original article, the need to identify patients at higher risk for endocarditis because of valvular abnormalities was suggested as a rationale for performing the clinical examination. The recommendations for endocarditis prophylaxis

have changed. Patients with murmurs from structural abnormalities of a native valve do not automatically require antibiotic prophylaxis to prevent infective endocarditis.¹

CHANGES IN THE REFERENCE STANDARD

The reference standard is the echocardiogram or the results from a cardiac catheterization that assess valvular competency.

RESULTS OF THE LITERATURE REVIEW

Diastolic murmurs are always important, requiring ascertainment of the underlying abnormality. Most studies of the detection of AR assess the performance of cardiologists or the ability to distinguish patients with serious AR from those with less significant impairment. The sensitivity and specificity of a variety of peripheral hemodynamic findings, popularized in many textbooks of physical diagnosis and cardiology, have not been adequately assessed.

Since the original review, 1 study² assessed the ability of cardiologists to identify the presence of AR (Table 32-4). In this study, 100 consecutive patients referred for evaluation of a systolic murmur of unknown cause underwent a standard cardiac examination by a cardiologist, who described the murmur and assigned a clinical diagnosis. Mild or greater AR was identified with high specificity but low sensitivity. Compared with studies cited in the original review, the lower sensitivity might reflect a challenging sample with a high prevalence of multiple valvular lesions and a predominance of mild regurgitation among those patients with AR.

Table 32-4 Likelihood Ratios of the Physical Examination for Detecting Aortic Regurgitation

	Patient Population	LR+ (95% CI)	LR- (95% CI)
Overall cardiac examination ²	Referred for evaluation of systolic murmur	5.1 (1.4-19)	0.82 (0.67-1.0)
Third heart sound (to identify severe AR) ³	Patients with isolated aortic insufficiency, referred for echocardiography	5.9 (1.4-25)	0.83 (0.73-0.95)

Abbreviations: AR, aortic regurgitation; CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

One additional study³ evaluated the ability of the clinical examination to distinguish severe AR from less severe disease. The presence of an S_3 , recorded by a physician referring a patient for cardiac ultrasonography, predicted severe AR (defined as a regurgitant fraction $\geq 40\%$), with a likelihood ratio of 5.9. The absence of an S_3 was not useful for ruling out severe AR (negative likelihood ratio, 0.83) (Table 32-4).³

EVIDENCE FROM GUIDELINES

The American College of Cardiology and American Heart Association guidelines⁴ (2003) state that Doppler echocardiography to exclude valvular regurgitation in asymptomatic patients with normal physical examination results is not indicated.

CLINICAL SCENARIO—RESOLUTION

Your patient has a typical diastolic murmur of AR. This finding alone warrants echocardiography because it is highly suggestive of an underlying abnormality. If the patient has AR, the grade 3 intensity of the murmur and the third heart sound increase the likelihood of severe AR. The collapsing radial pulse (Corrigan pulse) is of uncertain usefulness. Putting together all of your findings, you advise your patient that you are concerned that he might have important valvular disease. You provide information regarding endocarditis prophylaxis and arrange for an echocardiogram.

AORTIC REGURGITATION—MAKE THE DIAGNOSIS

PRIOR PROBABILITY

One study of randomly selected elderly (75–86 years old) Finnish persons found a 29% prevalence of mild or greater AR.⁵ Evaluation of more than 3000 men and women (aged 54 ± 10 years) in the Framingham heart study detected AR of trace or greater severity in 13.0% of men and 8.5% of women.⁶ Increasing age was associated with higher prevalence of AR.

POPULATION FOR WHOM THE SIGNS SHOULD BE EVALUATED

- Any patient undergoing cardiac auscultation
- A variety of medical and traumatic conditions are associated with AR:
- Rheumatic fever
 - Endocarditis
 - Conditions associated with aortic valve leaflet abnormalities (eg, Marfan syndrome, rheumatoid arthritis, ankylosing spondylitis)
 - Diseases that affect the aortic root (eg, hypertension, syphilis, inherited connective tissue disorders, aortic aneurysm)

PHYSICAL EXAMINATION SIGNS USEFUL IN THE DIAGNOSIS OF AORTIC REGURGITATION

The presence of a typical murmur of AR (an early diastolic, decrescendo murmur) should prompt echocardiographic evaluation (Table 32-5). Many eponymic peripheral pulse findings are associated with AR, but they are not useful for screening or for distinguishing the severity of regurgitation.

Table 32-5 Likelihood Ratio for Typical Murmur to Predict Aortic Regurgitation or an S_3 to Predict Severe Aortic Regurgitation

Finding (Type of Clinician)	Severity by Echocardiogram or Cardiac Catheterization	LR+ (Range or Point Estimate With 95% CI)	LR- (Range or Point Estimate With 95% CI)
Typical murmur ^{7,8} (cardiologist)	Mild or greater	8.8-32	0.2-0.3
	Moderate or greater	4.0-8.3	0-0.1
Murmur intensity ⁹ (generalist or cardiologist) ^a	Grade 3	4.5 (1.6-14)	
	Grade 2	1.1 (0.5-2.4)	
	Grade 1	0 (0-0.9)	
	No murmur	0 (0-1.1)	
Third heart sound ⁴ (cardiologist)	Severe	5.9 (1.4-25)	0.83 (0.73-0.95)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

^aAll patients had aortic regurgitation, so the likelihood ratios here are for severe aortic regurgitation associated with the murmur intensity.

REFERENCE STANDARD TESTS

Echocardiography and angiography.

REFERENCES FOR THE UPDATE

- Wilson W, Taubert KA, Gewitz M, et al. Prevention of infective endocarditis: guidelines from the American Heart Association: a guideline from the American Heart Association Rheumatic Fever, Endocarditis, and Kawasaki Disease Committee, Council on Cardiovascular Disease in the Young, and the Council on Clinical Cardiology, Council on Cardiovascular Surgery and Anesthesia, and the Quality of Care and Outcomes Research Interdisciplinary Working Group. *Circulation*. 2007;116(15):1736-1754.
- Attenhofer Jost CH, Turina J, Mayer K, et al. Echocardiography in the evaluation of systolic murmurs of unknown cause. *Am J Med*. 2000;108(8):614-620.^a
- Tribouilloy CM, Enriquez-Sarano M, Mohty D, et al. Pathophysiologic determinants of third heart sounds: a prospective clinical and Doppler echocardiographic study. *Am J Med*. 2001;111(2):96-102.^a

4. Cheitlin MD, Armstrong WF, Aurigemma GP, et al. ACC/AHA/ASE 2003 guideline update for the clinical application of echocardiography: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. http://www.acc.org/qualityandscience/clinical/guidelines/echo/index_clean.pdf. Accessed June 4, 2008.
5. Lindroos M, Kupari M, Heikkila J, Tilvis R. Prevalence of aortic valve abnormalities in the elderly: an echocardiographic study of a random population sample. *J Am Coll Cardiol.* 1993;21(5):1220-1225.
6. Singh JP, Evans JC, Levy D, et al. Prevalence and clinical determinants of mitral, tricuspid, and aortic regurgitation (the Framingham study). *Am J Cardiol.* 1999;83(6):897-902.
7. Aronow WS, Kronzon I. Correlation of prevalence and severity of aortic regurgitation detected by pulsed Doppler echocardiography with a murmur of aortic regurgitation in elderly patients in a long-term health care facility. *Am J Cardiol.* 1989;63(1):128-129.
8. Grayburn PA, Smith MD, Handshoe R, Friedman BJ, DeMarie AN. Detection of aortic insufficiency by standard echocardiography, pulsed Doppler echocardiography, and auscultation. *Ann Intern Med.* 1986;104(5):599-605.
9. Desjardins VA, Enriquez-Sarano M, Tajik AJ, Bailey KR, Seward JB. Intensity of murmurs correlates with severity of valvular regurgitation. *Am J Med.* 1996;100(2):149-156.

^aFor the Evidence to Support the Update for this topic,
see <http://www.JAMAevidence.com>.

This page intentionally left blank

EVIDENCE TO SUPPORT THE UPDATE:

32

Murmur, Diastolic

TITLE Echocardiography in Evaluating Systolic Murmurs of Unknown Cause.

AUTHORS Attenhofer Jost CH, Turina J, Mayer K, et al.

CITATION Am J Med. 2000;108(8):614-620.

QUESTION How well can cardiologists identify pathologic murmurs by auscultation and palpation alone?

DESIGN Consecutive patients were prospectively identified at referral for evaluation of a systolic murmur of unknown cause. Each participant was independently examined by 2 cardiologists from a pool of 8 and blinded to supplementary data and echocardiography results. Two-dimensional/Doppler echocardiography was performed as the gold standard in all participants.

SETTING Cardiology division in Switzerland.

PATIENTS One hundred patients referred for evaluation of systolic murmur of unknown cause were enrolled. Patients were excluded if they had a previously documented echocardiographic examination. The mean age of the participants was 55 years, with SD 22, and 57% were women.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

Full cardiac examination, with or without dynamic auscultation, was performed by 1 staff cardiologist and 1 cardiology associate. Only the results of the staff cardiologist's examinations were used in the analysis, and no comparison with the associate's findings is presented in the article. Murmurs were classified by Levine grade according to predefined characteristics, and the murmurs were classified as functional or organic according to the examiner's clinical expertise. All patients underwent transthoracic 2-dimensional and Doppler echocardiography, and valvular stenosis and regurgitation were classified according to standard criteria.

MAIN OUTCOME MEASURES

Descriptive statistics, sensitivity, specificity.

Table 32-6 Likelihood Ratio for Aortic Regurgitation According to the Presence of a Diastolic Murmur in Patients Referred for Systolic Murmurs

	AR by Echocardiography			
	Sensitivity	Specificity	LR+ (95% CI)	LR- (95% CI)
AR by clinical examination	0.21	0.96	5.1 (1.4-19)	0.82 (0.67-1.0)

Abbreviations: AR, aortic regurgitation; CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

Main Results

Twenty-eight of the patients referred for systolic murmurs had aortic regurgitation (AR). The degree of regurgitation was mild in 22 cases (79%) and associated with another echocardiographic lesion in 15 (54%) cases. The examiners made a clinical diagnosis of aortic insufficiency in 9 patients (Table 32-6).

CONCLUSION

LEVEL OF EVIDENCE Level 3.

STRENGTHS Prospective, consecutive patients.

LIMITATIONS Small, referral population referred for evaluation of a murmur. The echocardiographers were not blinded to the clinical findings.

The clinical examination was useful for ruling in AR but not for ruling out regurgitation. The negative likelihood ratio obtained in this study is higher than in a number of previous studies performed in a variety of settings. The difficult population in this study might explain this finding: patients were referred for evaluation of *systolic* murmurs, and those with AR had predominantly mild disease and approximately half had additional lesions. The ability of the cardiologist to identify those with AR (likelihood ratio, 5.1) despite the referral indication of a systolic murmur is impressive. These data support the clinical suggestion that finding a diastolic murmur requires an echocardiogram to assess AR.

Reviewed by David Cescon, MD,
and Edward Etchells, MD

TITLE Pathophysiologic Determinants of Third Heart Sounds: A Prospective Clinical and Doppler Echocardiographic Study.

AUTHORS Tribouilloy CM, Enriquez-Sarano M, Mohty D, et al.

CITATION *Am J Med.* 2001;111(2):96-102.

QUESTION Does an audible S₃ predict severe hemodynamic alterations in cardiology patients?

DESIGN Patients were identified at referral for echocardiography. Clinical data were obtained by noting the results of a clinical examination performed by the referring physician (a cardiologist or internist), who was unaware of the study. Transthoracic echocardiography was performed on all patients.

SETTING Cardiology referral center in the United States.

PATIENTS One hundred twenty-one patients with aortic regurgitation (AR) (mean age, 57 years; SD, 18 years; 66% men; 15% New York Heart Association classes III-IV) were included in the study. These patients were prospectively enrolled from among patients referred by their personal physician for echocardiography for any indication and found to have isolated mitral regurgitation or AR. Exclusion criteria included previous valve surgery, associated valvular stenosis, acute myocardial infarction, congenital or pericardial disease, or a change in cardiovascular status since clinical examination was performed.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

Documentation of the presence or absence of an S₃ was abstracted from each patient's chart as documented by the referring physician. Two-dimensional/Doppler echocardiography was performed on all patients by an echocardiographer. Severe regurgitation was defined as a regurgitant fraction of 40% or greater.

MAIN OUTCOME MEASURES

Sensitivity, specificity, and likelihood ratio for the ability of an S₃ to identify patients with severe regurgitation.

MAIN RESULTS

Fourteen patients with AR had an S₃. Of the 121 patients with AR, 61 were classified as severe according to the echocardiogram (Table 32-7).

Table 32-7 Likelihood Ratio for the Presence of a Third Heart Sound to Predict Severe Aortic Regurgitation (AR) as Opposed to Less Severe AR

Test	Sensitivity	Specificity	LR+ (95% CI)	LR- (95% CI)
S ₃ to identify severe AR	0.20	0.97	5.9 (1.4-25)	0.83 (0.73-0.95)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

CONCLUSION

LEVEL OF EVIDENCE Level 3.

STRENGTHS Large sample size.

LIMITATIONS The data were collected retrospectively, and it is not clear whether the echocardiographer was blinded to the clinical examination. The patients with an S₃ may have been selectively referred for echocardiograms, but this would have led to an inflated sensitivity and underestimated specificity, which is the opposite of what the investigators found.

The presence of an S₃ is highly specific for severe regurgitation in patients with isolated valvular regurgitation, and its presence reflects hemodynamically significant regurgitation reflected by left ventricular dysfunction. The detection of an S₃ should prompt further evaluation. However, the absence of an S₃ is not useful in ruling out significant AR.

Reviewed by David Cescon, MD,
and Edward Etchells, MD

CHAPTER 33

Does This Patient Have an Abnormal Systolic Murmur?

Edward Etchells, MD, MSC

Chaim Bell, MD

Kenneth Robb, MD

CLINICAL SCENARIOS

CASE 1 You are asked to see a 64-year-old man who has been admitted to the orthopedic service after a packing crate tipped over on his leg, producing an unstable fracture of his distal tibia and fibula. You see him as he is being prepared for surgery. The patient previously had a normal exercise tolerance and no cardiac symptoms. You conduct a complete cardiac examination, observing a grade 2 systolic murmur, loudest at the lower left sternal border, which does not radiate to the right carotid artery. The S_2 has normal intensity, and you do not hear a fourth heart sound (S_4). The carotid artery pulsation has a normal rate of increase and normal volume. The orthopedic surgeon is concerned about the murmur because a recent patient had a postoperative myocardial infarction (MI) and was subsequently diagnosed with aortic stenosis. The surgeon wonders whether surgery should be delayed until an echocardiogram is obtained to rule out aortic stenosis.

CASE 2 Your next patient is a 34-year-old woman without cardiovascular symptoms who has normal exercise tolerance. She has a grade 2 systolic murmur that begins late in systole and is loudest at the lower left sternal border. When the patient is examined in a standing position, the murmur increases in intensity, and you detect a loud systolic click just before the onset of the murmur. The rest of the cardiovascular examination result is normal. You suspect mitral valve prolapse (MVP), but you wonder how confident you should feel about the diagnosis.

WHY IS THE CLINICAL EXAMINATION IMPORTANT IN EVALUATING SYSTOLIC MURMURS?

Systolic murmurs can be an important clue to a structural cardiac abnormality (Table 33-1). The use of noninvasive cardiac testing, such as echocardiography, has increased dramatically. It is estimated that 2% of the general population undergoes noninvasive cardiac diagnostic evaluation.¹ In lieu of performing routine echocardiography on patients with systolic murmurs, a careful clinical examination may eliminate the need for additional tests in selected patients.

THE ANATOMIC AND PHYSIOLOGIC ORIGINS OF SYSTOLIC MURMURS

Heart murmurs are produced when turbulent blood flow causes prolonged auditory vibrations of cardiac structures. The intensity of the murmur depends on many factors, including blood viscosity and blood flow velocity and turbulence. In addition, the distance between the vibrations and the stethoscope, the angle at which the vibrations meet the

Table 33-1 Selected Causes of Systolic Murmurs

Abnormal cardiac structure
Aortic stenosis
Hypertrophic cardiomyopathy
Mitral regurgitation
Mitral valve prolapse
Ventricular septal defect
Pulmonic stenosis
Tricuspid regurgitation
Atrial septal defect
Normal cardiac structure, increased flow
Anemia
Thyrotoxicosis
Sepsis
Renal failure with volume overload

stethoscope, and the transmission qualities of the tissue between the vibration and the stethoscope affect murmur intensity.²

In this article, we will arbitrarily define an abnormal systolic murmur as one associated with abnormal cardiac structure. We will not consider the diagnosis of systolic murmurs caused by abnormally increased blood flow across normal cardiac structures, such as in anemia or thyrotoxicosis. However, clinicians must consider the diagnosis of abnormally increased blood flow in patients with systolic murmurs.

HOW TO EXAMINE FOR SYSTOLIC MURMURS

Most clinicians agree that a complete clinical history and physical examination, including a detailed cardiac examination, is an essential step in the assessment of systolic murmurs. Clinicians will interpret a systolic murmur in an asymptomatic 24-year-old woman with iron deficiency anemia differently from a systolic murmur in a 76-year-old woman with fever, weight loss, and digital infarctions after recent dental surgery.

Although a complete cardiac examination is important, the reliability and accuracy of many components of the cardiac examination for systolic murmurs have not been adequately evaluated. For example, the only adequately evaluated individual element of the cardiac history related to murmurs is effort syncope, which refers to a transient loss of consciousness during effort or exertion. This article focuses on features of the cardiac physical examination for systolic murmurs that have been adequately evaluated for precision and accuracy. A complete description of the cardiac physical examination of systolic murmurs is beyond the scope of this article but can be found in many textbooks.

The cardiac physical examination includes nonauscultatory and auscultatory components. Adequately evaluated nonauscultatory components include carotid artery palpation, apical-carotid delay, and brachioradial delay. To assess the carotid pulse, the clinician applies both light and firm

pressure over the artery and assesses both the rate of increase and the pulse volume. Experts suggest that examiners pay special attention to the peak of pulsation. A normal rate of increase feels like a sharp tap, whereas an abnormal rate of increase feels like a nudge. An abnormal rate of increase can also feel like a weak tap, followed by a nudge or push.³ Surprisingly, no clear guidelines exist for interpreting carotid volume. Suggested methods include palpating the artery with both hands and all fingers, or palpating with the thumb only.⁴ We can only offer that a normal carotid volume is easily felt with light palpation, whereas a reduced carotid volume is difficult to feel even with firm palpation.

Brachioradial delay and apical-carotid delay may be important findings for detecting aortic stenosis. For brachioradial delay, the examiner palpates simultaneously the right brachial artery of the patient with the right thumb and the right radial artery of the patient with the left index and middle finger. The examiner should use only light pressure on the brachial artery to avoid dampening the pulse waveform. The examiner attempts to detect a delay between the brachial artery and the radial artery pulsations; any palpable delay is considered abnormal.⁵ For apical-carotid delay, the examiner simultaneously palpates the precordial apex pulsation and the right carotid artery. The examiner attempts to detect a delay between the apical and the carotid artery pulsation; any palpable delay is abnormal.⁶

In contrast to the cardiac history and nonauscultatory examination, many components of routine cardiac auscultation have been adequately evaluated. During routine auscultation, the examiner attempts to detect a systolic murmur, which can be defined as a systolic noise with a duration longer than a heart sound.⁷ Examiners describe the grade, radiation (Table 33-2), onset, duration, and timing of peak murmur intensity (Figure 33-1). The Levine grading system⁸ facilitates description of intensity: a grade 1 murmur is not heard immediately on auscultation but only after the examiner has focused on systole for a few seconds, a grade 2 murmur is heard immediately on auscultation but is not loud, a grade 3 murmur is heard immediately on auscultation and is loud, and a palpable precordial vibration, called a thrill, signifies a grade 4 murmur. Other murmur characteristics, such as pitch and tonal quality, have not been adequately evaluated.

Other evaluated relevant features on routine auscultation include the intensity of the S₂, the S₄, and systolic clicks. The intensity of S₂ can be graded as normal, decreased, or absent. A normal S₂ should be easily heard in the second right and left intercostal spaces next to the sternum and should be louder than the first heart sound (S₁) in these areas.³ Abnormal splitting of the S₂ in relation to cardiac murmurs has not been adequately evaluated.

An S₄ is a low-pitched sound occurring just before systole, sometimes described as a presystolic sound. The S₄ from the left ventricle is best heard with the bell of the stethoscope lightly applied to the patient in the left lateral decubitus position.

Systolic clicks are high-pitched sounds with a duration similar to that of heart sounds. Systolic clicks (previously termed nonejection clicks) are associated with MVP. They generally occur later than 40 to 60 ms after the S₁, and

patient position greatly affects their timing. When a patient stands, a systolic click moves closer to the S_1 . Ejection sounds (previously termed *ejection clicks*) come from aortic or pulmonary valves opening in early systole, approximately 40 to 60 ms after the S_1 . The S_1 and an ejection sound together have roughly the cadence of saying “pa-da” or “pa-ta” quickly.³ Patient position causes no appreciable change in the timing of ejection sounds.

After routine auscultation, the clinician may wish to further assess a systolic murmur using special maneuvers. If the maneuver is intended to increase the intensity of the murmur, then the clinician should listen at the edge of the murmur’s radiation, where the murmur is barely audible. This will make it easier to detect an increase in murmur intensity. Similarly, if the maneuver is intended to decrease the intensity of the murmur, then the clinician should listen at the point of maximal intensity.

Maneuvers that primarily increase the venous return include quiet inspiration and sustained abdominal pressure. These maneuvers are intended to increase the intensity of right-sided heart murmurs, such as tricuspid regurgitation (TR) or pulmonic stenosis. For the quiet inspiration maneuver, the examiner determines the effect of quiet inspiration on the intensity of the murmur.⁹ The examiner should not ask the patient to breathe deeply, because the murmur will be obscured by the breath sounds. For the sustained abdominal pressure maneuver, the examiner exerts firm, sustained pressure inward and cephalad below the right costal margin. The intensity of the murmur is observed during several cardiac cycles.¹⁰

Transient arterial occlusion primarily increases systemic arterial resistance. This maneuver increases the intensity of left-sided regurgitant murmurs, such as mitral regurgitation (MR) or ventricular septal defect. The examiner inflates simultaneously 2 sphygmomanometers placed around each of the patient’s upper arms to approximately 20 to 40 mm above the previously recorded systolic blood pressure of the patient. Twenty seconds after cuff inflation, any changes in murmur intensity are observed.¹¹

Maneuvers that increase both venous return and systemic arterial resistance include standing to squatting and passive leg elevation. These maneuvers are intended to decrease the intensity of the murmur of hypertrophic cardiomyopathy and MVP. For the standing to squatting maneuver, the clinician sits to the side of the patient and instructs him or her to rapidly squat from the standing position. Changes in murmur intensity are noted immediately after squatting.¹² For the passive leg elevation maneuver, an assistant passively elevates both of the patient’s legs to approximately 45 degrees while the patient is supine. Changes in murmur intensity are observed 15 to 20 seconds after leg elevation.¹³

The Valsalva maneuver decreases venous return and increases systemic arterial resistance.³ The Valsalva maneuver decreases the intensity of aortic stenosis murmurs. The patient strains against a closed glottis for 20 seconds, and changes in murmur intensity are observed just before the end of the 20-second period.¹³ Patients may inadvertently do a

Table 33-2 Typical Location of Maximal Intensity and Radiation for Various Types of Abnormal Systolic Murmurs

Location of Maximal Intensity	Radiation	Typical for
Second right intercostal space	Right carotid artery Right clavicle	Aortic stenosis
Fifth or sixth left intercostal space mid left thorax	Left anterior axillary line Left axilla	Mitral regurgitation (including mitral regurgitation caused by mitral valve prolapse)
Lower left sternal border	Lower right sternal border Epigastrium	Tricuspid regurgitation
Fifth left intercostal space mid left thorax	Lower left sternal border	Hypertrophic cardiomyopathy

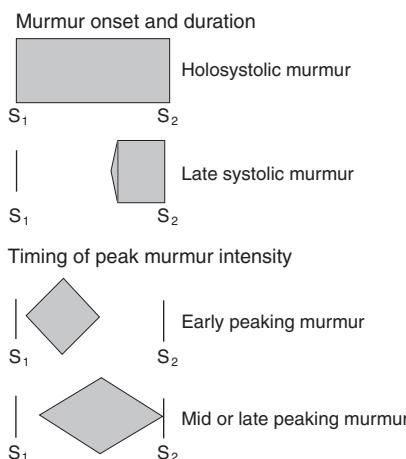


Figure 33-1 Select Features of Systolic Murmurs

In the holosystolic murmur, the murmur begins just after the first heart sound (S_1) and continues throughout the systole. In the late systolic murmur, the murmur begins at the middle of the systole or later and ends at the second heart sound (S_2). In an early peaking murmur, peak intensity is before the middle of the systole. In a mid- or late-peaking murmur, peak intensity is at the middle of the systole or later.

Valsalva during other maneuvers, such as sustained abdominal pressure or standing to squatting, so clinicians should ensure that patients breathe normally during these latter maneuvers.

PRECISION OF THE EXAMINATION RELATED TO SYSTOLIC MURMURS

Precision refers to agreement among clinicians regarding a particular clinical finding. The precision of the clinical examination for systolic murmurs has been evaluated in usual clinical circumstances by auscultating patients¹⁴⁻¹⁷ or in controlled nonclinical circumstances by listening to prerecorded audiotapes.¹⁸ Studies using audiotapes will yield higher estimates of precision, as will studies consisting of only normal

patients or very abnormal patients. Most of the available precision studies include patients with various causes of abnormal systolic murmurs, although one study included only patients with mild or moderate aortic stenosis.¹⁷ The experience of observers likely affects precision; all but one study¹⁶ used cardiologists as the examiners.

The only evaluated historical variable for diagnosing murmurs is effort syncope, which had a κ of 1.0 (simple agreement, 100) in one small study ($n = 22$).¹⁷ This study excluded patients with other types of syncope that could be confused with effort syncope, so it was relatively easy for the cardiologists to agree on the presence or absence of effort syncope.

One study found that the agreement between cardiology trainees on the carotid upstroke was poor, but data to calculate simple agreement or κ values were not provided.¹⁹ The precision of physical findings is summarized in Table 33-3.

The Bottom Line for Precision

- The precision of examining for any systolic murmur is moderate using audiotapes (κ , 0.48) but only fair in the clinical setting (κ , 0.30). The precision of examining for a loud (grade 2 or louder) systolic murmur is good using audiotapes (κ , 0.74) but only fair in the clinical setting (κ , 0.29).
- The precision of examining for a late-peaking systolic murmur is excellent (κ , 0.74).
- The precision of examining for a systolic click is good (simple agreement, 85%).

ACCURACY OF THE EXAMINATION RELATED TO SYSTOLIC MURMURS

To develop a structured search strategy, we used pertinent articles already in our files. Our strategy was deliberately broad to minimize the possibility of overlooking important articles. We then searched MEDLINE (English language) from 1966 through January 1996, using our structured search strategy (available on request). We manually reviewed

potentially relevant articles that we identified; we also reviewed the reference lists of these articles. We contacted authors of relevant studies for additional information.

Studies were reviewed by 2 independent readers (E.E. and C.B.). Disagreements between reviewers were resolved by discussion before a final quality grade was assigned. Quality grades were assigned using published guidelines (see Table 1-7 for a summary of Evidence Grades and Levels).²⁰ Grade A studies involve the independent, blind comparison of sign or symptom with a gold standard of diagnosis among a large number of consecutive patients suspected of having the target condition. Grade B studies involve the independent, blind comparison of sign or symptom with a gold standard of diagnosis among a small number of consecutive patients suspected of having the target condition. Grade C studies involve the independent, blind comparison of sign or symptom with a gold standard of diagnosis among nonconsecutive patients suspected of having the target condition; nonindependent comparison of sign or symptom with a gold standard of diagnosis among a sample of patients who obviously had the target condition plus, perhaps, normal individuals; or nonindependent comparison of a sign or symptom with a standard of uncertain validity.

Many of the studies were conducted in cardiology clinics, so the prevalence of abnormalities in these studies will be higher than in usual practice. For example, a study of patients undergoing cardiac catheterization for suspected aortic stenosis found a prevalence of aortic stenosis of 73%, so a positive clinical examination result virtually ruled in aortic stenosis. In usual practice, the prevalence of aortic stenosis would be much lower, so a positive clinical examination result would not rule in aortic stenosis, but rather indicate the need for further testing with echocardiography.

IS THIS AN ABNORMAL MURMUR?

Clinicians are primarily concerned whether a systolic murmur indicates a cardiac abnormality. In this context, the goal of the clinical examination is not an exact diagnosis, but rather identification of patients needing further testing to confirm or quantify an abnormality.

Several studies evaluated the accuracy of the entire clinical examination, including the medical history, physical examination, electrocardiogram, and chest radiograph; none has evaluated the history and physical examination alone.²¹⁻²⁴ In each study, cardiologists used the clinical examination to classify a systolic murmur as normal, possibly abnormal, or abnormal. Patients then underwent an echocardiogram or cardiac catheterization as the reference standard test. The most common abnormalities detected were valvular stenosis or regurgitation, atrial or ventricular septal defects, MVP, and cardiac hypertrophy. The study results, which are summarized in Table 33-4, indicate that cardiologists are efficient at identifying abnormal and normal murmurs.

The Bottom Line for Abnormal Murmur

- A clinical assessment of “normal murmur” by a cardiologist significantly reduces the likelihood of a cardiac abnormality.

Table 33-3 Precision of the Clinical Examination of Systolic Murmurs

Finding	Examiner	No.	κ^a	Simple Agreement, %
No murmur vs grades 1-4	Cardiologists (tapes) ¹⁸	100	0.48	70
	Cardiologists ¹⁴	100	0.30	54
	Cardiologists ¹⁵	80	...	86
	Cardiologists ¹⁶	32	...	97
	Noncardiologists ¹⁶	32	...	78
No murmur/grade 1 vs grades 2-4	Cardiologists (tapes) ¹⁸	100	0.74	87
	Cardiologists ¹⁴	100	0.29	76
Acoustic shape (late peaking vs not late peaking)	Cardiologists ¹⁷	22	0.74	95
Midsystolic click	Cardiologists ¹⁵	80	...	85

^aEllipses indicate data not available.

- A clinical assessment of “abnormal murmur” by a cardiologist significantly increases the likelihood of a cardiac abnormality.

Aortic Stenosis

Effort syncope is the only adequately studied individual historical variable. Presence of effort syncope in patients with a systolic murmur effectively rules in aortic stenosis (positive likelihood ratio [LR+], ∞ ; 95% confidence interval [CI], 1.3- ∞) but absence of effort syncope is not helpful (negative likelihood ratio [LR-], 0.76; 95% CI, 0.67-0.86) (grade C study).¹⁷

Several studies have examined the accuracy of the physical examination for detecting aortic stenosis. In these studies, echocardiography or cardiac catheterization confirmed aortic stenosis. Definitions of aortic stenosis varied, with peak instantaneous gradients ranging from as low as 25 mm Hg to as high as 50 mm Hg or aortic valve areas ranging from as low as 0.7 cm² to as high as 1.1 cm².

Many physical findings may increase or decrease the likelihood of aortic stenosis.^{17,25,26} Table 33-5 lists the findings beginning with the highest positive LRs from the largest studies with the best methodologic quality. All of the studies used cardiologist examiners.

Two studies are notable for their high methodologic quality and large sample sizes. The first study²⁵ involved 781 consecutive, unselected elderly patients who were nursing home residents. Each study participant received an examination by a single senior cardiologist, followed by an echocardiogram. Overall, 68 patients (9%) had aortic stenosis defined as a peak instantaneous Doppler gradient of 25 mm Hg or greater. This study provides a reasonable estimate of the accuracy of the clinical examination in an elderly population. Many of the patients had no symptoms and no audible murmur, which may have elevated the estimates of specificity and the positive LRs for some of the findings.

The second study²⁶ evaluated 231 consecutive patients referred for cardiac catheterization for various reasons, including suspected aortic stenosis. Cardiology fellows or cardiologists examined patients before cardiac catheterization. Overall, 113 patients (49%) had aortic stenosis, defined as a valve area of 0.8 cm² or less or a peak gradient of 50 mm Hg or greater, at cardiac catheterization. This study population was highly selected, so the prevalence of aortic stenosis was much higher than would be expected in usual clinical practice.

The accuracy of special maneuvers was evaluated by 2 trained cardiologists, who examined 50 nonconsecutive participants with a variety of heart diseases, including aortic stenosis, MR, ventricular septal defect, hypertrophic cardiomyopathy, pulmonic stenosis, and TR.²⁷ No maneuver was useful for ruling in aortic stenosis (data not shown), but certain findings from the Valsalva maneuver reduced the likelihood of aortic stenosis (Table 33-5).

A potentially useful multivariate decisional aid for diagnosing aortic stenosis was developed using split-sample validation (Table 33-6).²⁶ The study showed an excellent positive LR for patients with point scores higher than 10. One of the variables in this model was aortic valve calcification on the lateral chest radiograph.

Table 33-4 Accuracy of Clinical Examination for Detecting Abnormal Systolic Murmur

Overall Clinical Assessment	LR (95% CI)	Quality Grade ^a
Abnormal Murmur		
Study 1 ^b	∞ (14- ∞)	A
Study 2 ^c	∞ (2.8- ∞)	C
Study 3 ^d	3.8 (2.8-5.4)	C
Possibly Abnormal Murmur		
Study 1 ^b	2.3 (0.7-5.9)	A
Study 2 ^e	1.3 (1.2-1.4)	C
Normal Murmur		
Study 1 ^b	0 (0-0.4)	A
Study 2 ^c	0.01 (0-0.02)	C
Study 3 ^d	0.05 (0.01-0.20)	C
Study 4 ^c	0.3 (0.1-0.6)	C

Abbreviations: CI, confidence interval; LR, likelihood ratio.

^aSee Table 1-7 for a summary of Evidence Grades and Levels.

^bOf 103 patients, 93 had normal murmurs. The study was conducted among pregnant patients. Reference standard: echocardiogram.

^cOf 30 patients, 16 had normal murmurs. Reference standard: echocardiogram.

^dOf 1059 patients, 100 had normal murmurs. Reference standard: cardiac catheterization.

^eOf 532 patients, 378 had normal murmurs. Reference standard: echocardiogram.

Although the preceding results are encouraging, 2 small studies had less impressive results. The first study included 75 patients with severe multivalvular disease who were undergoing cardiac catheterization and found that a cardiologist’s clinical diagnosis of aortic stenosis was only reasonably accurate (LR+, 3.7; 95% CI, 2.2-7.0; LR-, 0.23; 95% CI, 0.11-0.44).²⁸ Many of these patients had severe multivalvular disease, which may have made an exact diagnosis more difficult. A study on 35 elderly patients with systolic murmurs who were examined by a geriatrician found that a clinical diagnosis of neither “aortic stenosis present” (LR+, 2.4; 95% CI, 0.72-6.9) nor “aortic stenosis absent” (LR, 0.7; 95% CI, 0.30-1.1) was accurate.²⁹ This study suggests that assessments by cardiologists may be better than assessments by noncardiologists.

The Bottom Line for Aortic Stenosis

- The presence of any of the following clinical findings significantly increases the likelihood of aortic stenosis: effort syncope, slow rate of increase of the carotid pulse, timing of peak murmur intensity in late or midsystole, decreased intensity or absent S₂, apical-carotid delay, or brachioradial delay.
- The absence of any of the following clinical findings significantly reduces the likelihood of aortic stenosis: any systolic murmur or murmur radiation to the right carotid artery.
- Combinations of the following clinical variables can be useful to rule in or rule out aortic stenosis: decreased carotid volume, delayed carotid upstroke, decreased or absent S₂, murmur loudest at second right intercostal space, and valve calcification on chest radiograph.

Table 33-5 Accuracy of the Physical Examination for Detecting Aortic Stenosis

Finding	Reference Standard (No. of Patients)	LR+ (95% CI)	LR- (95% CI)	Quality Grade ^a
Slow rate of increase of carotid pulse				
Study 1 ²⁵	Cardiac catheterization (781)	130 (33-560)	0.62 (0.51-0.75)	A
Study 2 ²⁶	Cardiac catheterization (231)	2.8 (2.1-3.7)	0.18 (0.11-0.30)	C ^b
Study 3 ¹⁷	Cardiac catheterization (106)	6.4 (0.8-45)	0.73 (0.59-0.90)	C
Timing of peak murmur intensity				
Late peaking ²⁵	Cardiac catheterization (781)	101 (25-410)	0.31 (0.22-0.44)	A
Midpeaking ¹⁷	Cardiac catheterization (106)	8.0 (2.7-23)	0.13 (0.07-0.24)	C
Decreased intensity or absent second heart sound				
Study 1 ²⁵	Cardiac catheterization (781)	50 (24-100)	0.45 (0.34-0.58)	A
Study 2 ²⁶	Cardiac catheterization (231)	3.1 (2.1-4.3)	0.36 (0.26-0.49)	C ^b
Apical carotid delay ⁶	Cardiac catheterization (44)	∞ (2.4-∞)	0.05 (0.01-0.31)	C
Brachioradial delay ⁵	Echocardiogram (58)	6.8 (3.2-14)	0.0 (0.0-0.3)	C
Fourth heart sound ²⁵	Cardiac catheterization (781)	2.5 (2.1-3.0)	0.26 (0.14-0.49)	A
Presence of any murmur ²⁵	Cardiac catheterization (781)	2.4 (2.2-2.7)	0 (0-0.13)	A
Reduced carotid volume				
Study 1 ²⁶	Cardiac catheterization (231)	2.3 (1.7-3.0)	0.31 (0.21-0.46)	C ^b
Study 2 ¹⁷	Cardiac catheterization (106)	2.2 (1.2-4.2)	0.39 (0.22-0.69)	C
Radiation to right carotid				
Study 1 ²⁵	Cardiac catheterization (781)	1.4 (1.3-1.5)	0.10 (0.13-0.40)	A
Study 2 ²⁶	Cardiac catheterization (231)	1.5 (1.3-1.7)	0.05 (0.01-0.20)	C ^b
With Valsalva maneuver intensity is decreased ²⁷	Cardiac catheterization (50)	1.2 (0.8-1.6)	0 (0-1.6)	C

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

^aSee Table 1-7 for a summary of Evidence Grades and Levels.

^bGrade A study except cardiac catheterization interpreted with knowledge of clinical findings.

Table 33-6 Multivariable Decision Rule for Suspected Aortic Stenosis²⁶

Point Score	Aortic Stenosis ^a		
	Yes	No	LR (95% CI) ^b
14	7	0	∞ (0.6-∞)
10-13	22	1	8.0 (1.6-46)
7-9	22	3	2.7 (1.0-8.0)
2-6	11	15	0.27 (0.15-0.49)
0	1	4	0.10 (0.01-0.58)
Total	63	23	...
Variable	Point Score		
Reduced carotid volume	2		
Slow rate of increase of carotid pulse	3		
Murmur loudest at second right intercostal space	2		
Decreased or absent second heart sound	3		
Valve calcification on chest radiograph	4		
Maximum score	14		

Abbreviations: CI, confidence interval; LR, likelihood ratio.

^aDefined as peak transvalvular gradient of > 50 mm Hg at cardiac catheterization.

^bEllipsis indicates not applicable.

Mitral Regurgitation

We report the accuracy of the clinical examination for detecting moderate to severe regurgitation confirmed through echocardiography or cardiac catheterization (Table 33-7). Detection of moderate to severe MR, even in asymptomatic patients, may influence recommendations for echocardiographic monitoring³⁰ or medical treatment.³¹

If a cardiologist hears a murmur in the mitral area (mid left thorax, fifth intercostal space), then the likelihood of MR is increased slightly, but absence of a murmur significantly reduces the likelihood of MR.^{15,32,33} Similarly, a late systolic or holosystolic murmur slightly increases the likelihood of MR, but absence of such a murmur significantly reduces the likelihood of MR. In the setting of acute MI, absence of a murmur is less useful for ruling out acute MR (LR-, 0.66; 95% CI, 0.25-1.0).³⁴ Transient arterial occlusion was accurate for ruling in and ruling out left-sided regurgitant murmurs, such as MR and ventricular septal defect.²⁷

Internal medicine house staff are less accurate than cardiologists for detecting the murmur of MR, with positive LRs ranging from 1.1 (for interns) to 4.6 (for medical students) and negative LRs ranging from 0.7 (for junior residents) to 1.0 (for interns and senior residents)³⁵ (grade A study).

Table 33-7 Accuracy of the Clinical Examination for Detecting Mitral Regurgitation

Finding		Reference Standard (No. of Patients)	LR+ (95% CI)	LR- (95% CI)	Quality Grade ^a
Murmur in mitral area	Study 1 ³³ Study 2 ³²	Echocardiogram: moderate to severe MR (394) Cardiac catheterization: moderate to severe MR (35)	3.9 (3.0-5.1) 3.6 (1.9-7.7)	0.34 (0.23-0.47) 0.12 (0.02-0.50)	C C
Late or holosystolic murmur ¹⁵		Echocardiogram: moderate to severe MR (80)	1.8 (1.2-2.5)	0 (0-0.8)	C
Any murmur during acute MI ³⁴		Cardiac catheterization: moderate to severe MR (206)	4.7 (1.3-11)	0.66 (0.25-1.0)	C
With transient arterial occlusion, murmur increases in intensity ²⁷		Cardiac catheterization: severity not stated ^b	7.5 (2.5-23)	0.28 (0.13-0.60)	C

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio; MI, myocardial infarction; MR, mitral regurgitation.

^aSee Table 1-7 for a summary of Evidence Grades and Levels.

^bOutcome of interest was left-sided regurgitant lesions, including MR or ventricular septal defect.

The Bottom Line for Mitral Regurgitation

- For cardiologists, absence of a mitral area murmur or a late systolic/holosystolic murmur significantly reduces the likelihood of MR, except in the setting of acute MI.
- Cardiologists can accurately distinguish left-sided regurgitant murmurs, such as MR and ventricular septal defect, using transient arterial occlusion.
- Noncardiologists' assessments for MR are considerably less accurate.

Tricuspid Regurgitation

Cardiologists are reasonably accurate for diagnosing the murmur of moderately severe to severe TR in patients ($n = 21$, with TR; $n = 295$, without TR) referred for echocardiography (LR+, 10.1; 95% CI, 5.8-18; LR-, 0.41; 95% CI, 0.24-0.70) (grade C).³³ Special maneuvers may also be helpful for diagnosing TR and other right-sided lesions such as pulmonic stenosis. One study ($n = 10$, with TR or pulmonic stenosis; $n = 40$, without TR or pulmonic stenosis) using cardiologist examiners found that an increase in murmur intensity with inspiration significantly increased the likelihood of a right-sided valvular lesion, whereas the absence of increased intensity made these conditions less likely (LR+, 8.0; 95% CI, 3.5-18; LR-, 0.0; 95% CI, 0-0.43) (grade C).²⁷ In another study, patients with severe MR ($n = 15$) or TR ($n = 15$) were examined by experienced cardiologists before cardiac catheterization.¹⁰ To distinguish TR from MR, increased murmur intensity on inspiration had a positive LR of ∞ (95% CI, 3.1- ∞) and a negative LR of 0.20 (95% CI, 0.07-0.45). For the finding of increased murmur intensity with sustained abdominal pressure, the positive LR was ∞ (95% CI, 2.5- ∞) and the negative LR was 0.33 (95% CI, 0.15-0.58) (grade C).

The Bottom Line for Tricuspid Regurgitation

- Cardiologists can accurately detect the murmur of TR.
- Cardiologists can accurately rule in and rule out TR with the quiet inspiration and sustained abdominal pressure maneuvers.

Hypertrophic Cardiomyopathy

There are limited data on the accuracy of clinical examination for hypertrophic cardiomyopathy (also termed *idiopathic*

hypertrophic subaortic stenosis). Many studies evaluate phonocardiography or intracardiac tracings rather than auscultation,³⁶⁻⁴⁰ whereas others include fewer than 15 patients.⁴¹⁻⁴⁵ One study evaluated carotid sinus pressure, which is not routinely recommended for the clinical examination.⁴⁶

Special maneuvers may help distinguish the murmur of hypertrophic cardiomyopathy.²⁷ Using cardiologist examiners, if a murmur decreased in intensity with passive leg elevation, then hypertrophic cardiomyopathy was significantly more likely (LR+, 8.0; 95% CI, 3.0-21), whereas if the murmur did not decrease in intensity, the likelihood was significantly reduced (LR-, 0.22; 95% CI, 0.06-0.77). If murmur intensity was decreased or unchanged with standing to squatting, then hypertrophic cardiomyopathy was significantly more likely (LR+, 4.5; 95% CI, 2.3-8.6), whereas if the murmur increased in intensity, the likelihood of hypertrophic cardiomyopathy was significantly reduced (LR-, 0.13; 95% CI, 0.02-0.81) (grade C).

The Bottom Line for Hypertrophic Cardiomyopathy

Cardiologists can rule in or rule out hypertrophic cardiomyopathy by evaluating for decreased murmur intensity with passive leg elevation or increased murmur intensity when the patient goes from a squatting to standing position.

Mitral Valve Prolapse

The accuracy of the clinical examination for diagnosing MVP cannot be defined, because clinical findings alone are sufficient for the diagnosis of MVP. A patient with a systolic click and a systolic murmur meets the diagnostic criteria for MVP even if the patient has a normal echocardiogram result.^{47,48}

However, we can examine the relationship between clinical findings and echocardiographic findings (Table 33-8).⁴⁹⁻⁵³ With cardiologist examiners, a systolic click accompanied by a systolic murmur helped to rule in echocardiographic MVP. The accuracy of an isolated systolic click is variable, possibly because of unreliability of the clinical examination and differences between studies regarding the definition of echocardiographic MVP. An isolated systolic murmur has little effect on the likelihood of echocardiographic MVP, whereas absence of both a systolic click and a murmur appears to reduce the likelihood of echocardiographic MVP. Noncardiologists are less accurate than cardiologists for all of these findings.

Table 33-8 Accuracy of the Clinical Examination for Detecting Echocardiographic Mitral Valve Prolapse

Finding	Clinician (No. of Patients)	LR (95% CI)	Quality Grade ^a
Systolic click and murmur			
Study 1 ⁴⁹	Cardiologists (401)	19 (4.6-80)	C
Study 2 ⁵⁰	Noncardiologists (104)	2.4 (1.0-5.7)	C
Systolic click			
Study 1 ⁴⁹	Cardiologists (401)	12 (5.4-25)	C
Study 2 ⁵⁰	Noncardiologists (104)	1.3 (0.7-2.2)	C
Nonejection click, with or without a murmur			
Study 1 ⁵¹	Cardiologists (155)	3.8 (2.3-6.8)	A
Study 2 ⁵²	Cardiologists (140)	1.7 (1.3-2.1)	C
Murmur, with or without a systolic click			
Study 1 ⁵²	Cardiologists (140)	1.9 (1.3-3.0)	C
Study 2 ⁵³	Noncardiologists (259)	1.2 (0.9-1.5)	C
Murmur only			
Study 1 ⁵⁰	Cardiologists (401)	2.4 (1.0-5.7)	C
Study 2 ⁵¹	Noncardiologists (104)	0.7 (0.3-1.3)	C
No murmur, no systolic click			
Study 1 ⁵¹	Cardiologists (155)	0.04 (0.02-0.11)	A
Study 2 ⁵²	Cardiologists (140)	0.26 (0.12-0.54)	C
Study 3 ⁴⁹	Cardiologists (401)	0.21 (0.15-0.29)	C
Study 4 ⁵⁰	Noncardiologists (104)	0.53 (0.23-1.20)	C

Abbreviations: CI, confidence interval; LR, likelihood ratio.

^aSee Table 1-7 for a summary of Evidence Grades and Levels.**Table 33-9** Accuracy of the Clinical Examination for Predicting Adverse Clinical Outcomes Related to Mitral Valve Prolapse^a

Finding	Clinician (No. of Patients)	LR (95% CI)	Quality Grade ^b
Holosystolic murmur			
Study 1 ⁵⁶	Cardiologists (316)	18 (6.6-51)	C
Study 2 ⁵⁷	Cardiologists (321)	5.1 (2.2-9.9)	C
Late systolic murmur or click and murmur			
Study 1 ⁵⁸	Cardiologists (316)	1.2 (0.7-1.7)	C
Study 2 ⁵⁷	Cardiologists (321)	0.8 (0.3-1.5)	C
Click and holosystolic murmur ⁵⁷	Cardiologists (321)	0.8 (0.2-2.4)	C
Any click or isolated click			
Study 1 ⁵⁸	Cardiologists (316)	0.4 (0.2-0.8)	C
Study 2 ⁵⁷	Cardiologists (321)	0.26 (0.05-1.1)	C
No click/no murmur			
Study 1 ⁵⁴	Cardiologists (237)	0 (0-4.1)	C
Study 2 ⁵⁸	Cardiologists (316)	0 (0-1.4)	C

Abbreviations: CI, confidence interval; LR, likelihood ratio.

^aOutcomes include death (cardiac or all-cause, depending on study), stroke, endocarditis, or progressive mitral regurgitation requiring surgery. Most outcomes were progressive mitral regurgitation requiring surgery.^bSee Table 1-7 for a summary of Evidence Grades and Levels.

Mitral valve leaflet redundancy or thickening is the echocardiographic variable most strongly associated with adverse clinical outcomes.^{54,55} In one study, neither a systolic click (LR+, 2.8; 95% CI, 1.8-4.6; LR-, 0.76; 95% CI, 0.69-0.84) nor a systolic murmur (LR+, 1.3; 95% CI, 1.1-1.5; LR-, 0.57; 95% CI, 0.43-0.76) affected the likelihood of echocardiographic mitral valve leaflet thickening or redundancy (grade C study).⁵⁶

Several longitudinal studies of patients with echocardiographic MVP have related baseline clinical findings to the development of adverse clinical events, including cardiac death, progressive MR requiring surgery, endocarditis, and systemic embolism.^{57,58} A holosystolic murmur without a systolic click significantly increased the likelihood of an adverse event, whereas absence of both a systolic click and murmur was associated with no adverse events. Other clinical findings had little effect on the likelihood of adverse events (Table 33-9).

The Bottom Line for Mitral Valve Prolapse

- A systolic click, with or without systolic murmur, is sufficient for the diagnosis of MVP.
- If a cardiologist hears a systolic click, with or without a murmur, then the likelihood of echocardiographic MVP is significantly increased. The absence of both a systolic click and murmur significantly reduces the likelihood of echocardiographic MVP.
- In patients with echocardiographic MVP, a holosystolic murmur without a systolic click significantly increases the likelihood of long-term complications, whereas absence of both a systolic click and murmur significantly reduces the likelihood of long-term complications.

WHEN TO EXAMINE FOR SYSTOLIC MURMURS

We are unaware of data by which one might give an evidence-based recommendation regarding the examination for systolic murmurs. Auscultation for systolic murmurs should probably be carried out in any patient for whom a complete cardiovascular database is necessary.

ARE SYSTOLIC MURMURS EVER NORMAL?

In unreferred young adults, the prevalence of systolic murmurs ranges from 5% to 52%^{8,59-61}; echocardiography result is normal in 86% to 100%.⁶²⁻⁶⁴ Echocardiography result is normal in 90% to 94% of pregnant women with systolic murmurs who are referred for testing.^{21,24,65} In elderly medical outpatients or residents of long-term care facilities, the prevalence of systolic murmurs ranges from 29% to 60%⁶⁶⁻⁶⁸; echocardiography is normal in 44% to 100%.^{24,25,29,69,70} This wide range of normal in the elderly can be partially explained by various study definitions of normal echocardiograms. Commonly detected abnormalities in the elderly were left ventricular systolic dysfunction, aortic stenosis, and MR. Other studies include aortic valve sclerosis as an abnormality, although the clinical importance of aortic valve sclerosis is uncertain.

A venous hum⁷¹ and a mammary souffle are both normal conditions that present either as systolic murmurs or, more commonly, as continuous murmurs.

HOW TO IMPROVE SKILLS IN EXAMINING THIS AREA

The characteristics of murmurs can be learned using cardiovascular auscultatory tapes or cardiac patient simulators, although the effectiveness of these aids is uncertain.^{72,73} Most audiotapes are accompanied by phonocardiographic and expert cardiologist analyses, so these tapes can help clinicians to calibrate their ears to those of experts.

Most commercially available stethoscopes have similar acoustic properties, although some have poor performance at low frequencies.⁷⁴ Good stethoscope maintenance is essential because dirt or cracked tubing⁷⁵ will significantly reduce accuracy. Large earpieces are better because small earpieces can be occluded by the sharp bony angle at the external auditory meatus.³

At the bedside, eliminate background noise whenever possible. If background noise is unavoidable, try to repeat your examination in a quieter setting.

Finally, relate your clinical findings to the results of assessments by a colleague, a cardiologist, or an echocardiogram whenever possible. Resolving disagreements between your assessments and those of others is an excellent way of upgrading your clinical skills.

RECOMMENDATIONS FOR FURTHER RESEARCH

Most studies used cardiologists or senior cardiology fellows to conduct the clinical examinations. There are few data on the precision and accuracy of the clinical examination conducted by noncardiologists. Some studies include inappropriately narrow spectrums of patients, such as only patients with moderate and severe aortic stenosis.^{5,6,17} Further studies should focus on a broad spectrum of patients from primary or secondary care settings, particularly patients older than 40 years when the prevalence of abnormal murmurs is significantly increased.

CLINICAL SCENARIOS—RESOLUTIONS

CASE 1 Your first patient, who is awaiting urgent surgery for an open fracture, had a systolic murmur that did not radiate to the right carotid artery. The likelihood of aortic stenosis is significantly reduced by this finding. In addition, the carotid artery pulsation had normal volume, the S₂ intensity was normal, and there was no S₄. These findings also help to reduce the likelihood of aortic stenosis. You are confident in your assessment because it was conducted in a quiet room with a comfortable and cooperative patient. You can advise the surgeon that aortic stenosis is unlikely.

CASE 2 Your second patient has a systolic click and a systolic murmur, strongly suggesting MVP. If you are an experienced auscultator, then these findings significantly increase

the likelihood that the echocardiogram will show evidence of MVP. However, even if the echocardiogram result is normal, you already have enough evidence to diagnose MVP. You may wish to obtain an echocardiogram at a later date to determine the severity of the valvular abnormality.

Author Affiliations at the Time of the Original Publication

Division of General Internal Medicine and Clinical Epidemiology, Department of Medicine, University of Toronto, Toronto, Ontario, Canada; and The Toronto Hospital, Toronto, Ontario, Canada.

Acknowledgments

We thank Wilbert Aronow, MD, for providing additional data and methodologic information about his studies, Eugene Oddone, MD, MHS, for his helpful comments on earlier drafts of this article, and Ms Sharon Smith for her assistance in the preparation of the manuscript and tables.

REFERENCES

- Ballard DJ, Khanderia BK, Tajik AJ, Seward JB, Weber VP, Melton LJ. Population-based study of echocardiography: time trends in utilization and diagnostic profile of an evolving technology. *Int J Technol Assess Health Care*. 1989;5(2):249-261.
- Rushmer RF, Sparkman DR, Polley RFL, et al. Variability in detection and interpretation of heart murmurs. *Am J Dis Child*. 1952;83(6):740-754.
- Constant J. *Essentials of Bedside Cardiology*. 2nd ed. Totowa, NJ: Humana Press; 2003:30, 121-153.
- Marriot HJ. *Bedside Cardiac Diagnosis*. Philadelphia, PA: JB Lippincott; 1991:19.
- Leach RM, McBrien DJ. Brachioradial delay: a new clinical indicator of the severity of aortic stenosis. *Lancet*. 1990;335(8699):1199-1201.
- Chun PKC, Dunn BE. Clinical clue of severe aortic stenosis: simultaneous palpation of the carotid and apical impulses. *Arch Intern Med*. 1982;142(13):2284-2288.
- Levine SA. The systolic murmur: its clinical significance. *JAMA*. 1933; 101:436-438.
- Freeman AR, Levine SA. The clinical significance of the systolic murmur: a study of 1000 consecutive cardiac cases. *Ann Intern Med*. 1933;6:1371-1385.
- Rivero-Carvalho JM. Signo para el diagnostico de las insuficiencias tricuspideas. *Arch Inst Cardiol Mex*. 1946;16:531-540.
- Maisel A, Atwood JE, Goldberger AL. Hepatojugular reflux: useful in the bedside diagnosis of tricuspid regurgitation. *Ann Intern Med*. 1984;101 (6):781-782.
- Lembo NJ, Dell'Italia LJ, Crawford MH, O'Rourke RA. Diagnosis of left-sided regurgitant murmurs of transient arterial occlusion: a new maneuver using blood pressure cuffs. *Ann Intern Med*. 1986;105(3):368-370.
- Sharpey-Schafer EP. Effects of squatting on the normal and failing circulation. *BMJ*. 1956;1(4975):1072-1074.
- Ravin A, Craddock LD, Wolf PS, Shander D. *Auscultation of the Heart*. 3rd ed. Chicago, IL: Mosby-Year Book; 1977:2:45-67.
- Taranta A, Spagnuolo M, Snyder R, Gerbarg DS, Hofler JJ. Auscultation of the heart by physicians and by computer. In: *Data Acquisition and Processing in Biology and Medicine*, Vol 3. New York, NY: Macmillan Publishing Co; 1964:23-52.
- Panidis IP, McAllister M, Ross J, Mintz GS. Prevalence and severity of mitral regurgitation in the mitral valve prolapse syndrome: a Doppler echocardiographic study of 80 patients. *J Am Coll Cardiol*. 1986;7(5):975-981.
- Raftery EB, Holland WW. Examination into the heart: an investigation into variation. *Am J Epidemiol*. 1967;85(3):438-444.
- Forssell G, Jonasson R, Orinius E. Identifying severe aortic valvular stenosis by bedside examination. *Acta Med Scand*. 1985;218(4):397-400.

18. Dobrow RJ, Calatayud JB, Abraham S, Caceres CA. A study of physician variation in heart sound interpretation. *Med Annu Dist Columbia*. 1964;33:305-308.
19. Spodick DH, Sugiuira T, Doi Y, Paladino D, Haffty BG. Rate of rise of the carotid pulse: an investigation of observer error in a common clinical measurement. *Am J Cardiol*. 1982;49(1):159-162.
20. Holleman DR, Simel DL. Does the clinical examination predict airflow limitation? *JAMA*. 1995;273(4):313-319.
21. Mishra M, Chambers JB, Jackson G. Murmurs in pregnancy: an audit of echocardiography. *BMJ*. 1992;304(6839):1413-1414.
22. Ahu IM. Functional systolic murmurs. *Indian Heart J*. 1982;34(4):241-244.
23. Lockhart PB, Crist D, Stone PH. The reliability of the medical history in the identification of patients at risk for infective endocarditis. *J Am Dent Assoc*. 1989;119(3):417-422.
24. Xu M, McHaffie DJ. Nonspecific systolic murmurs: an audit of the clinical value of echocardiography. *N Z Med J*. 1993;106(950):54-56.
25. Aronow WS, Kronzon I. Correlation of prevalence and severity of valvular aortic stenosis determined by continuous-wave Doppler echocardiography with physical signs of aortic stenosis in patients aged 62 to 100 years with aortic systolic ejection murmurs. *Am J Cardiol*. 1987;60(4):399-401.
26. Hoagland PM, Cook EF, Wynne J, Goldman L. Value of noninvasive testing in adults with suspected aortic stenosis. *Am J Med*. 1986;80(6):1041-1050.
27. Lembo NJ, Dell'Italia LJ, Crawford MH, O'Rourke RA. Bedside diagnosis of systolic murmurs. *N Engl J Med*. 1988;318(24):1572-1578.
28. Jaffe WM, Roche AHG, Coverdale HA, McAlister HF, Ormiston JA, Greene ER. Clinical evaluation versus Doppler echocardiography in the quantitative assessment of valvular heart disease. *Circulation*. 1988;78(2):267-275.
29. McKillop GM, Stewart DA, Burns JMA, Ballantyne D. Doppler echocardiography in elderly patients with ejection systolic murmurs. *Postgrad Med J*. 1991;61(794):1059-1061.
30. Gaasch WH, John RM, Aurigemma GP. Managing asymptomatic patients with chronic mitral regurgitation. *Chest*. 1995;108(3):842-847.
31. Seneviratne B, Morre GA, West PD. Effect of captopril on functional mitral regurgitation in dilated heart failure: a randomized double blind placebo controlled trial. *Br Heart J*. 1994;72(1):63-68.
32. Meyers DG, McCall D, Sears TD, Olson TS, Felix GL. Duplex pulsed Doppler echocardiography in mitral regurgitation. *J Clin Ultrasound*. 1986;14(2):117-121.
33. Rahko PS. Prevalence of regurgitant murmurs in patients with valvular regurgitation detected by Doppler echocardiography. *Ann Intern Med*. 1989;111(6):466-472.
34. Lehmann KG, Francis CK, Dodge HT; TIMI Study Group. Mitral regurgitation in early myocardial infarction: incidence, clinical detection, and prognostic implications. *Ann Intern Med*. 1992;117(1):10-17.
35. Kinney EL. Causes of false-negative auscultation of regurgitant lesions: a Doppler echocardiographic study of 294 patients. *J Gen Intern Med*. 1988;3(5):429-434.
36. Buda A, Mackenzie G, Wigle D. The Mueller maneuver in muscular subaortic stenosis [abstract]. *Circulation*. 1977;56(suppl 3):III-138.
37. Nellen M, Gotsman MS, Vogelpoel L, Beck W, Schrire V. Effects of prompt squatting on the systolic murmur in idiopathic hypertrophic obstructive cardiomyopathy. *BMJ*. 1967;3(5558):140-143.
38. Braunwald E, Oldham HN Jr, Ross J Jr, Linhart JW, Mason DT, Fort I III. The circulatory response of patients with idiopathic hypertrophic subaortic stenosis to nitroglycerin and the Valsalva maneuver. *Circulation*. 1964;29:422-431.
39. Kramer DS, French WJ, Criley JM. The postextrasystolic murmur response to gradient in hypertrophic cardiomyopathy. *Ann Intern Med*. 1986;104(6):772-776.
40. Petrin TJ, Tavel ME. Idiopathic hypertrophic subaortic stenosis as observed in a large community hospital. *J Am Geriatr Soc*. 1979;27(1):43-46.
41. Rosenblum R, Delman AJ. Valsalva's maneuver and the systolic murmur of hypertrophic subaortic stenosis. *Am J Cardiol*. 1965;15:868-870.
42. Stefadourous MA, Mucha E, Frank MJ. Paradoxical response of the murmur of idiopathic hypertrophic subaortic stenosis to the Valsalva maneuver. *Am J Cardiol*. 1976;37(1):89-92.
43. Bartall H, Amber S, Desser KB, Benchimol A. Normalization of the external carotid pulse tracing of hypertrophic subaortic stenosis during Muller's maneuver. *Chest*. 1978;74(1):77-78.
44. Cassidy J, Aronow WS, Prakash R. The effect of isometric exercise on the systolic murmur of patients with idiopathic hypertrophic subaortic stenosis. *Chest*. 1975;67(4):395-397.
45. Battle WE, Siegel FA, Fox LM. The older patient with idiopathic hypertrophic subaortic stenosis. *Geriatrics*. 1977;1:61-69.
46. Klein HO, DiSegni E, Dean H, Beker B, Bakst A, Kaplinksy E. Increased intensity of the murmur of hypertrophic obstructive cardiomyopathy with carotid sinus pressure. *Chest*. 1988;93(4):814-820.
47. Perloff JK, Child JS, Edwards JE. New guidelines for the clinical diagnosis of mitral valve prolapse. *Am J Cardiol*. 1986;57(13):1124-1129.
48. Devereux RB, Kramer-Fox R, Shear K, Kligfield P, Pini R, Savage DD. Diagnosis and classification of severity of mitral valve prolapse: methodologic, biologic, and prognostic consideration. *Am Heart J*. 1987;113(5):1265-1280.
49. Devereux RB, Kramer-Fox R, Brown WT, et al. Relation between clinical features of the mitral valve prolapse syndrome and echocardiographically documented mitral valve prolapse. *J Am Coll Cardiol*. 1986;8(4):763-772.
50. Olive KE, Grassman ED. Mitral valve prolapse: comparison of diagnosis by physical examination and echocardiography. *South Med J*. 1990;83(11):1266-1269.
51. Abbas AS, DeCristofaro D, Anabtawi J, Irwin L. Mitral valve prolapse: comparative value of M-mode, two-dimensional and Doppler echocardiography. *J Am Coll Cardiol*. 1983;2(6):1219-1223.
52. Barron JT, Manrose DL, Liebson PR. Comparison of auscultation with two-dimensional and Doppler echocardiography in patients with suspected mitral valve prolapse. *Clin Cardiol*. 1988;11(6):401-406.
53. Retchin SM, Waugh RA, Fletcher RH. The impact of echocardiographic results on treatment decisions for patients suspected of mitral valve prolapse. *Med Care*. 1987;25(7):652-657.
54. Nishimura RA, McGoon MD, Shub C, Miller FA Jr, Ilstrup DM, Tajik AJ. Echocardiographically documented mitral-valve prolapse: long-term follow-up of 237 patients. *N Engl J Med*. 1985;313(21):1305-1309.
55. Orenica AJ, Petty GW, Khanderia BK, et al. Risk of stroke with mitral valve prolapse in population-based cohort survey. *Stroke*. 1995;26(1):7-13.
56. Marks AR, Choong CY, Sanfilippo AJ, Ferre M, Weyman AE. Identification of high-risk and low-risk subgroups of patients with mitral valve prolapse. *N Engl J Med*. 1989;320(16):1031-1036.
57. Tofler OB, Tofler GH. Use of auscultation to follow patients with mitral systolic clicks and murmurs. *Am J Cardiol*. 1990;66(19):1355-1358.
58. Zuppiroli A, Rinaldi M, Kramer-Fox R, Favilli S, Roman MJ, Devereux RB. Natural history of mitral valve prolapse. *Am J Cardiol*. 1995;75(15):1028-1032.
59. McCracken D, Everett JE. An investigation of the incidence of cardiac murmurs in young healthy women. *Practitioner*. 1976;216(1293):308-309.
60. Stewart IMG. Systolic murmurs in 525 healthy young adults. *Br Heart J*. 1951;13(4):561-565.
61. Vaughan JP. Blood pressure and heart murmurs in a rural population in the United Republic of Tanzania. *Bull World Health Organ*. 1979;57(1):89-97.
62. Cotter L, Logan RL, Poole A. Innocent systolic murmurs in healthy 40-year-old men. *J R Coll Physicians Lond*. 1980;14(2):128-129.
63. Markiewicz W, Stoner J, London E, Hunt SA, Popp RL. Mitral valve prolapse in one hundred presumably healthy young females. *Circulation*. 1976;53(3):464-473.
64. Tan CC, Hiew TM. Innocent systolic murmurs in healthy young males: a study of the characteristics and prevalence in the SAF pre-enlistee population. *Singapore Med J*. 1988;29(4):337-340.
65. Northcote RJ, Knight PV, Ballantyne D. Systolic murmurs in pregnancy: value of echocardiographic assessment. *Clin Cardiol*. 1985;8(6):327-328.
66. Bethel CS, Crow EW. Heart sounds in the aged. *Am J Cardiol*. 1963;11:763-767.
67. Bruns DL, Van Der Hauwaert LG. The aortic systolic murmur developing with increased age. *Br Heart J*. 1958;20(3):370-378.
68. Griffiths RA, Sheldon MG. The clinical significance of systolic murmurs in the elderly. *Age Ageing*. 1975;4(2):99-104.
69. Perez GL, Jacob M, Bhat PK, Rao DB, Luisada AA. Incidence of murmurs in the aging heart. *J Am Geriatr Soc*. 1976;24(1):29-31.
70. Knight PV, Martin BJ, Ballantyne D. Echocardiographic diagnoses in elderly patients with systolic murmurs and cardiac disease. *Age Ageing*. 1986;15(3):169-173.
71. Bujak W, Gioia F, Cayler GG. An innocent thrill: a common finding with an innocent murmur. *JAMA*. 1976;235(22):2417.
72. Oddone EZ, Waugh RA, Samsa G, Corey R, Feussner JR. Teaching cardiovascular examination skills: results from a randomized controlled trial. *Am J Med*. 1993;95(4):389-396.
73. Ewy GA, Felner JM, Juul D, Mayer JW, Sajid AW, Waugh RA. Test of a cardiology patient simulator with students in fourth year electives. *J Med Educ*. 1987;62(9):738-743.
74. Abella M, Formolo J, Penney DG. Comparison of the acoustic properties of six popular stethoscopes. *J Acoust Soc Am*. 1992;91(4 pt 1):2224-2228.
75. Orton D, Stryker R. Sick stethoscope syndrome. *JAMA*. 1986;256(20):2817.

Prepared by David Cescon, MD, and Edward Etchells, MD, MSC

Reviewed by Eugene Oddone, MD

CLINICAL SCENARIO

A 62-year-old man scheduled for elective total knee replacement has been referred to you for preoperative assessment of a systolic murmur. The orthopedic surgeon detected a systolic murmur and wants to rule out aortic stenosis (AS) before surgery. The patient has no cardiovascular symptoms. On auscultation, you hear normal first and second heart sounds (S_1 and S_2). There is a grade 3 early systolic murmur, loudest at the lower left sternal border, which does not radiate to either the right clavicle or carotids. You detect a normal volume and normal rate of increase of the carotid pulse. The rest of the clinical examination results, including those for the electrocardiogram (ECG) and chest radiograph, are normal.

Original Review

Etchells EE, Bell C, Robb K. Does this patient have an abnormal systolic murmur? *JAMA*. 1997;277(7):564-571.

UPDATED LITERATURE SEARCH

Our literature search combined the parent search strategy for The Rational Clinical Examination with the following terms: "systolic and murmur," "heart valve diseases," "aortic valve stenosis," "pulmonary valve stenosis," "mitral valve prolapse," "mitral valve insufficiency," "tricuspid valve insufficiency," "hypertrophic cardiomyopathy," and "heart murmurs." Results were limited to English-language publications in the MEDLINE database from 1996 to July 2004. The titles and abstracts of the search results were screened, case reports were excluded, and 28 potentially relevant primary studies and review articles were retrieved. We scanned the reference list of each article for additional studies. For accuracy studies, we retained those of adult subjects that included sensitivity and specificity data of physical findings and had a quality score of level 3 or greater. We excluded level 3 studies with fewer than 100 patients. Five new studies were ultimately included in this update.¹

NEW FINDINGS

1. Cardiologists are able to distinguish normal ("innocent") murmurs from abnormal murmurs by the physical examination alone.
2. Emergency department physicians are able to detect normal murmurs by clinical evaluation (including physical examination; medical history; ECG, chest radiograph, and laboratory test results; and previously recorded chart data).
3. The presence of a holosystolic murmur, loud murmur, decreased carotid upstroke, or systolic thrill makes it much more likely that a systolic murmur represents an underlying cardiac abnormality rather than a functional murmur.
4. In patients for whom examiners did not know whether a murmur was present before examination, emergency department physicians and cardiologists identified valvular heart disease with good accuracy.
5. Absence of murmur radiation to the right clavicle makes moderate to severe AS much less likely.
6. The presence of any 3 of the following findings makes moderate to severe AS much more likely: maximal murmur intensity in second right intercostal space, reduced carotid pulse volume, slow rate of increase of carotid pulse, and reduced or absent second heart sounds (S_2).
7. When mitral regurgitation (MR) is identified, murmur intensity equal to or more than grade 3 makes severe regurgitation more likely.

IMPROVEMENTS IN DATA PRESENTED IN THE ORIGINAL PUBLICATION

The newer studies do not alter the results reported in the original publication but do provide new information on the role of individual auscultatory findings.

In the original article, the need to identify patients at higher risk for endocarditis because of valvular abnormalities was suggested as a rationale for performing the clinical examination. The recommendations for endocarditis prophylaxis have changed. Patients with murmurs from structural abnormalities of a native valve do not automatically require antibiotic prophylaxis to prevent infective endocarditis.¹

CHANGES IN THE REFERENCE STANDARD

The reference standard is an echocardiogram or a cardiac catheterization that assesses valvular competency.

RESULTS OF THE LITERATURE REVIEW

Precision

Since the original review, 2 published studies involving non-cardiologist examiners have evaluated the precision of various physical examination maneuvers in actual patients.^{2,3} In a large study of medical patients presenting to the emergency department, there was substantial agreement on the presence of systolic murmurs ($\kappa = 0.8$). The precision of examining for a loud murmur ($\kappa = 0.59$) and for an S_2 in the clinical setting is moderate ($\kappa = 0.54$), whereas the precision of other findings is only fair. In both of these studies, the various findings were not evaluated independently, so the examiners' opinions may have been influenced by the presence or absence of related findings.

Accuracy

Distinguishing Abnormal From Normal (Innocent) Murmurs

Two new studies evaluated examiners' ability to distinguish murmurs caused by an underlying cardiac abnormality from those generated by structurally normal hearts (innocent murmurs). One of these studies evaluated the accuracy of the entire clinical evaluation (including physical examination, medical history, echocardiogram, chest radiograph, laboratory tests, and data from old charts) by noncardiologist emergency department physicians, and one evaluated the accuracy of the cardiologist's physical examination alone.⁴

In a study of high methodologic quality, Reichlin et al² evaluated the performance of emergency department physicians' clinical assessments of patients with systolic murmurs. Although these noncardiologists are somewhat less accurate at distinguishing normal from innocent murmurs

than cardiologists, a normal clinical assessment significantly reduces the likelihood of a cardiac abnormality (negative likelihood ratio [LR-], 0.29; 95% confidence interval [CI], 0.17-0.45).

The second study assessed the ability of cardiologists to distinguish innocent from pathologic murmurs by physical examination alone in patients referred for evaluation of a systolic murmur. The cardiologists' overall assessments of significant heart disease (defined as moderate to severe valvular heart disease, congenital shunt, or an intraventricular gradient identified by echocardiography) performed with a positive likelihood ratio (LR+) of 11 (95% CI, 5.0-26) and LR- of 0.22 (95% CI, 0.10-0.41). In addition, several clinical signs were assessed to appraise their performance in categorizing significant systolic murmurs confirmed by echocardiography. The most frequently detected findings, and those that were most useful, are shown in Table 33-10.

Patients with mild AS or regurgitation are not included in the calculation of these LRs. Patients with a loud, plateau-shaped, or holosystolic murmur are more likely to have significant lesions than functional murmurs or mild valvular heart disease. Similarly, the absence of holosystolic or loud murmur suggests that there are no significant lesions. However, an echocardiogram must be obtained when the clinician wants to determine whether the murmur represents moderate to severe AS or regurgitation, a congenital shunt, or an intraventricular pressure gradient.

Identifying Valvular Heart Disease by Physical Examination

The ability to distinguish innocent from pathologic murmurs is important in stratifying patients for referral for echocardiography. However, the ability to make this distinction does not reflect examiners' true ability to determine the presence of valvular heart disease: by excluding patients with no audible murmur, the specificity of the physical examination for valvular disease is underestimated.

In the study by Reichlin et al,² the inclusion criteria required that at least 2 of 3 screening physicians agree that a subject had a murmur: 203 patients were enrolled from 852 screened, whereas 582 were excluded because no systolic murmur was heard. There was excellent agreement among examiners about the presence of a murmur, with disagreement in only 18 patients (2%). The exclusion of those patients with no murmur is an example of verification bias. Verification bias occurs when the gold standard test is not applied to all the potentially eligible patients to confirm their disease status. In this case, patients without systolic murmurs were excluded from the analysis and had no echocardiogram to confirm the absence of structural heart disease. Typically, selective inclusion creates an overestimate of the sensitivity and an underestimate of the specificity of the clinical assessment. However, because the study provides complete information on all patients, we are able to correct for verification bias, with the assumption that patients with no murmur truly had no valvular disease. Recalculation yields an LR+ of 14 (95% CI, 10-19) for a clinical assessment suggesting an abnormal murmur and a LR- of 0.21 (95% CI, 0.13-0.34) when either no murmur was heard or the murmur was deemed normal. Because some patients without systolic murmurs can

Table 33-10 Ability of Findings to Identify Patients With Significant Cardiac Lesions vs Functional Systolic Murmur

Clinical Sign	LR for a Significant Systolic Murmur ^a	
	LR+ (95% CI)	LR- (95% CI)
Holosystolic murmur (n = 26)	8.7 (2.3-33)	0.19 (0.08-0.43)
Loud murmur (n = 29)	6.5 (2.3-19)	0.08 (0.02-0.31)
Plateau-shaped murmur (n = 20)	4.1 (1.4-12)	0.48 (0.30-0.77)
Loudest at the apex (n = 30)	2.5 (0.58-11)	0.84 (0.65-1.1)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

^aThe LR+ is the likelihood ratio when the finding is present and indicates an increased likelihood that the systolic murmur is associated with moderate to severe aortic stenosis or mitral regurgitation, congenital shunt, or intraventricular pressure gradients. The LR- is the likelihood ratio when the finding is absent and shows the likelihood that a significant lesion will be present when the finding is absent.

still have AS or MR, these corrected LRs represent the best possible clinical performance.

Another study using cardiologist examiners addressed the performance of a complete cardiovascular physical examination without additional information in a population of asymptomatic individuals. The patients were not selected because of an auscultated abnormality.⁵ A murmur was heard in 63 patients, with 17 murmurs classified as abnormal; transesophageal echocardiography identified valvular abnormalities in 33 patients. In this population, the cardiovascular physical examination alone performed with an LR+ of 38 (95% CI, 9.5-154) and LR- of 0.31 (95% CI, 0.18-0.52).

Thus, these 2 studies provide information on the clinician's ability to identify valvular heart disease irrespective of the presence of a murmur, which better reflects an initial assessment in clinical practice. Although the populations of patients studied are different and the emergency department assessment includes supplementary information, the examiners' overall performance in these studies is similar. When an abnormal murmur is identified, the pooled LR+ for echocardiographic valvular disease is 15 (95% CI, 11-20; results homogenous with $P = .11$; $I^2 = 48\%$; 95% CI, 0%-86%); when no murmur is heard or the murmur is determined to be "normal," the pooled LR- is 0.25 (95% CI, 0.17-0.36; results homogenous with $P = .29$; $I^2 = 16\%$; 95% CI, 0%-55%).⁶

Aortic Stenosis

One new grade 2 study ($n = 123$),³ performed by noncardiologists, prospectively evaluated individual findings and combinations of findings for the detection of moderate or severe AS (defined as an aortic valve area less than 1.2 cm^2 or peak transvalvular gradient of 25 mm Hg or more). A slow carotid upstroke was the most important individual finding for ruling in AS (LR+, 9.2; 95% CI, 3.4-24) (Table 33-11). The 2-step process for using combinations of findings begins with examination for the presence of a murmur over the right clavicle. If this murmur is absent, AS is considerably less likely (LR-, 0.1; 95% CI, 0.02-0.44). When a murmur radiates to the right clavicle, 4 associated findings are sought: highest intensity of murmur at second right intercostal space, reduced intensity of S_2 , reduced carotid volume, and slow carotid upstroke. When zero to 2 of these associated findings are present, the result is indeterminate (LR, 1.8; 95% CI, 0.93-2.9), whereas if 3 to 4 of these findings are present, the likelihood of AS is significantly increased (LR, 40; 95% CI, 6.6-239).

Mitral Regurgitation

One study evaluated the accuracy of isolated findings in predicting severe MR,⁷ defined as a regurgitant fraction of 40% or more detected by echocardiography (Table 33-12). The clinical findings of interest were abstracted from the patients' personal charts, as recorded by the patients' own physicians (cardiologists and internists), who were unaware of the study. The study evaluated the relationship between the intensity of the murmur and the severity of regurgitation, and demonstrated a significant correlation.

Mitral Valve Prolapse

No new high-quality studies added to the information in the original review. The absence of a murmur and click rules out

Table 33-11 Accuracy of the Physical Examination for Detecting Aortic Stenosis

Finding	LR+ (95% CI)	LR- (95% CI)
Slow carotid upstroke	9.2 (3.4-24)	0.56 (0.32-0.8)
Murmur radiating to right carotid	8.1 (4-16)	0.29 (0.12-0.57)
Reduced or absent S_2	7.5 (3.2-17)	0.50 (0.27-0.76)
Murmur over right clavicle	3.0 (2-4.1)	0.10 (0.02-0.44)
Any systolic murmur	2.6 (1.8-3.5)	0 (0-0.45)
Reduced carotid volume	2.0 (1-3.2)	0.64 (0.34-0.99)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

Table 33-12 Accuracy of the Physical Examination for Detecting Severe Mitral Regurgitation⁷

Finding	LR+ (95% CI)
Murmur grades 4-5	14 (3.3-56)
Murmur grade 3	3.5 (2.1-5.7)
Murmur grades 0-2	0.19 (0.11-0.33)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio.

mitral valve prolapse (MVP) (LR, 0.04). The presence of a nonejection click (a high-pitched sound of short duration in mid or late systole) with or without a murmur slightly increases the likelihood of echocardiographic MVP (LR 3.8).⁸

EVIDENCE FROM GUIDELINES

The American College of Cardiology/American Heart Association guidelines (2003)⁹ recommend echocardiography to evaluate heart murmurs in patients with cardiovascular symptoms or in asymptomatic patients with clinical features that suggest a moderate or greater probability that the murmur is reflective of underlying structural heart disease. Echocardiography is not recommended in asymptomatic adults whose murmur has been identified as functional or innocent by an experienced observer.⁹

CLINICAL SCENARIO—RESOLUTION

Your patient's murmur did not radiate to the right clavicle. This finding makes AS much less likely (LR, 0.1). There are no other concerning features that raise the possibility of other serious structural heart disease, including the ECG and chest radiograph. If you are an experienced clinician, this reduces the likelihood of important structural heart disease (LR, 0-0.1). If you are less experienced and not certain of your overall assessment that the murmur is "functional," concentrating on whether the murmur is holosystolic or "loud" and whether the patient has a decreased carotid upstroke or systolic thrill may yield more useful information than your clinical gestalt. Conditions that can cause increased blood flow through a structurally normal heart should be excluded, such as anemia, renal failure, and thyrotoxicosis.

SYSTOLIC MURMURS—MAKE THE DIAGNOSIS

Systolic murmurs are common, and echocardiography is normal in the majority of asymptomatic individuals with murmurs. Clinical evaluation offers the potential to identify those patients with increased likelihood of underlying structural disease and to avoid costly echocardiographic evaluation in all patients with systolic murmurs.

PRIOR PROBABILITY

One study of randomly selected elderly Finnish persons (aged 75–86 years) found a prevalence of moderate to severe AS of 8.8% in women and 3.6% in men.¹⁰ The prevalence in younger patients ought to be less. The Framingham Heart Study showed that echocardiographic evidence of MR is common and a function of both age and sex.¹¹ A useful approximation for the prevalence of mild to moderate MR is 15% from age 40 to 60 years for both men and women. After age 60, women have a prevalence of about 25% compared with men, who have an increasing frequency of MR that approximates 40% by age 80 years. The prevalence of MVP is about 2.5%.^{12,13}

POPULATION FOR WHOM A SYSTOLIC MURMUR SHOULD BE ASSESSED

- It is sensible to listen for a systolic murmur in every patient for whom a complete cardiac database is necessary.
- Once a patient with a systolic murmur is identified, the clinical examination helps identify those more likely to have significant underlying cardiac lesions. However, a cardiac echocardiogram is required to determine whether the finding represents a significant or less significant cardiac lesion.
- The presence of a murmur can be heard with a variety of underlying lesions such as myocardial ischemia, endocarditis, and disturbances that cause a high flow rate.

IDENTIFYING NORMAL (INNOCENT) MURMURS

Cardiologists and emergency physicians are accurate at distinguishing abnormal from innocent murmurs (Tables 33-13 and 33-14).

Table 33-13 Likelihood Ratio for the Overall Examination for Detecting Valvular Disease

	LR for Valvular Disease	
	LR+ (95% CI)	LR- (95% CI)
Cardiologists ⁵	38 (9.5–154)	0.31 (0.18–0.52)
Emergency department physicians ²	14 (10–19)	0.21 (0.13–0.34)
Summary	15 (11–20)	0.25 (0.17–0.36)

Abbreviations: CI, confidence interval; LR, likelihood ratio; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

Because the overall performance of generalist physicians has not been described, attention to individual findings may be even more useful than the overall clinical impression when a murmur is auscultated.

Table 33-14 Likelihood Ratios of Individual Findings for Identifying Murmurs That Are Significant

Clinical Sign	LR for a Significant Systolic Murmur ^a	
	LR+ (95% CI)	LR- (95% CI)
Systolic thrill (n = 8)	12 (0.76–205)	0.73 (0.58–0.93)
Holosystolic murmur (n = 26)	8.7 (2.3–33)	0.19 (0.08–0.43)
Loud murmur (n = 29)	6.5 (2.3–19)	0.08 (0.02–0.31)
Plateau-shaped murmur (n = 20)	4.1 (1.4–12)	0.48 (0.30–0.77)
Loudest at the apex (n = 30)	2.5 (0.58–11)	0.84 (0.65–1.1)
Radiation to the carotid (n = 9)	0.91 (0.28–3.0)	1.0 (0.78–1.3)

Abbreviations: CI, confidence interval; LR, likelihood ratio; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

^aModerate to severe aortic stenosis or mitral regurgitation, congenital shunt, or intraventricular pressure gradient.

AORTIC STENOSIS

The presence of AS requires detection of a systolic murmur, generally radiating to the right clavicle. For such patients, evaluate the S₂ to determine whether it is reduced in intensity, feel the carotid artery to assess whether the volume is reduced and the upstroke slower than normal, and assess whether the murmur is loudest in the second right intercostal space (Table 33-15).

Table 33-15 Likelihood Ratios of Combinations of Findings for Aortic Stenosis

Clinical Findings ^a	LR (95% CI) for Moderate or Greater Aortic Stenosis
Systolic murmur over right clavicle + 3–4 associated findings	40 (6.6–239)
Systolic murmur over right clavicle + 0–2 associated findings	1.8 (0.93–2.9)
No systolic murmur over right clavicle	0.1 (0.02–0.44)

Abbreviations: CI, confidence interval; LR, likelihood ratio.

^aReduced or absent second heart sound, reduced carotid volume, slow rate of increase of carotid pulse, and maximal murmur intensity in second right intercostal space.

MITRAL REGURGITATION AND MITRAL VALVE PROLAPSE

Although cardiologists are accurate at identifying echocardiographic MR (Table 33-16), the performance of generalist physicians has not been evaluated as well. Once MR is identified, the intensity of the murmur helps to identify the severity of the regurgitation.⁶

Table 33-16 Likelihood Ratio for the Murmur Intensity to Identify Severe Mitral Regurgitation

Finding	LR+ (95% CI)
Murmur grades 4-5	14 (3.3-56)
Murmur grade 3	3.5 (2.1-5.7)
Murmur grade 0-2	0.19 (0.11-0.33)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio.

The absence of a murmur and click rules out MVP (LR, 0.04), whereas the presence of a systolic click, with or without a murmur, slightly increases the likelihood of echocardiographic MVP (LR, 3.8).

REFERENCE STANDARD TEST

Echocardiography or cardiac angiography.

REFERENCES FOR THE UPDATE

- Wilson W, Taubert KA, Gewitz M, et al. Prevention of infective endocarditis: guidelines from the American Heart Association: a guideline from the American Heart Association Rheumatic Fever, Endocarditis, and Kawasaki Disease Committee, Council on Cardiovascular Disease in the Young, and the Council on Clinical Cardiology, Council on Cardiovascular Surgery and Anesthesia, and the Quality of Care and Outcomes Research Interdisciplinary Working Group. *Circulation*. 2007;116(15):1736-1754.
- Reichlin S, Dieterle T, Camli C, Leimenstoll B, Schoenenberger RA, Martina B. Initial clinical evaluation of cardiac systolic murmurs in the ED by noncardiologists. *Am J Emerg Med*. 2004;22(2):71-75.^a
- Etchells E, Glenns V, Shadowitz S, Bell C, Siu S. A bedside clinical prediction rule for detecting moderate or severe aortic stenosis. *J Gen Intern Med*. 1998;13(10):699-704.
- Attenhofer Jost CH, Turina J, Mayer K, et al. Echocardiography in the evaluation of systolic murmurs of unknown cause. *Am J Med*. 2000;108(8):614-620.^a
- Roldan CA, Shively BK, Crawford MH. Value of the cardiovascular physical examination for detecting valvular heart disease in asymptomatic subjects. *Am J Cardiol*. 1996;77(15):1327-1331.^a
- Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. *Stat Med*. 2002;21(11):1539-1558.
- Desjardins VA, Enriquez-Sarano M, Tajik AJ, Bailey KR, Seward JB. Intensity of murmurs correlates with severity of valvular regurgitation. *Am J Med*. 1996;100(2):149-156.^a
- Abbas AS, DeCristofaro D, Anabtawi J, Irwin L. Mitral valve prolapse: comparative value of M-mode, two dimensional and Doppler echocardiography. *J Am Coll Cardiol*. 1983;2(6):1219-1223.
- Cheitlin MD, Armstrong WF, Aurigemma GP, et al. ACC/AHA/ASE 2003 guideline update for the clinical application of echocardiography: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines, p. 8. http://www.acc.org/qualityandscience/clinical/guidelines/echo/index_clean.pdf. Accessed June 4, 2008.
- Lindroos M, Kupari M, Heikkila J, Tilvis R. Prevalence of aortic valve abnormalities in the elderly: an echocardiographic study of a random population sample. *J Am Coll Cardiol*. 1993;21(5):1220-1225.
- Singh JP, Evans JC, Levy D, et al. Prevalence and clinical determinants of mitral, tricuspid, and aortic regurgitation (the Framingham Study). *Am J Cardiol*. 1999;83(6):897-902.
- Freed LA, Levy D, Levine RA, et al. Prevalence and clinical outcome of mitral valve prolapse. *N Engl J Med*. 1999;341(1):1-7.
- Theal M, Sleik K, Anand S, Yi Q, Yusuf S, Lonn E. Prevalence of mitral valve prolapse in ethnic groups. *Can J Cardiol*. 2004;20(5):511-515.

^aFor the Evidence to Support the Update for this topic, see <http://www.JAMAevidence.com>.

This page intentionally left blank

EVIDENCE TO SUPPORT THE UPDATE:

33

Murmur, Systolic

TITLE A Bedside Clinical Prediction Rule for Detecting Moderate or Severe Aortic Stenosis.

AUTHORS Etchells E, Glenns V, Shadowitz S, Bell C, Siu S.

CITATION *J Gen Intern Med.* 1998;13(10):699-704.

QUESTION Can a clinical prediction rule using simple physical examination findings accurately detect aortic stenosis (AS) in a broad spectrum of patients?

DESIGN Consecutive patients were prospectively enrolled when they were referred for echocardiography. Two examiners (a third-year medical resident and a staff general internist) performed the maneuvers on all enrolled patients. An echocardiographer, blinded to the findings, identified all patients with moderate or greater AS.

SETTING General medical/cardiology wards in an urban university hospital in Toronto.

PATIENTS One hundred twenty-three patients admitted to the general medicine and cardiology wards. The majority had some history of congestive heart failure, angina, or myocardial infarction. The median age was 68 years, 58% were men, and 56% had Canadian Cardiovascular Society class I symptoms at the study. Exclusion criteria were age younger than 50 years, cardiac care unit/intensive care unit admission, unstable angina within 48 hours, history of cardiovascular surgery or valve replacement, severe dyspnea at rest, or inability to consent.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

Two examiners, blinded to echocardiographic findings, independently performed a structured physical examination and focused medical history on all enrolled patients. Transthoracic echocardiography was performed on all patients by an echocardiographer blinded to the clinical findings, who identified moderate to severe AS, defined as aortic valve area of 1.2 cm² or smaller or peak transvalvular gradient of 25 mm Hg or higher.

MAIN OUTCOME MEASURES

Sensitivity, specificity, and likelihood ratios; κ for interobserver variability.

MAIN RESULTS

Seventeen patients (14%) were found to have AS, with complete physical examination data available for 15.

CONCLUSION

LEVEL OF EVIDENCE Level 2.

STRENGTHS Prospective data collection with valid reference standard and confirmed independence of clinical examination.

LIMITATIONS This study included only 17 patients with the condition of interest.

Table 33-17 Likelihood Ratios for Findings to Predict Aortic Stenosis

Test	Sensitivity	Specificity	LR+ (95% CI)	LR- (95% CI)
Slow carotid upstroke (n = 12)	0.47	0.95	9.2 (3.4-24)	0.56 (0.32-0.8)
Murmur radiating to right carotid (n = 20)	0.73	0.91	8.1 (4-16)	0.29 (0.12-0.57)
Reduced S ₂ (n = 15)	0.53	0.93	7.5 (3.2-17)	0.50 (0.27-0.76)
Murmur over right clavicle (n = 45)	0.93	0.69	3.0 (2-4.1)	0.10 (0.02-0.44)
Any systolic murmur (n = 52)	1.0	0.64	2.6 (1.8-3.5)	0 (0-0.45)
Reduced carotid volume (n = 35)	0.53	0.73	2.0 (1.0-3.2)	0.64 (0.34-0.99)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

Table 33-18 Combination of Findings for Predicting Aortic Stenosis

	LR (95% CI)
Murmur over clavicle + 3-4 associated findings ^a (n = 7)	40 (6.6-239)
Murmur over clavicle + 0-2 associated findings (n = 38)	1.8 (0.93-2.9)
No murmur over right clavicle (n = 69)	0.1 (0.02-0.44)

Abbreviations: CI, confidence interval; LR, likelihood ratio.

^aAssociated findings include reduced second heart sound (S₂), reduced carotid volume, slow carotid upstroke, and murmur loudest at second right intercostal space.

This study validates several physical examination maneuvers as performed by generalist physicians in a broad spectrum of older general medical inpatients (Table 33-17). These patients are typical of those admitted into hospitals or referred for cardiovascular evaluation. The use of moderate to severe AS as the finding of interest is a clinically significant endpoint. The study confirms that the absence of any murmur or the absence of a murmur over the right clavicle is the best finding for ruling out AS. A reduced carotid upstroke by palpation, a murmur radiating to the right carotid, or S₂ that is reduced in intensity increases the likelihood the most. In contrast to previous studies, a murmur radiating to the right carotid is useful for identifying patients with AS if detected, but AS can still exist without the presence of a murmur radiating to the carotid.

The examiners participating in the study underwent a brief training period (30 minutes) and performed a standardized physical examination. As a result, the performance of the examination might be lower among examiners without the training, although the brief training period could be easily replicated. In addition, because the findings are assessed as part of a standardized physical examination, it is impossible to evaluate their independence. In other words, an examiner who observes that one of the findings is present might be more influenced and likely to describe other abnormal findings.

The authors also created and prospectively evaluated combinations of findings (Table 33-18), which performed with excellent accuracy: a lack of a murmur radiating to the right clavicle effectively rules out AS of moderate or greater severity, whereas the presence of such a murmur in association with 3 or more other findings rules in the diagnosis.

Table 33-19 Reliability of Findings for Aortic Stenosis

Finding	Generalized κ (Lower 95% CI)
S ₂ (normal vs decreased)	0.54 (0.46)
Loud murmur (>II/VI) second RICS	0.45 (0.37)
Radiation to right clavicle	0.36 (0.28)
Radiation to right carotid	0.33 (0.25)
Delayed carotid upstroke	0.26 (0.18)
Reduced carotid volume	0.24 (0.16)
Presence of any systolic murmur	0.19 (0.11)

Abbreviations: CI, confidence interval; RICS, right intercostal space.

The reliability assessment of individual maneuvers is useful and demonstrates that individual findings have reliabilities that vary from slight to moderate (Table 33-19).

Reviewed by David Cescon, MD

TITLE Echocardiography in Evaluating Systolic Murmurs of Unknown Cause.

AUTHORS Attenhofer Jost CH, Turina J, Mayer K, et al.

CITATION Am J Med. 2000;108(8):614-620.

QUESTION How well can cardiologists identify pathologic murmurs by auscultation or palpation alone?

DESIGN Consecutive patients were prospectively identified at referral for evaluation of a systolic murmur of unknown cause. Each subject was independently examined by 2 cardiologists from a pool of 8, blinded to supplementary data and echocardiography results. Two-dimensional (2D)/Doppler echocardiography was performed as the gold standard in all participants. It is not clear whether the ultrasonographers were blinded to the clinical examination.

SETTING Cardiology division in Switzerland.

PATIENTS One hundred patients referred for evaluation of systolic murmur of unknown cause were enrolled. Patients were excluded if they had a previously documented echocardiographic examination. The mean age of subjects was 55 ± 22 years, and 57% were women.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

Full cardiac examination with or without dynamic auscultation as deemed appropriate by 2 blinded cardiologist examiners. Murmurs were classified by Levine grade and described and characterized as functional or organic according to the examiners' clinical expertise. All patients underwent transthoracic 2D/Doppler echocardiography; valvular stenosis and regurgitation were classified according to standard criteria.

MAIN OUTCOME MEASURES

Raw data, sensitivity, specificity.

MAIN RESULTS

Twenty-one patients had a "functional" murmur and were considered normal. Of the 79 patients with "organic" murmurs, 29 patients had aortic stenosis (AS) of various severity and 30 patients had mitral regurgitation (MR). Although the patients were referred for evaluation of systolic murmurs,

Table 33-20 Likelihood Ratios for Overall Assessment of a Valvular Lesion of Any Severity

	LR+ (95% CI)	LR- (95% CI)
Aortic stenosis (n = 33)	2.1 (1.1-3.9)	0.78 (0.61-0.95)
Mitral regurgitation (n = 33)	2.3 (1.5-3.6)	0.43 (0.23-0.71)
Aortic regurgitation (n = 9)	5.1 (1.5-3.9)	0.82 (0.63-0.95)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

echocardiography revealed aortic regurgitation in 28. The data in Table 33-20 indicate the likelihood of the finding when the cardiologists' overall assessment results were positive.

The cardiologists' overall clinical assessments of significant heart disease (defined as moderate to severe valvular heart disease, congenital shunt, or an intraventricular gradient) performed with a positive likelihood ratio (LR+) of 11 (95% confidence interval [CI], 5.0-26) and negative likelihood ratio (LR-) of 0.22 (95% CI, 0.10-0.41). The characteristics of the murmur and response to a few maneuvers were assessed to identify their performance in categorizing significant systolic murmurs confirmed by echocardiography (Table 33-21).

A loud (diagnostic odds ratio, 81) or holosystolic murmur (diagnostic odds ratio, 46) was the most accurate finding for identifying those patients with a significant murmur vs those with a functional murmur.

No patient had a diminished carotid upstroke, so this finding cannot be assessed from the data. A diminished second heart sound (S_2) was assessed, but the finding was heard in 5 patients only. One maneuver, the response to Valsalva, was assessed. Typically, patients with AS or MR would have a decreased intensity with the initiation of the maneuver, whereas patients with hypertrophic cardiomyopathy would have an increase. The maneuver in this study did not help identify patients with significant lesions (LR+, 1.2; 95% CI, 0.66-2.2; and LR-, 0.84; 95% CI, 0.50-1.4), but no patients with hypertrophic cardiomyopathy were found.

Table 33-21 Likelihood Ratio of Signs for a Significant Systolic Murmur

Clinical Sign	LR for a Significant Systolic Murmur vs a Functional Murmur	
	LR+ (95% CI)	LR- (95% CI)
Systolic thrill (n = 8)	12 (0.76-205)	0.73 (0.58-0.93)
Holosystolic murmur (n = 26)	8.7 (2.3-33)	0.19 (0.08-0.43)
Loud murmur (n = 29)	6.5 (2.3-19)	0.08 (0.02-0.31)
Plateau-shaped murmur (n = 20)	4.1 (1.4-12)	0.48 (0.30-0.77)
Loudest at the apex (n = 30)	2.5 (0.58-11)	0.84 (0.65-1.1)
Radiation to the carotid (n = 9)	0.91 (0.28-3.0)	1.0 (0.78-1.3)

Abbreviations: CI, confidence interval; LR, likelihood ratio; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

CONCLUSION

LEVEL OF EVIDENCE Level 3.

STRENGTHS Prospective, consecutive patients.

LIMITATIONS Small referral population referred for evaluation of a murmur. The echocardiographers were not blinded to the clinical findings. The CIs around some of these findings are large.

For the individual clinical signs, we could calculate the LR comparing patients with a significant murmur vs those with a functional murmur. This analysis ignores the patients who had less significant cardiac lesions as the explanation for their systolic murmur (eg, mild AS or MR). Thus, clinicians must understand that although these findings might identify patients more likely to have a significant vs a functional murmur, an echocardiogram must be done to determine whether the findings are associated with a significant or less-significant cardiac lesion.

The results suggest that a cardiologist's examination is useful even when the referring clinician is uncertain that a murmur is innocent. Because these patients are likely the most difficult to examine, the results for the cardiologist might be a "worst-case" scenario for the LRs. We can anticipate that for all patients with systolic murmurs, the LRs would suggest greater accuracy.

The presence of a variety of findings increases the likelihood that a systolic murmur will be significant. Loud, plateau-shaped, holosystolic murmurs with a thrill will have a high likelihood of emanating from significant cardiac abnormalities. These individual findings might work better than the clinician's overall clinical assessment for assessing systolic murmurs for patients in whom the diagnosis might not be readily apparent from the physical examination findings. An important caveat is that this analysis suggests only the presence of a *significant* lesion as defined by the authors as opposed to a functional murmur. Thus, the presence of findings with a high LR+ means that the clinician must request an echocardiogram to determine whether the underlying cardiac lesions are significant or less significant. Similarly, the absence of a loud or holosystolic murmur makes a significant lesion less likely, but an echocardiogram would be required to identify patients with less significant lesions.

The results of this study should be interpreted in light of the clinical population—patients referred for evaluation of systolic murmurs that likely included those for whom the referring clinician was uncertain of the diagnosis. The data in the table do not represent the LRs for a specific diagnosis (eg, AS), but for any significant lesion associated with a systolic murmur.

The response to Valsalva does not help identify significant AS or mitral regurgitant murmurs, but this maneuver could still be important for identifying significant hypertrophic cardiomyopathy.

Reviewed by David Cescon, MD, and Edward Etchells, MD, MSC

TITLE Intensity of Murmurs Correlates With Severity of Valvular Regurgitation.

AUTHORS Desjardins VA, Enriquez-Sarano M, Tajik AJ, Bailey KR, Seward JB.

CITATION *Am J Med.* 1996;100(2):149-156.

QUESTION Does the intensity of regurgitant murmurs on clinical examination correlate with the degree of echocardiographic regurgitation?

DESIGN Investigators prospectively enrolled 210 consecutive patients undergoing Doppler echocardiography who were found to have chronic isolated mitral or aortic regurgitation. Results of a physical examination performed within 2 weeks of echocardiography by the patient's own physician (179 cardiologists, 31 general internists), who was unaware of the study, were abstracted from chart data.

SETTING Echocardiography laboratory in a major US center.

PATIENTS Two hundred ten consecutive patients prospectively identified with chronic, isolated mitral regurgitation (MR) or aortic insufficiency (AI) of mild or greater severity. Exclusion criteria included previous valve repair or replacement, associated valvular stenosis or acute regurgitation, and lack of physical examination performed by the referring physician within 2 weeks of echocardiography. For the 40 patients with isolated AI, the mean age was 58 ± 16 years, 65% were men, 8% were in atrial fibrillation, and the mean regurgitant fraction was $36\% \pm 16\%$. For the 170 patients with MR, the mean age was 64 ± 13 years, 54% were men, 21% were in atrial fibrillation, and the mean regurgitant fraction was $36\% \pm 18\%$ by Doppler echocardiography.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

Quantitative Doppler and 2-dimensional echocardiography were performed on all patients before enrollment. It is not clear whether the echocardiographers were blinded to clinical data. Severe regurgitation was defined as a regurgitant fraction of 40% or higher. The clinical examination documenting murmur severity was performed independently by each patient's personal physician, who was not aware of the study and did not receive any special training or instruction regarding standardization of murmur grading.

MAIN OUTCOME MEASURES

Raw data, correlation coefficients (r). Likelihood ratios were calculated from the data provided.

MAIN RESULTS

The intensity of the murmur predicts the severity of MR (Table 33-22).

Table 33-22 Likelihood Ratios for the Presence of Severe Mitral Regurgitation as a Function of the Murmur Intensity

Murmur Grade	LR (95% CI)
4 Or 5	14 (3.3-56)
3	3.5 (2.1-5.7)
0-2	0.19 (0.11-0.33)

Abbreviations: CI, confidence interval; LR, likelihood ratio.

CONCLUSION

LEVEL OF EVIDENCE Level 2.

STRENGTHS The population included in this study represents a difficult sample because all had some degree of regurgitation. The study examines a relevant clinical question because the ability to correlate the intensity of a regurgitant murmur with the degree of regurgitation is a useful clinical tool.

LIMITATIONS Only patients with isolated lesions were included. The results demonstrate that the evaluation of murmur intensity of isolated MR by internists and cardiologists is a useful diagnostic test: a loud murmur (grade 4 or greater) is a good predictor of severe MR, whereas a murmur of grade 2 or less effectively rules out the presence of severe MR.

This study simulated normal clinical conditions without special training or standardized instructions to the examiner. These results are valid only in chronic, isolated MR and cannot be applied to the acute setting or to patients with complex murmurs.

Reviewed by David Cescon, MD, and Edward Etchells, MD, MSC

TITLE Initial Clinical Evaluation of Cardiac Systolic Murmurs in the Emergency Department by Noncardiologists.

AUTHORS Reichlin S, Dieterle T, Camli C, Leimenstoll B, Schoenenberger RA, Martina B.

CITATION *Am J Emerg Med.* 2004;22(2):71-75.

QUESTION How well do noncardiologists distinguish innocent systolic murmurs from those produced by valvular heart disease in a typical emergency department evaluation?

DESIGN Medical patients presenting to the emergency department were prospectively identified and evaluated for the presence of a systolic murmur. If 2 of 3 physicians, including 1 study physician, agreed on the presence of a murmur, the patient was enrolled in the study.

SETTING Emergency department of a university teaching hospital in Switzerland.

PATIENTS Two hundred three patients were enrolled from 852 medical patients screened in the emergency department. The patients were typical medical patients, with mean age of 64.7 (\pm 22.3) years, and 58% were women. A significant percentage of the enrolled patients, had chest pain at presentation, and the majority had a pathologic electrocardiogram (ECG) (61%) or chest radiograph (53%) in the emergency department.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

The emergency department attending physician's clinical evaluation (including medical history, physical examination, ECG, chest radiograph, and laboratory tests) sought to distinguish normal from abnormal murmurs in all enrolled patients. Transthoracic echocardiography was performed to identify valvular heart disease in all enrolled subjects within 24 hours by 2 cardiologists blinded to the results of the clinical evaluation.

MAIN OUTCOME MEASURES

Sensitivity, specificity, and likelihood ratios (LRs).

MAIN RESULTS

Seventy-one of 203 patients had structural heart disease evident on echocardiography. Twenty-one patients were excluded because there was no informed consent (17) or the echocardiography was not performed (4), leaving 582 patients with no systolic murmur. Of the entire sample size, there was disagreement for only 18 patients, for whom a third examiner settled the discordance.

The κ statistic for the presence of a murmur was 0.8; the κ statistic for murmur grades 0 to 2 vs those greater than grade 2 was 0.59.

CONCLUSION

LEVEL OF EVIDENCE Level 1.

STRENGTHS Prospective, consecutive patients with independent application of the reference standard in a population typical for those in whom distinguishing a normal from an abnormal systolic murmur by clinical examination is an important clinical question. Because the patients provided information on all potentially eligible patients, we can correct for verification bias.

LIMITATIONS Entrance criteria required that 2 of 3 examiners agree that a murmur was present. Although this may decrease generalizability, it improves our confidence that a murmur was present.

This large, high-quality study evaluated the utility of the clinical evaluation by noncardiologists. The examiners in this study had access to all available clinical information, including patient charts that documented previously identified valvular heart disease in 10% of patients; however, this represents a realistic clinical scenario.

The level of agreement among examiners in identifying the presence of a systolic murmur of intensity greater than grade II/VI documented in this study compares favorably to that of previous studies involving cardiologists examining patients.

This study provides complete information on all patients, allowing us to correct for verification bias by making certain assumptions about the patients for whom both clinicians did not hear a murmur or for whom there was a disagreement about the presence of a murmur. The majority of patients who did not undergo echocardiography did not have a systolic murmur, as judged by 2 examiners. If we assume that none of these patients truly had valvular heart disease, the LRs are as shown in Table 33-23. These LRs estimate the efficiency of the

Table 33-23 Likelihood Ratio of the Overall Examination for an Abnormal Murmur

Clinical Evaluation	Patients	Sensitivity	Specificity	LR+ (95% CI)	LR- (95% CI)
Overall examination suggests abnormal murmur, corrected for verification bias	All patients			14 (10-19)	0.21 (0.13-0.34)
Overall examination, uncorrected for verification bias	Only patients with systolic murmurs	0.80	0.69	2.6 (2.0-3.4)	0.29 (0.17-0.45)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

clinicians to identifying aortic or mitral valvular disease among all patients. Because most patients do not have valvular heart disease, the specificity of the examination is excellent.

The LRs reported by the investigators, uncorrected for verification bias, show the performance of the clinical examination among patients known to have a systolic murmur. In clinical practice, these patients would be more reflective of those referred for echocardiography to determine the presence of a systolic murmur.

Reviewed by David Cescon, MD, and Edward Etchells, MD, MSC

TITLE Value of the Cardiovascular Physical Examination for Detecting Valvular Heart Disease in Asymptomatic Subjects.

AUTHORS Roldan CA, Shively BK, Crawford MH.

CITATION *Am J Cardiol.* 1996;77(15):1327-1331.

QUESTION How useful is the physical examination in detecting the presence or absence of valvular heart disease in asymptomatic individuals?

DESIGN Nonconsecutive patients were prospectively identified for inclusion and were examined by a cardiologist blinded to other data. An echocardiographer, blinded to clinical findings, identified valvular heart disease.

SETTING Outpatient clinic in the United States.

PATIENTS The population consisted of 75 patients with connective tissue diseases and 68 healthy volunteers. The patients with connective tissue diseases had systemic disease without cardiac symptoms and constituted a group of patients for whom most physicians would auscultate the heart to detect asymptomatic cardiac disease associated with their underlying disorder (systemic lupus erythematosus, ankylosing spondylitis, rheumatoid arthritis, antiphospholipid antibody syndrome).

The mean age of participants was 38 ± 11 years, 56 were men, and none had cardiovascular symptoms. Only 5% of subjects were known to have murmur or valvular heart disease.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

Subjects were randomly sequenced for a complete physical examination, including dynamic auscultation by a cardiologist blinded to other data. The cardiologist recorded the findings for jugular venous pulse; the palpated carotid pulse; the palpated precordial maximal impulse; the presence of a right ventricular lift; abnormalities of the second, third, and fourth heart sounds; clicks; and ejection sounds. The dynamic auscultation included

evaluation of murmur change with respiration, Valsalva maneuver, handgrip, and changes in body position.

Transesophageal echocardiography was performed on all subjects by an echocardiographer blinded to the clinical examination and other data. Diagnosis of valvular disease was based on established criteria.

MAIN OUTCOME MEASURES

Sensitivity, specificity.

MAIN RESULTS

Thirty-three patients had echocardiographic evidence of valvular abnormalities, the majority (24 of 33) of which were mitral valve regurgitant lesions or prolapse. The predictive value of the individual findings is reported, but none occurred in more than 8% of patients, providing broad confidence intervals. It is difficult to disentangle the individual findings from the overall assessment because the individual components and categorization of individual murmurs were based on the total evaluation (Table 33-24).

CONCLUSION

LEVEL OF EVIDENCE Level 3.

STRENGTHS Prospective, blinding of examination, and gold standard test.

LIMITATIONS Cardiologist examiner may limit generalizability to generalist physicians. Nonconsecutive patients.

The study population is unique in that these patients were not selected according to an auscultated abnormality. They represent a combination of healthy patients and patients with noncardiac disease, all of whom might undergo auscultation in the course of “routine” medical care. By including healthy patients, a high specificity for the examination could be expected because most patients would not have abnormal findings and would not have cardiac abnormalities shown by echocardiogram.

This study evaluated physical examination by a cardiologist alone, without supplementary information or investiga-

Table 33-24 Likelihood Ratio for the Overall Clinical Examination to Identify Patients With Abnormal Cardiac Valves

Test	Valvular Heart Disease by Echocardiography			
	Sensitivity	Specificity	LR+ (95% CI)	LR- (95% CI)
Overall clinical assessment for a valvular abnormality	0.70	0.98	38 (9.5-154)	0.31 (0.18-0.52)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

tions in a healthy population at risk for valvular heart disease. It is useful that the report includes the actual individual components used by the cardiologists to determine their overall clinical assessment. The cardiologists heard a surprising number of murmurs, but when they described a murmur

as abnormal, the likelihood of an echocardiograph abnormality increased greatly.

Reviewed by David Cescon, MD, and Edward Etchells, MD, MSC

This page intentionally left blank

CHAPTER 34

Does This Patient Have Myasthenia Gravis?

Katalin Scherer, MD

Richard S. Bedlack, MD, PhD

David L. Simel, MD, MHS

CLINICAL SCENARIOS

CASE 1 A 45-year-old man has a 2-month history of fluctuating double vision, a droopy right eye that improves with rest, and a complaint that food gets stuck halfway down. Your examination confirms severe right eyelid ptosis that dramatically improves with rest. His right eye adduction and up gaze are markedly impaired. The left eye demonstrates complete horizontal ophthalmoplegia. The limb muscle strength and reflexes are normal. You wonder whether there is an accurate and clinically useful bedside test to help confirm the diagnosis of myasthenia gravis.

CASE 2 A 69-year-old man has a 2-month history of intermittent spells of double vision, generalized weakness that worsens toward the evening, and unspecified dizziness. Although he has normal strength and reflexes and no ophthalmoplegia, he does report fluctuating diplopia during the examination. As in case 1, you must decide whether to perform additional bedside tests, obtain electrodiagnostic or acetylcholine antibody testing, or pursue a broader diagnostic evaluation of the various causes of dizzy spells and fatigue.

WHY IS THIS AN IMPORTANT QUESTION TO ANSWER WITH A CLINICAL EXAMINATION?

Myasthenia gravis is an autoimmune disease associated with circulating acetylcholine receptor antibodies, modification of the synaptic cleft, and destruction of the postsynaptic neuromuscular membrane. The clinical hallmark of the disease is fatigable weakness. The clinical severity ranges from mild, purely ocular, forms to severe generalized weakness and respiratory failure. Myasthenia gravis is a rare disease; its prevalence in the United States is reported at 14.2 in 100000. Prevalence rates have been increasing steadily during the past decades, likely because of decreased mortality, longer survival, and higher rates of diagnosis.¹⁻³ Men older than 50 years have the highest incidence in the population, with the peak at approximately aged 70 years. Women have 2 incidence peaks: one at approximately aged 20 to 40 years and one at approximately aged 70 years.^{4,5}

Clinicians must be alert to the symptoms and signs of myasthenia gravis because it is an eminently treatable disease, and the earlier treatment is started, the better the clinical response.⁶⁻⁸ Only 54% to 69% of patients with myasthenia gravis are diagnosed within 1 year of onset, and the mean time to diagnosis is more than 1 year.^{3,9,12} Untreated patients are at risk for deterioration and “crisis,” which occurs when weakness becomes severe enough to require mechanical ventilation.^{13,14} Left untreated, reversible and fatigable weakness may become fixed. An erroneous diagnosis of myasthenia gravis may expose patients to unnecessary diagnostic procedures and treatments.

The acetylcholine receptor antibody test is the most specific diagnostic test for myasthenia gravis. This test has reasonable sensitivity in generalized myasthenia gravis (80%-96%), but up to 50% of patients with purely ocular myasthenia have seronegative test results.¹⁵⁻¹⁹ Single-fiber electromyography, performed by highly trained experts at specialized centers, is highly sensitive for disorders of the neuromuscular junction but is not specific for myasthenia gravis.

The purpose of this review was to determine the value of clinical symptoms and signs, as well as the results of simple provocative clinical tests, in deciding whether myasthenia gravis should be considered as a diagnosis and in enabling the physician to determine whether further confirmatory testing (including the highly specific and sensitive antibody test) is warranted.

Anatomic and Physiologic Origins of the Symptoms and Signs Used to Answer This Question

In the normal neuromuscular junction, acetylcholine is released into the synaptic cleft, diffuses to the postsynaptic membrane, binds to ligand-sensitive ion channels (nicotinic acetylcholine receptors), and causes an excitatory postsynaptic end-plate potential. If the threshold depolarization is achieved, an action potential will spread along the muscle fiber membrane, causing muscle contraction. Acetylcholine is cleared from the synaptic cleft by presynaptic reuptake and by the metabolic action of acetylcholinesterase (Figure 34-1).

The failure of transmission at many neuromuscular junctions in myasthenia results in diminished end-plate potentials that are insufficient to generate action potentials in a number of muscle fibers.²⁰ This results in fatigable weakness of striated

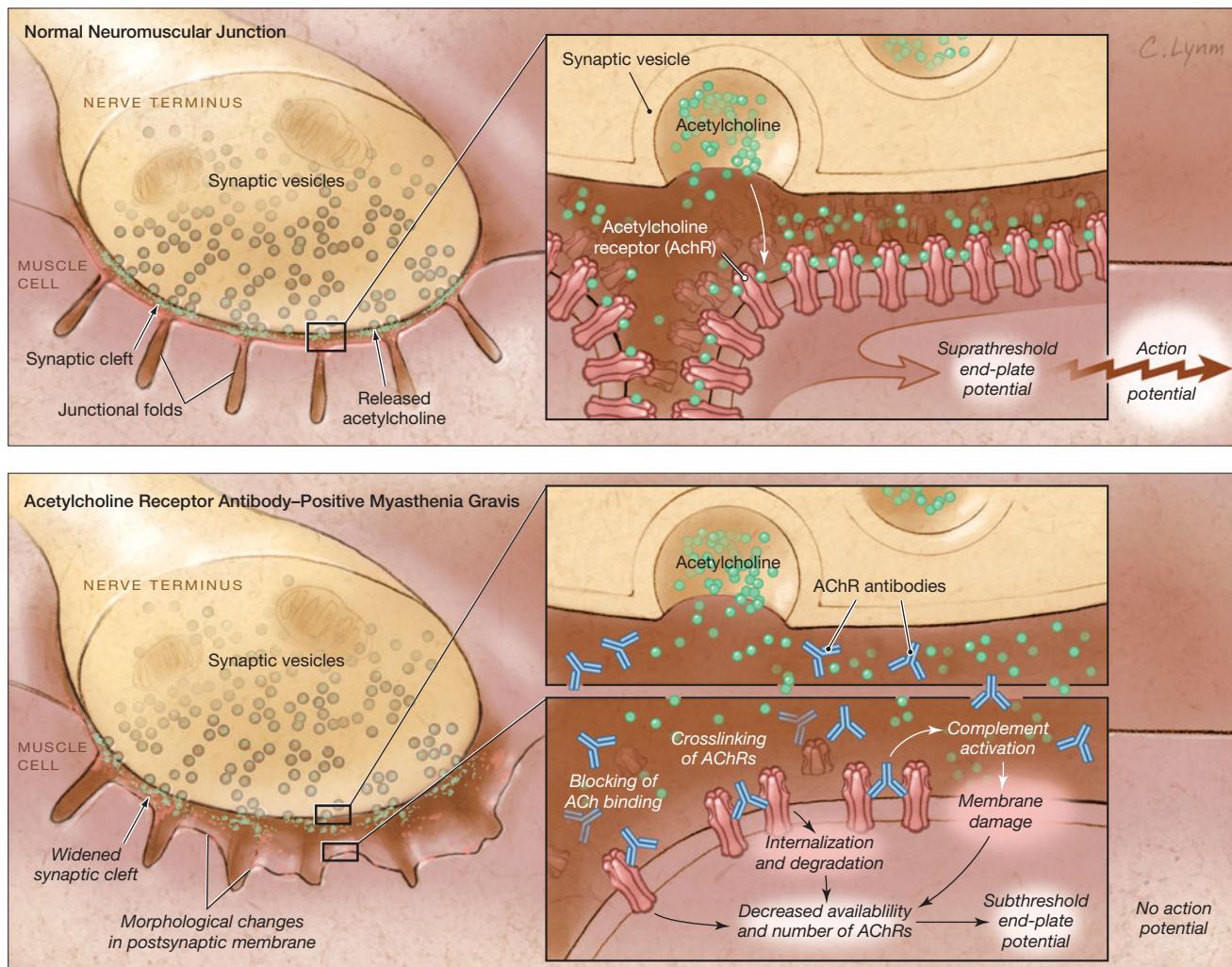


Figure 34-1 Neuromuscular Junction

In patients with acetylcholine receptor (AChR) antibody-positive myasthenia gravis, circulating antibodies bind to the AChRs, which may block acetylcholine binding, lead to cross-linking of receptors promoting internalization and degradation, and induce postsynaptic membrane damage via complement activation. The number and availability of receptors are reduced such that end-plate potentials are insufficient to generate action potentials in a number of muscle fibers, causing weakness.

muscles, which is the basis for the clinical diagnosis. Sustained or repetitive muscle contraction causes fatigue and weakness of myasthenic muscles. Cooling a weak muscle improves neuromuscular transmission.²¹ Rest and acetylcholinesterase inhibitors transiently increase acetylcholine levels in the synaptic cleft. The change in strength after these manipulations can be assessed during the clinical examination.

Symptoms and Signs and How to Elicit Them

Patients with myasthenia gravis complain of weakness in specific muscles. Up to 65% of patients initially have ocular symptoms of double vision and drooping of the eyelids. Less than one-fourth of patients present with bulbar weakness (ie, in lower cranial nerve–innervated oropharyngeal muscles) and report slurred or nasal speech, alterations of the voice (eg, softness, breathiness, hoarseness), and difficulty chewing or swallowing. Limb weakness is an unusual initial complaint (14%-27%) and should be differentiated from nonspecific generalized fatigue. Patients may report shortness of breath. The symptoms of myasthenia are typically better on awakening or after rest and become progressively worse with prolonged use of the affected muscles or later in the day.^{3,22-24}

Reduced muscle power by manual testing in specific muscles that worsens with repetition and improves with rest is the characteristic examination finding in myasthenia. Most muscles with voluntary activation have a large variability of strength even under normal conditions because of effort. Evaluating extremity strength greatly depends on the experience of the examiner. Ptosis and extraocular muscle deficits are relatively free of a voluntary component and provide a more objective measure.

Fatigable and rapidly fluctuating asymmetric ptosis is a hallmark of myasthenia gravis. The rapid fluctuation results from improvement during even very short periods of rest, such as blinking. Besides fast variability in the degree of ptosis, it may altogether shift quickly from one eye to the other, known as “shifting ptosis.”²² Ptosis should be evaluated with the patient sitting comfortably, the head held in primary position without tilting. The patient fixates on a distant object (eg, a spot on the wall) and is asked to refrain from blinking and to relax the forehead muscles. Frontalis contraction, a mostly involuntary compensatory mechanism, is a common and characteristic sign in myasthenic patients with ptosis. Relaxing the forehead muscles may be difficult for some patients. The examiner measures palpebral fissure width at eye level during forward gaze and again during prolonged upward or lateral gaze for 30 seconds.^{22,25} The more ptotic eyelid should be used for additional provocative tests, such as the ice pack, rest, and sleep tests.

The ice pack test is performed by placing a latex glove finger filled with crushed ice over the more ptotic eyelid for 2 minutes. During the rest test, the patient places a glove filled with cotton (a placebo) over the more ptotic eyelid while holding the eyes closed for 2 minutes. During the sleep test, the patient is left in a quiet dark room with the eyes closed for 30 minutes. Complete or almost complete resolution of ptosis or at least a 2-mm increase in palpebral fissure width

constitutes a positive response to these maneuvers. It is important to evaluate the improvement immediately after the tests because the lids may quickly begin to droop again.

The curtain sign (also known as “enhanced ptosis” or “paradoxical ptosis”) is usually observed in patients with some initial ptosis. The patient looks straight ahead and refrains from blinking. The examiner holds one eye open, which results in the other lid starting to droop more (like a curtain falling). The lid twitch sign occurs when the patient opens the eyes after gentle closure or follows the examiner’s finger down and then back up to eye level. The lids overshoot or twitch for a fraction of a second before settling into position and starting to droop.²⁶

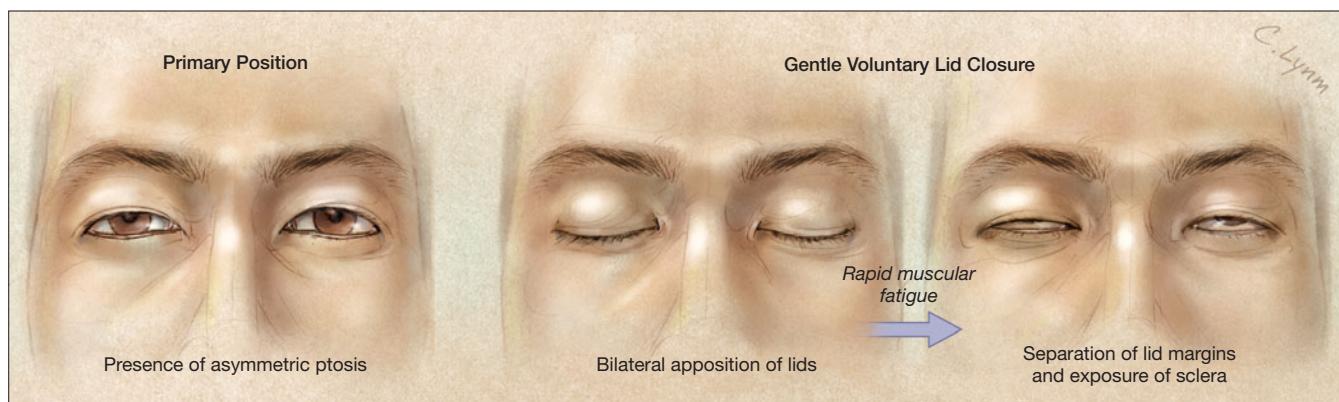
Asymmetric weakness of extraocular muscles is commonly observed in myasthenia when sustained lateral gaze or up gaze worsens or induces double vision. The cover-uncover test may be performed to bring out subtle extraocular weakness. As the patient fixates on an object in the distance, the examiner covers one eye while observing for deviation of the uncovered eye during lateral and then upward gazing. With extraocular weakness, the uncovered eye will drift. The examination is completed by repeating the procedure for the opposite eye. Quiver eye movements are fast, small-twitch, “lightning-like” or “jerk-like” movements of the eyes on changing direction of gaze. They are said to occur even in the setting of profound ophthalmoplegia.²⁷

Although patients rarely complain of facial weakness, it is often found on examination. Severe facial weakness results in a characteristic transverse smile. Orbicularis oculi weakness is demonstrated as the examiner tries to separate the eyelids against forced eye closure. Orbicularis oculi fatigue may be observed on gentle eye closure. After complete initial apposition of the lid margins, they separate within seconds and the white of the sclera starts to show (positive peek sign) (Figure 34-2).²⁸ The iris should not be visible because of the eyeballs being rolled up (Bell phenomenon). The iris may be visible if the patient is not trying to close the eyes voluntarily (in the case of a conversion reaction and functional weakness) or in case of severe ophthalmoplegia.

Tongue and pharyngeal weakness will result in the patient’s speech becoming slurred or nasal, especially with prolonged speaking. Other commonly weak muscles include neck flexors, deltoids, hip flexors, finger/wrist extensors, and foot dorsiflexors. The muscles should be repeatedly tested against manual resistance, with a brief rest between repetitions. Having the patient hold the head above the pillow in the supine position and having the patient hold the arms outstretched in abduction at the shoulder for 1 minute are ways to test for fatigability of neck flexors and deltoids, respectively. Involvement is often asymmetric. The remainder of the neurologic examination results, including those for deep tendon reflexes and sensory examination, must be normal.

Anticholinesterase Tests

Edrophonium chloride is a fast- and short-acting acetylcholinesterase inhibitor that may be administered in the office setting to diagnose myasthenia gravis (Box 34-1). Its effect

**Figure 34-2** Peek Sign

Orbicularis oculi weakness may be indicated by a positive peek sign after gentle eyelid closure. After complete initial apposition of the lid margins, they quickly (within 30 seconds) start to separate, and the sclera starts to show (ie, a positive peek sign). The presence of a peek sign increases the likelihood of myasthenia gravis (likelihood ratio, 30; 95% confidence interval, 3.2–278), but absence of the peek sign does not rule it out.

Box 34-1 Edrophonium Test

Establish reliable peripheral intravenous access.

Prepare a syringe with 2 mg of atropine (available in ampoules of 0.4 or 1 mg/mL) as a precaution.

Prepare 1 mL (10 mg) of edrophonium in a tuberculin syringe (edrophonium is available in a 10 mg/mL solution in a 1-mL ampoule [10 mg] or in a 10-mL vial [total of 100 mg]).

Inject 2 mg (0.2 mL) slowly for 15 seconds while observing for an objective improvement in target muscles.

Improvement should occur within 30 seconds and disappear in 5 minutes; if there is no response or no significant adverse effects, administer the remaining edrophonium (8 mg [0.8 mL]), for a total dose of 10 mg.

Atropine should be injected (0.5 or 1 mg) in case of clinically significant bradycardia, respiratory distress, or syncope.^a

^aRoutine administration of atropine simultaneously with edrophonium for the purpose of diagnostic testing for myasthenia gravis is not recommended. Bartley and Bullock²⁹ recommend using a 3-way stopcock, with the edrophonium-containing syringe attached to the direct port and the atropine-containing syringe attached to the side port so that atropine may be quickly injected in case of severe adverse effects.

tive contraindications. Using a 3-way stopcock setup may be feasible in a patient already equipped with a peripheral intravenous line (eg, in an intensive care unit). One concern with such a setup is the possibility of an accidental mix-up of the syringes, with resultant injection errors—the syringes should always be labeled clearly. Because of the short action of the drug, the examiner must be able to quickly assess for improvement. Evaluating extraocular muscle abnormalities or changes in manual muscle testing requires skill and time; therefore, most experts recommend performing the edrophonium test only when the patient has easily observed baseline weakness in specific muscles.³² Some authors suggest that a clearly ptotic eyelid or visibly abnormal extraocular muscles are the only acceptable findings to observe for objective endpoints.²⁷ Unequivocal improvement in ptosis or extraocular muscles constitutes a positive response. The administering physician (especially one with less experience) should consider blinding the edrophonium administration to avoid expectation bias.

Neostigmine bromide is an anticholinesterase agent used to treat myasthenia gravis. Parenteral preparations are available in vials containing 0.25-, 0.5-, and 1-mg/mL doses. The recommended dose for the diagnosis of myasthenia is 0.02 mg/kg given intramuscularly. A standard dose of 1 or 1.5 mg may be used. The response should be evaluated 30 minutes after injection, at peak effect. The half-life after intramuscular administration is 50 to 90 minutes. Adverse effects, precautions, and need for good intravenous access (to administer atropine in case of an adverse event) are the same as for edrophonium.

Pyridostigmine bromide is an analog of neostigmine, with a slightly longer duration of action and fewer adverse effects. It is the most commonly used anticholinesterase agent for the symptomatic treatment of myasthenia gravis. It has been used for diagnosis in patients in whom edrophonium or neostigmine is relatively contraindicated, although it is not generally used for diagnostic purposes.³³ It is available for injection in 2-mL vials containing 5 mg/mL. A 2-mg intra-

usually occurs within 30 seconds and lasts less than 5 minutes. Most myasthenic muscles respond to the test dose of 2 mg, but many will require more. Adverse effects are rare and usually mild (excess salivation, sweating, abdominal cramps, or fecal incontinence). Serious adverse effects, such as bradycardia, asystole, and bronchoconstriction, occur infrequently (<0.2%) but warrant that the patient receive cardiac monitoring during the test and that a bag-mask be available should the patient need ventilatory assistance.^{30,31} Reactive airway disease or cardiac bradyarrhythmias are rela-

muscular or intravenous dose is equivalent to 60 mg orally. Precautions should be exercised just as with edrophonium and neostigmine.

METHODS

Search Strategy and Quality Review

English-language articles in the MEDLINE database from January 1966 through January 2005 were searched using the terms “myasthenia gravis,” “diagnosis,” and “test.” One of the authors (K.S.) identified potential articles by screening the retrieved titles and abstracts (when available) and searching through the bibliographies of the retrieved articles. Two authors (K.S. and R.S.B.) independently reviewed the retrieved articles. An article

was included when agreement existed that the study had met our inclusion criteria.

Eligible studies evaluated a particular symptom or sign in patients with myasthenia gravis and in controls. Studies requiring sophisticated equipment or subspecialty trained physicians (otolaryngology, ophthalmology, etc) were excluded. Studies based on small numbers of patients were not excluded, because most series are comparatively small in the literature. Of 640 total articles, the search identified 33 potential articles. Of these, 15 met inclusion criteria and form the basis of this review.^{28,33-46} Quality of evidence in each study was classified according to a published classification scheme for levels of evidence developed for The Rational Clinical Examination series (Table 34-1).⁴⁷ Only 2 studies included an independent blinded comparison of signs and symptoms to a criterion standard.^{34,36}

Table 34-1 Characteristics of Studies That Include Patients With Myasthenia Gravis, as Well as Controls

Source, y	Enrollment	Patient Selection	Patients With Myasthenia Gravis, No./Overall (%)	Diagnostic Criteria for Myasthenia Gravis	Symptom or Sign Studied (Inclusion Criteria)	Enrollment Site
Evidence Level 2 ^a						
Kubis et al, ³⁴ 2000	Prospective	Consecutive	10/25 (40)	AChRAb or SFEMG	Ice test, rest test (ptosis)	Neuro-ophthalmology clinic
Evidence Level 3 ^a						
Ertas et al, ³⁶ 1994	Prospective	Unclear	12/27 (44)	Overall clinical impression	Ice test, edrophonium, or neostigmine test (ptosis)	Neurology clinic
Czaplinski et al, ³⁵ 2003	Prospective	Unclear	5/10 (50)	AChRAb and RNS	Ice test, edrophonium test (ptosis)	Neurology clinic
Evidence Level 4 ^a						
Sethi et al, ⁴⁰ 1987	Unclear	Unclear	10/17 (59)	Overall clinical impression	Ice test, edrophonium test (ptosis)	Neurology clinic
Odel et al, ⁴¹ 1991	Unclear	Unclear	42/68 (62)	Edrophonium test	Sleep test (ptosis or ophthalmoplegia)	Ophthalmology clinic
Golnik et al, ³⁹ 1999	Prospective	Unclear	20/40 (50)	AChRAb or edrophonium test	Ice test (ptosis)	Neuro-ophthalmology clinic
Ellis et al, ³⁷ 2000	Prospective	Consecutive	15/30 (50)	Overall clinical impression	Ice test (ptosis or abnormal extraocular movements)	Ophthalmology clinic
Lertchavanakul et al, ³⁸ 2001	Prospective	Unclear	20/40 (50)	EMG or neostigmine test	Ice test (ptosis)	Ophthalmology clinic
Evidence Level 5 ^a						
Osserman and Kaplan, ⁴⁴ 1952	Prospective	Unclear	15/50 (30)	Overall clinical impression	Edrophonium test	Neurology clinic, hospital
Yee et al, ⁴⁵ 1976	Prospective	Unclear	10/18 (56)	Edrophonium or neostigmine test	Quiver eye movements (ophthalmoplegia)	Ophthalmology clinic
Osher and Griggs, ²⁸ 1979	Prospective	Consecutive	25/275 (9)	Unclear	Peek sign (orbicularis oculi fatigue)	Ophthalmology clinic
Nicholson et al, ⁴³ 1983	Prospective	Consecutive	46/75 (61)	Overall clinical impression with 1 positive test result	Edrophonium test	AChRAb laboratory
Batocchi et al, ⁴² 1997	Prospective	Consecutive	39/72 (54)	Overall clinical impression with 2 positive test results	Edrophonium test (ptosis, ophthalmoplegia)	Ophthalmology clinic
Padua et al, ³³ 2000	Prospective	Consecutive	29/69 (42)	AChRAb + SFEMG or RNS + AChE	Edrophonium or pyridostigmine test	Neurology clinic
Weijnen et al, ⁴⁶ 2000	Unclear	Unclear	60/80 (75)	Overall clinical impression	Food in mouth after swallowing, unintelligible speech after prolonged speaking	Oromaxillofacial surgery clinic

Abbreviations: AChE, acetylcholine esterase inhibitor; AChRAb, acetylcholine receptor antibody; EMG, electromyography; RNS, repetitive nerve stimulation; SFEMG, single-fiber electromyography.

^aSee Table 1-7 for a description of Evidence Levels.

Statistical Methods

Sensitivity was defined as the proportion of patients with myasthenia gravis who had the particular symptom or sign; specificity, as the proportion of nonmyasthenic patients without the sign or symptom. The positive likelihood ratio (LR_+) was defined as the likelihood of a positive test result (or presence of a sign or symptom) in a myasthenic patient compared with the likelihood of a positive test result in a nonmyasthenic patient, that is, the increase in odds that the patient has myasthenia gravis when the test result is positive (or when the sign or symptom is present). LR_+ is expressed as sensitivity/(1 – specificity). The negative likelihood ratio

(LR_-) is the likelihood of a negative test result (or absence of a sign or symptom) in a myasthenic patient compared with the likelihood of a negative test result (or absence of a sign or symptom) in a nonmyasthenic patient, that is, the decrease in odds that the patient has myasthenia gravis when the test result is negative (or when the sign or symptom is absent). LR_- is expressed as (1 – sensitivity)/specificity. Summary LRs were derived using random-effects measures that provide conservative confidence intervals (CIs) around the estimates.⁴⁸⁻⁵⁰

RESULTS

Table 34-2 Clinical Signs and Symptoms and Results of Clinical Tests in the Prediction of Myasthenia Gravis

Source, y	LR (95% CI)	
	Positive	Negative
Symptoms		
Food in mouth after swallowing		
Weijnen et al, ⁴⁶ 2000	13 (0.85-212)	0.70 (0.58-0.84)
Unintelligible speech after prolonged speaking		
Weijnen et al, ⁴⁶ 2000	4.5 (1.2-17)	0.61 (0.46-0.80)
Signs		
Peek sign		
Osher and Griggs, ²⁸ 1979	30 (3.2-278)	0.88 (0.76-1.0)
Quiver eye movements		
Yee et al, ⁴⁵ 1976	4.1 (0.22-75)	0.82 (0.57-1.2)
Simple Office Tests		
Ice test		
Kubis et al, ³⁴ 2000	28 (1.8-427)	0.14 (0.03-0.62)
Ertas et al, ³⁶ 1994	31 (2.0-472)	0.04 (0-0.61)
Czaplinski et al, ³⁵ 2003	11 (0.77-158)	0.09 (0.01-1.3)
Sethi et al, ⁴⁰ 1987	12 (0.83-185)	0.24 (0.08-0.72)
Golnik et al, ³⁹ 1999	33 (2.1-515)	0.22 (0.10-0.50)
Ellis et al, ³⁷ 2000	31 (2.0-475)	0.03 (0-0.46)
Lertchavanakul et al, ³⁸ 2001	39 (2.5-605)	0.07 (0.01-0.33)
Summary	24 (8.5-67)	0.16 (0.09-0.27)
Anticholinesterase test		
Ertas et al, ³⁶ 1994	28.0 (1.8-436)	0.12 (0.03-0.54)
Czaplinski et al, ³⁵ 2003	9.0 (0.61-133)	0.27 (0.07-1.1)
Sethi et al, ⁴⁰ 1987	12 (0.83-185)	0.24 (0.08-0.72)
Osserman and Kaplan, ⁴⁴ 1952	70 (4.4-1096)	0.03 (0-0.46)
Nicholson et al, ⁴³ 1983	54 (3.5-850)	0.10 (0.04-0.24)
Batocchi et al, ⁴² 1997	67 (4.3-1053)	0.01 (0-0.16)
Padua et al, ³³ 2000	9.7 (3.8-25)	0.04 (0.01-0.28)
Summary	15 (7.5-31)	0.11 (0.06-0.21)
Rest test		
Kubis et al, ³⁴ 2000	16 (0.98-261)	0.52 (0.29-0.95)
Sleep test		
Odel et al ⁴¹	53 (3.4-832)	0.01 (0-0.16)

Abbreviations: CI, confidence interval; LR, likelihood ratio.

Accuracy of Symptoms for the Diagnosis of Myasthenia Gravis

Only 1 eligible study was identified, and it evaluated 2 symptoms.⁴⁶ The history was taken from patients via a questionnaire. The presence of food remaining in the mouth after swallowing increases the likelihood of myasthenia gravis, but the wide CI indicates that the finding is not reliable. Speech becoming unintelligible during prolonged speaking has an LR of 4.5 (95% CI, 1.2-17). Neither normal swallowing nor normal speech rules out myasthenia gravis (LR, 0.70; 95% CI, 0.58-0.84; and LR, 0.61; 95% CI, 0.46-0.80, respectively).

Accuracy of Signs for the Diagnosis of Myasthenia Gravis

Two eligible studies were identified and reported on 1 sign each.^{28,45} The presence or absence of quiver eye movements increased the likelihood of myasthenia gravis, but the broad CIs around the LR indicate that the examiner may not rely on the finding. The presence of the peek sign might be more useful (LR, 30; 95% CI, 3.2-278) but also has broad CIs.

Accuracy of Simple Office Tests for the Diagnosis of Myasthenia Gravis

Seven studies investigated the ice test, and all had similar findings.³⁴⁻⁴⁰ The overall prevalence (prior probability) of myasthenia gravis in these studies was 49% (92 of 189 patients total). All but 1 of these studies were carried out prospectively. The LR for a positive ice test result suggests that the finding is useful (summary LR, 24; 95% CI, 8.5-67). A negative ice test result lessens

the likelihood of myasthenia gravis (summary LR, 0.16; 95% CI, 0.09-0.27).

Two studies evaluated the precision (ie, interobserver variation) of the ice test. Kubis et al³⁴ used the signed rank test to evaluate interobserver variability and found no significant difference between observers ($P = .79$). Ertas et al³⁶ reported complete agreement among their observers. Neither of the studies evaluated the intraobserver variation.

Seven studies reported the results of anticholinesterase tests, and all had similar findings.^{33,35,36,40,42-44} Five of these studies evaluated the edrophonium test; one study included response to pyridostigmine, and another included response to neostigmine as an alternative. All but 1 of these studies were prospective, and 3 were carried out on consecutive patients. One hundred fifty-six (49%) of 320 patients had myasthenia gravis. The likelihood of myasthenia gravis increases for a positive test result (summary LR, 15; 95% CI, 7.5-31), whereas the lack of improvement makes myasthenia gravis much less likely (summary LR, 0.11; 95% CI, 0.06-0.21).

Two studies evaluated the sleep or rest test on 93 patients, including 52 (56%) with myasthenia gravis.^{34,41} An abnormal rest test result increases the likelihood of myasthenia, but the wide CI indicates uncertainty about the true significance. A positive sleep test result may be more useful (LR, 53; 95% CI, 3.4-832). Both the rest and sleep test make the probability of myasthenia unlikely when the result is normal (LR, 0.52; 95% CI, 0.29-0.95; and LR, 0.01; 95% CI, 0-0.16, respectively).

Are These Symptoms or Signs Ever Normal?

Fluctuating weakness (ie, reduced muscle power) that worsens with exertion and improves with rest or with application of ice or cold is never normal. It is important to differentiate fluctuating weakness from patients' reports of weakness, which most often refers to fatigue or exertion. True fluctuating weakness, as demonstrated by manual muscle testing, is the cardinal feature of myasthenia gravis. Other neuromuscular conditions (including amyotrophic lateral sclerosis and periodic paralyses) may be associated with fluctuating weakness; however, the fluctuation in myasthenia is more dramatic and occurs much more rapidly. Ptosis or diplopia may be present in a number of conditions (congenital exotropia or esotropia, strabismus, congenital ptosis, cranial nerve palsies, myopathies, progressive external ophthalmoplegia, brainstem lesions, and neurodegenerative disorders such as progressive supranuclear palsy), but the constant degree of involvement and associated neurologic findings (pupillary abnormalities, nystagmus, vertigo, sensory involvement) commonly exclude myasthenia gravis as a diagnosis. One must bear in mind that even the initially fluctuating weakness of myasthenia gravis will become fixed over time if severe enough. The hypomimia (masked facies) of parkinsonism may be mistaken for facial weakness, but on examination, no true weakness is found and associated features of parkinsonism are evident. It may be a challenge to differentiate true fatigable weakness caused by myasthenia gravis from conversion reactions. In the latter conditions, one may often find that various elements of the examination are inconsistent with

pathophysiologic conditions. Conversion reactions commonly produce giveaway weakness, in which an initial full resistance suddenly gives way under the hand of the examiner, as opposed to true weakness that gradually worsens or is present from the start. Ptosis produced by conversion reactions is commonly symmetrical and bilateral. Because it occurs with contraction of the orbicularis oculi, one can observe that the lower lid elevates. It may completely disappear with diverting the patient's attention. Eye closure weakness caused by poor effort results in the iris showing between the eyelids.

CLINICAL SCENARIOS—RESOLUTIONS

CASE 1 Fluctuating diplopia and ptosis are highly characteristic of myasthenia gravis. The presence of a positive rest test result may increase the likelihood of myasthenia. The physician must carefully question the patient regarding his complaint of food getting stuck halfway down. If it is food remaining in the mouth after swallowing, it may also increase the probability of myasthenia. The available evidence-based data, however, do not allow the examiner to rely on these findings to confirm the diagnosis. These positive test results should prompt the clinician to confirm the diagnosis with the acetylcholine receptor antibody test and to refer this patient to a specialist (neurologist or neuro-ophthalmologist).

CASE 2 The presentation of an elderly patient complaining of fluctuating double vision and weakness worsening toward the end of the day raises the possibility of myasthenia gravis. The lack of quiver eye movements, peek sign, or history of unintelligible speech after prolonged speaking or food in the mouth after swallowing does not significantly reduce the likelihood of myasthenia according to the studies we reviewed. This patient does not have any objective ptosis or visible diplopia, so provocative tests cannot be performed. A search should be undertaken for causes of nonspecific dizziness and generalized fatigue. If, however, he continues to complain of fluctuating double vision, he should be referred for specialist evaluation to rule out myasthenia despite normal physical examination findings.

THE BOTTOM LINE

The presence of certain historical features (speech becoming unintelligible after prolonged periods) or signs (peek sign) may be useful in confirming the diagnosis of myasthenia gravis, although their absence does not rule it out. The ice test, the sleep test, and response to anticholinesterase agents (especially the edrophonium test) are useful in confirming the diagnosis and reduce the likelihood when results are negative. A positive test result should prompt proceeding with acetylcholine receptor antibody testing and specialist referral for electrophysiologic tests and should help confirm the diagnosis in patients who have negative results for the acetylcholine receptor antibody panel.

This review has several limitations, and the results should be interpreted with caution. The results may not be general-

izable for a number of reasons. Myasthenia gravis is a rare disorder, and the number of studies evaluating its symptoms and signs are few. The studies included in this review examined only a few symptoms and signs in a selected group of patients with a confirmed diagnosis of myasthenia gravis. Because of possible verification bias in this selected population of patients with myasthenia (in whom confirmation of the diagnosis is more likely with clear-cut cases), it is expected that in the general population these tests have a lower sensitivity but even higher specificity. Because of the uncertainty regarding sensitivity, patients with persistent symptoms but normal physical examination findings should be referred to specialists for diagnosis. The specificity and sensitivity of the described tests may also greatly depend on the skill and experience of the examiner. Future studies are needed that evaluate not only intraobserver variability but agreement between experts and nonexperts. There are other signs that may be more useful than those tested historically and that await scientific study. This review underscores the need for more studies to evaluate symptoms and signs predictive of myasthenia to improve physicians' ability to recognize and evaluate patients at presentation.

Author Affiliations at the Time of the Original Publication

Department of Medicine, Duke University Medical Center, Durham, North Carolina (Dr Scherer); Departments of Medicine, Durham Veterans Affairs Medical Center and Duke University, Durham, North Carolina (Drs Bedlack and Simel).

Acknowledgments

We thank Donald B. Sanders, MD, Jonathan A. Edlow, MD, and John R. Lynch, MD, for critical review of the manuscript.

REFERENCES

- Phillips LH II. The epidemiology of myasthenia gravis. *Neurol Clin*. 1994;12(2):263-271.
- Phillips LH II, Torner JC. Epidemiologic evidence for a changing natural history of myasthenia gravis. *Neurology*. 1996;47(5):1233-1238.
- Giagheddu M, Puggioni G, Sanna G, et al. Epidemiological study of myasthenia gravis in Sardinia, Italy (1958-1986). *Acta Neurol Scand*. 1989;79(4):326-333.
- Sanders DB, Andrews PI, Howard JF, Massey JM. Seronegative myasthenia gravis. *Neurology*. 1997;48(suppl 5):S40-S45.
- Phillips LH II, Torner JC, Anderson MS, Cox GM. The epidemiology of myasthenia gravis in central and western Virginia. *Neurology*. 1992;42(10):1888-1893.
- Massey JM. Acquired myasthenia gravis. *Neurol Clin*. 1997;15(3):577-595.
- Bedlack RS, Sanders DB. Steroid treatment for myasthenia gravis: steroids have an important role. *Muscle Nerve*. 2002;25(1):117-121.
- Pascuzzi RM, Coslett HB, Johns TR. Long-term corticosteroid treatment of myasthenia gravis: report of 116 patients. *Ann Neurol*. 1984;15(3):291-298.
- Kalb B, Matell G, Pirskanen R, Lambe M. Epidemiology of myasthenia gravis: a population-based study in Stockholm, Sweden. *Neuroepidemiology*. 2002;21(5):221-225.
- Beghi E, Antozzi C, Batocchi AP, et al. Prognosis of myasthenia gravis: a multicenter follow-up study of 844 patients. *J Neurol Sci*. 1991;106(2):213-220.
- Weizer JS, Lee AG, Coats DK. Myasthenia gravis with ocular involvement in older patients. *Can J Ophthalmol*. 2001;36(1):26-33.
- Somnier FE, Trojaborg W. Neurophysiological evaluation in myasthenia gravis: a comprehensive study of a complete patient population. *Electroencephalogr Clin Neurophysiol*. 1993;89(2):73-87.
- Osserman KE, Kornfeld P, Cohen E, et al. Studies in myasthenia gravis: review of two hundred eighty-two cases at the Mount Sinai Hospital, New York City. *Arch Intern Med*. 1958;102(1):72-81.
- Simpson JE, Westerberg MR, Magee KR. Myasthenia gravis: an analysis of 295 cases. *Acta Neurol Scand*. 1966;42(suppl 23):1-27.
- Romi F, Skeie GO, Aarli JA, Gilhus NE. Muscle autoantibodies in subgroups of myasthenia gravis patients. *J Neurol*. 2000;247(5):369-375.
- Oh SJ, Kim DE, Kuruoglu R, Bradley RJ, Dwyer D. Diagnostic sensitivity of the laboratory tests in myasthenia gravis. *Muscle Nerve*. 1992;15(6):720-724.
- Howard FM Jr, Lennon VA, Finley J, Matsumoto J, Elveback LR. Clinical correlations of antibodies that bind, block, or modulate human acetylcholine receptors in myasthenia gravis. *Ann NY Acad Sci*. 1987;505:526-538.
- Garlepp MJ, Kay PH, Dawkins RL. The diagnostic significance of autoantibodies to the acetylcholine receptor. *J Neuroimmunol*. 1982;3(4):337-350.
- Lindstrom JM, Seybold ME, Lennon VA, Whittingham S, Duane DD. Antibody to acetylcholine receptor in myasthenia gravis: prevalence, clinical correlates, and diagnostic value. *Neurology*. 1976;26(11):1054-1059.
- Drachman DB. Myasthenia gravis. *N Engl J Med*. 1994;330(25):1797-1810.
- Borenstein S, Desmedt JE. Local cooling in myasthenia: improvement of neuromuscular failure. *Arch Neurol*. 1975;32(3):152-157.
- Oosterhuis HJ. The ocular signs and symptoms of myasthenia gravis. *Doc Ophthalmol*. 1982;52(3-4):363-378.
- Oosterhuis HJ. Studies in myasthenia gravis, I: a clinical study of 180 patients. *J Neurol Sci*. 1964;38:512-546.
- Sneddon J. Myasthenia gravis: the difficult diagnosis. *Br J Psychiatry*. 1980;136:92-93.
- Oosterhuis HJ. Acquired blepharoptosis. *Clin Neurol Neurosurg*. 1996;98(1):1-7.
- Cogan DG. Myasthenia gravis: a review of the disease and a description of lid twitch as a characteristic sign. *Arch Ophthalmol*. 1965;74:217-221.
- Cogan DG, Yee RD, Gittinger J. Rapid eye movements in myasthenia gravis, I: clinical observations. *Arch Ophthalmol*. 1976;94(7):1083-1085.
- Osher RH, Griggs RC. Orbicularis fatigue: the "peek" sign of myasthenia gravis. *Arch Ophthalmol*. 1979;97(4):677-679.
- Bartley GB, Bullock JD. The Tensilon test using a three-way stopcock and double syringe. *Ophthal Plast Reconstr Surg*. 1986;2(2):95-96.
- Ing EB, Ing SY, Ing T, Ramocki JA. The complication rate of edrophonium testing for suspected myasthenia gravis. *Can J Ophthalmol*. 2000;35(3):141-144.
- Van Dyk HJ, Florence L. The Tensilon test: a safe office procedure. *Ophthalmology*. 1980;87(3):210-212.
- Seybold ME. The office Tensilon test for ocular myasthenia gravis. *Arch Neurol*. 1986;43(8):842-843.
- Padua L, Stalberg E, LoMonaco M, Evoli A, Batocchi A, Tonali P. SFEMG in ocular myasthenia gravis diagnosis. *Clin Neurophysiol*. 2000;111(7):1203-1207.
- Kubis KC, Danesh-Meyer HV, Savino PJ, Sergott RC. The ice test versus the rest test in myasthenia gravis. *Ophthalmology*. 2000;107(11):1995-1998.
- Czaplinski A, Steck AJ, Fuhr P. Ice pack test for myasthenia gravis: a simple, noninvasive and safe diagnostic method. *J Neurol*. 2003;250(7):883-884.
- Ertas M, Arac N, Kumral K, Tuncbay T. Ice test as a simple diagnostic aid for myasthenia gravis. *Acta Neurol Scand*. 1994;89(3):227-229.
- Ellis FD, Hoyt CS, Ellis FJ, Jeffery AR, Sondhi N. Extraocular muscle responses to orbital cooling (ice test) for ocular myasthenia gravis diagnosis. *J AAPOS*. 2000;4(5):271-281.
- Lertchavanakul A, Gannerdsiri P, Hirunwiwatkul P. Ice test for ocular myasthenia gravis. *J Med Assoc Thai*. 2001;84(suppl 1):S131-S136.
- Golnik KC, Pena R, Lee AG, Eggenberger ER. An ice test for the diagnosis of myasthenia gravis. *Ophthalmology*. 1999;106(7):1282-1286.
- Sethi KD, Rivner MH, Swift TR. Ice pack test for myasthenia gravis. *Neurology*. 1987;37(8):1383-1385.
- Odell JG, Winterkorn JM, Behrens MM. The sleep test for myasthenia gravis: a safe alternative to Tensilon. *J Clin Neuroophthalmol*. 1991;11(4):288-292.
- Batocchi AP, Evoli A, Majolini L, et al. Ocular palsies in the absence of other neurological or ocular symptoms: analysis of 105 cases. *J Neurol*. 1997;244(10):639-645.
- Nicholson GA, McLeod JG, Griffiths LR. Comparison of diagnostic tests in myasthenia gravis. *Clin Exp Neurol*. 1983;19:45-49.

44. Osserman KE, Kaplan LI. Rapid diagnostic test for myasthenia gravis: increased muscle strength, without fasciculations, after intravenous administration of edrophonium (Tensilon) chloride. *JAMA*. 1952;150(4):265-268.
45. Yee RD, Cogan DG, Zee DS, Baloh RW, Honrubia V. Rapid eye movements in myasthenia gravis, II: electro-oculographic analysis. *Arch Ophthalmol*. 1976;94(9):1465-1472.
46. Weijnen FG, Kuks JB, van der Bilt A, van der Glas HW, Wassenberg MW, Bosman F. Tongue force in patients with myasthenia gravis. *Acta Neurol Scand*. 2000;102(5):303-308.
47. Holleman DR Jr, Simel DL. Does the clinical examination predict airflow limitation? *JAMA*. 1995;273(4):313-319.
48. Eddy DM, Hasselblad V, Shachter RD. *Meta-analysis by the Confidence Profile Method: The Statistical Synthesis of Evidence*. San Diego, CA: Academic Press; 1992.
49. Borenstein M, Rothstein H. *Comprehensive Meta-analysis v1.0.23* [computer program]. Englewood Cliffs, NJ: Biostat; 1999.
50. Jaeschke RZ, Meade M, Guatt G, Keenan S, Cook D. How to use diagnostic test articles in the intensive care unit: diagnosing weanability using f/Vt. *Crit Care Med*. 1997;25(9):1514-1521.

This page intentionally left blank

Prepared by Katalin Scherer, MD

Reviewed by Richard Bedlack, MD

Original Review

Scherer K, Bedlack RS, Simel DL. Does this patient have myasthenia gravis? *JAMA*. 2005;293(15):1906-1914.

The Update was prepared within 12 months of The Rational Clinical Examination article publication, so the “Make the Diagnosis” section summarizes findings published in the original review.

CLINICAL SCENARIOS

Case 1

A 45-year-old man has a 2-month history of fluctuating double vision, a droopy right eye that improves with rest, and a complaint that food gets stuck halfway down. Your examination confirms severe right eyelid ptosis that dramatically improves with rest. His right eye adduction and up gaze are markedly impaired. The left eye demonstrates complete horizontal ophthalmoplegia. The limb muscle strength and reflexes are normal. You wonder whether there is an accurate and clinically useful bedside test to help confirm the diagnosis of myasthenia gravis.

Case 2

A 69-year-old man has a 2-month history of intermittent spells of double vision, generalized weakness that worsens toward the evening, and unspecified dizziness. Although he has normal strength and reflexes and no ophthalmoplegia, he does report fluctuating diplopia during the examination. As in case 1, you must decide whether to perform additional bedside tests, obtain electrodiagnostic or acetylcholine antibody testing, or pursue a broader diagnostic evaluation of the various causes of dizzy spells and fatigue.

CLINICAL SCENARIOS—RESOLUTION

Case 1

Fluctuating diplopia and ptosis are highly characteristic of myasthenia gravis. The presence of a positive rest test result may increase the likelihood of myasthenia. The physician must carefully question the patient regarding his complaint of food getting stuck halfway down. If it is food remaining in the mouth after swallowing, it may also increase the probability of myasthenia. The available evidence-based data, however, do not allow the examiner to rely on these findings to confirm the diagnosis. These positive test results should prompt the clinician to confirm the diagnosis with the acetylcholine receptor antibody test and to refer this patient to a specialist (neurologist or neuro-ophthalmologist).

Case 2

The presentation of an elderly patient complaining of fluctuating double vision and weakness worsening toward the end of the day raises the possibility of myasthenia gravis. The lack of quiver eye movements, peek sign, or history of unintelligible speech after prolonged speaking or food in the mouth after swallowing does not significantly reduce the likelihood of myasthenia according to the studies we reviewed. This patient does not have any objective ptosis or visible diplopia, so provocative tests cannot be performed. A search should be undertaken for causes of nonspecific dizziness and generalized fatigue. If, however, he continues to complain of fluctuating double vision, he should be referred for specialist evaluation to rule out myasthenia despite normal physical examination findings.

MYASTHENIA GRAVIS—MAKE THE DIAGNOSIS

PRIOR PROBABILITY

Myasthenia gravis is a rare disease. The prevalence in the United States is reported at approximately 14.1 in 100000.¹⁻³ Men older than 50 years have the highest incidence, with the peak at approximately aged 70 years. Women have 2 incidence peaks: one at approximately aged 20 to 40 years and one at approximately aged 70 years.^{4,5} The prior probability of myasthenia gravis in the general population among patients presenting with symptoms is unknown. Because of the high prevalence of the disease in the included studies (close to 50%), the results may not be generalizable to the general population.

POPULATION FOR WHOM MYASTHENIA GRAVIS COULD BE CONSIDERED

- Patients with asymmetric fluctuating eyelid ptosis
- Patients with extraocular dysmotility not referable to a single nerve
- Patients with weakness of other specific muscles
- Young women of childbearing age and men and women aged approximately 70 years

DETECTING THE LIKELIHOOD OF MYASTHENIA GRAVIS

The clinical findings, when applied to the correct patient population, are important (Table 34-3).

Table 34-3 Detecting the Likelihood of Myasthenia Gravis

	LR (95% CI)
Makes the Diagnosis More Likely	
The presence of an abnormal sleep test result in a patient with symptoms	53 (3.4-832)
The presence of the peek sign in a patient with symptoms	30 (3.2-278)
The presence of an abnormal ice test result in a patient with symptoms	24 (8.5-67)
The presence of a positive response to an anticholinesterase test in a patient with symptoms	15 (7.5-31)
The presence of the history “speech becoming unintelligible during prolonged speaking” in a patient with symptoms	4.5 (1.2-17)
Reduces the Likelihood of Myasthenia	
The presence of a normal rest test result in a patient with symptoms	0.01 (0-0.16)
The absence of a positive response to an anticholinesterase test in a patient with symptoms	0.11 (0.06-0.21)
The presence of a normal ice test result in a patient with symptoms	0.16 (0.09-0.27)
The presence of a normal sleep test result in a patient with symptoms	0.52 (0.29-0.95)

Abbreviations: CI, confidence interval; LR, likelihood ratio.

REFERENCE STANDARD TESTS

The reference standard for definite myasthenia gravis is typical clinical presentation plus one of the following: elevated acetylcholine receptor antibody level or abnormal electrodiagnostic study results (repetitive nerve stimulation or single-fiber electromyography). These criteria should also be fulfilled in clinical practice for definite diagnosis.

REFERENCES FOR THE UPDATE

1. Phillips LH II. The epidemiology of myasthenia gravis. *Neurol Clin*. 1994;12(2):263-271.
2. Phillips LH II, Torner JC. Epidemiologic evidence for a changing natural history of myasthenia gravis. *Neurology*. 1996;47(5):1233-1238.
3. Giagheddu M, Puggioni G, Sanna G, et al. Epidemiological study of myasthenia gravis in Sardinia, Italy (1958-1986). *Acta Neurol Scand*. 1989;79(4):326-333.
4. Sanders DB, Andrews PI, Howard JE, Massey JM. Seronegative myasthenia gravis. *Neurology*. 1997;48(suppl 5):S40-S45.
5. Phillips LH II, Torner JC, Anderson MS, Cox GM. The epidemiology of myasthenia gravis in central and western Virginia. *Neurology*. 1992; 42(10):1888-1893.

CHAPTER 35

Is This Patient Having a Myocardial Infarction?

Akbar A. Panju, MBChB, FRCPC

Brenda R. Hemmelgarn, PhD, MD

Gordon H. Guyatt, MD, MSc, FRCPC

David L. Simel, MD, MHS

CLINICAL SCENARIOS

Are These Patients Having Myocardial Infarctions?

CASE 1 A 57-year-old man presents to the emergency department with squeezing retrosternal pain that started 1 hour ago. He is diaphoretic. His blood pressure is 110/70 mm Hg, his heart rate is 74/min, and he has an audible fourth heart sound. The electrocardiogram (ECG) reveals a 2-mm ST-segment elevation in leads V₁ to V₄.

CASE 2 A 70-year-old man, with a myocardial infarction (MI) 5 years previously, presents to the emergency department with severe tightness in the neck. The discomfort started 30 minutes ago and was associated with diaphoresis. His blood pressure is 90/60 mm Hg, his heart rate is 50/min, and the ECG reveals Q waves in V₁ to V₄ (present in the old ECG).

CASE 3 A 50-year-old woman presents to the emergency department with retrosternal burning of 1 hour's duration and nausea. Antacids provided no relief. The findings of the clinical examination were unremarkable. The ECG reveals 3-mm ST-segment elevation in leads II, III, and aVF and 1-mm ST-segment depression in leads I and aVL.

CASE 4 A 40-year-old woman presents to the emergency department with a 24-hour history of left-sided chest pain. The pain is worsened by exertion and movement. Her medical history is unremarkable. The examination reveals normal vital signs and tenderness with palpation of the left lower costal cartilages. An ECG result is normal.

WHY IS THIS AN IMPORTANT QUESTION TO ANSWER WITH A CLINICAL EXAMINATION?

There have been numerous technologic advancements made in the assessment of patients with symptoms suggestive of acute MI. These include evaluation of time-dependent changes in enzyme levels and biomarkers, as well as an assessment of wall-motion abnormality using echocardiography, radionuclide angiography, or nuclear imaging.

Despite this progress, a carefully conducted history-taking and physical examination remain the first components—and the cornerstones—in the initial assessment of patients presenting with suspected MI. The medical history and physical examination are critical in guiding the selection of further diagnostic and therapeutic interventions. Clinicians complement their clinical examination with a 12-lead ECG and biomarkers, which are additional data that provide the most definitive diagnosis of MI. We will focus on features of medical history, physical examination, and ECG that aid in increasing or decreasing the likelihood of acute MI. We include the ECG in our review because the clinician often

interprets the results at the patient's bedside as part of a prompt initial clinical evaluation.

For the purpose of clarification, we begin by describing the 3 diagnostic groupings of patients with acute chest pain currently used by clinicians and then contrast these with the categorization of chest pain as presence or absence of MI, as is evident in the literature. We then briefly describe signs and symptoms of MI, mechanisms of chest pain, and conditions that may present with symptoms suggestive of MI. After these introductory topics, a detailed account of the precision and accuracy of the medical history, physical examination, and ECG in the diagnosis of MI is provided.

DEFINITIONS

Cardiac ischemic chest pain presents in a spectrum of conditions, including angina, unstable angina, and MI. *Angina* is defined as a discomfort in the chest or adjacent areas, caused by myocardial ischemia, usually brought on by exertion, and associated with a disturbance of myocardial function, but without myocardial necrosis.¹ Various grading systems of the severity of angina pectoris have been developed. The classification proposed by the Canadian Cardiovascular Society,² outlined in [Table 35-1](#), is a practical one adopted in a variety of settings.

Unstable angina encompasses a spectrum of symptomatic manifestations of ischemic heart disease intermediate between stable angina and acute MI. According to historical features, ECG findings (with and without pain), and hemodynamic changes (low blood pressure, third heart sound, mitral regurgitation, and pulmonary crackles), guidelines have been developed to stratify patients with suspected unstable angina into high, intermediate, or low risk of complications after initial evaluation.³ These guidelines also recommend disposition based on initial assessment of risk.

The diagnosis of MI used in most studies is based on criteria proposed by the World Health Organization (WHO). In an attempt to standardize the diagnosis of acute MI, the

WHO requires evolutionary changes on serially obtained ECG tracings or an increase and decrease in biomarker levels, either with typical ischemic-type chest discomfort and an ECG result that was not normal or with an ECG progression labeled probable and associated with lesser symptoms.⁴

Diagnosis in Acute Chest Pain

Determining the correct diagnosis is imperative to administering the appropriate therapy. The available therapeutic options create the categories for patients presenting to the emergency department with chest pain or other symptoms suggesting cardiac ischemia. Three distinct management strategies determine the diagnostic groupings clinicians use currently ([Figure 35-1](#)).

For the first group of patients, which includes those with MI and ST-segment elevation or left bundle-branch block (LBBB) ([Figure 35-1](#), group A), current therapy consists of early percutaneous coronary interventions or thrombolytic therapy. A second group of patients includes those with MI but without ST-segment elevation or LBBB, or those with high-risk unstable angina ([Figure 35-1](#), group B). These patients require intensive monitoring, immediate administration of antiplatelet agents, and possibly antithrombotic therapy. The third group includes patients with low-risk unstable angina or nonischemic chest pain ([Figure 35-1](#), group C). Clinicians may consider either admitting these patients to an intermediate care setting or ward bed or discharging them home with plans for subsequent diagnostic testing to establish the cause of their symptoms. Economic pressures on the health care system have highlighted the importance of distinguishing the second from the third group of patients.

Ideally, we should have information that allows us to classify patients into one of these 3 groups. This is not, however, the issue addressed by most studies of the medical history and physical examination in the setting of acute chest pain. Rather, as shown in [Figure 35-2](#), studies typically classify patients with acute chest pain into 2 groups according to the presence (group 1) or absence (group 2) of MI. Specifically, all patients with MI ([Figure 35-1](#), groups A and B) are compared with all those without MI ([Figure 35-1](#), group C).

The results of studies that used the [Figure 35-2](#) design may mislead clinicians who need to discriminate among the 3 groups of patients as shown in [Figure 35-1](#). Clinical features that fail to distinguish patients with infarct or high-risk unstable angina from those with low-risk unstable angina or nonischemic chest pain might still be useful in the decision about whether to admit to a monitored bed in an acute care hospital. The study design that most investigators have chosen, depicted in [Figure 35-2](#), does not correlate with the current triage of chest-pain patients according to the therapeutic options available. Current therapeutic interventions for MI require the presence of ECG changes. Categorizing patients as in [Figure 35-2](#) will, however, provide clinically important information when we have interventions that are clearly useful in acute MI both with and without ECG changes. Our review will aid the reader in identifying features of the medical history, physical examination, and ECG that help differentiate acute MI

Table 35-1 Grading of Angina of Effort by the Canadian Cardiovascular Society

Grade	Description
I	"Ordinary physical activity does not cause angina," such as walking or climbing stairs. Angina with strenuous or rapid or prolonged exertion at work or recreation.
II	"Slight limitation of ordinary activity." Walking or climbing stairs rapidly, walking uphill, or walking or stair climbing after meals, in cold, in wind, or under emotional stress, or only during the few hours after awakening. Walking more than 2 blocks on the level and climbing more than 1 flight of ordinary stairs at a normal pace and in normal conditions.
III	"Marked limitation of ordinary physical activity." Walking 1 or 2 blocks on a level surface and climbing 1 flight of stairs in normal conditions and at a normal pace.
IV	"Inability to carry on any physical activity without discomfort—angina syndrome may be present at rest."

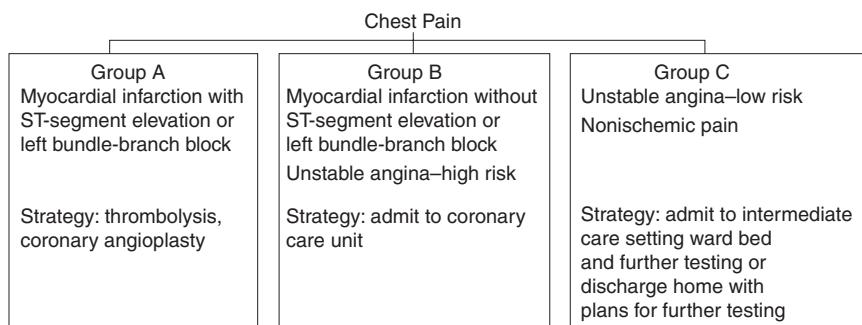


Figure 35-1 Diagnostic Groupings of Acute Chest Pain Based on Management Strategies

patients, both with and without ECG changes, from non-MI patients. Clinicians must avoid misinterpreting the diagnostic information we will present as if it were useful in differentiating among the 3 groups in Figure 35-1.

Relevant Signs and Symptoms

Patients with acute MI typically present with a characteristic combination of signs and symptoms, as outlined in standard textbooks of medicine. Pain is described as being the most common presenting complaint, and considerable emphasis is placed on the characteristics of the pain, including its location, duration, radiation, and quality. Location of the pain includes the central portion of the chest or epigastrium, with potential radiation to the arms, neck, jaw, or less commonly to the abdomen and back. Quality of the chest pain is characteristically described with adjectives such as squeezing, crushing, and pressure.

Other symptoms also may be present, including diaphoresis, nausea, vomiting, weakness, and syncope. Although certain features have been identified as being important in recognizing MI, follow-up data from the Framingham study cohort estimate that approximately 25% of infarcts may go unrecognized because of either lack of chest pain or atypical symptoms.⁵

Mechanism of Chest Pain in Myocardial Infarction

Three-fourths of all patients with recognized acute MI present with chest pain.⁶ Cardiac ischemic pain originates in the myocardium, where free nerve endings are the sensory receptors. Cardiac afferent impulses travel through fibers in the cardiac sympathetic nerves, the upper 5 sympathetic ganglia, the white

rami communicants, the gray rami, and then via the upper 4 or 5 thoracic roots. Cardiac afferent impulses project to the dorsal horn convergent neurons, travel via the spinothalamic tract to the thalamus, and subsequently to the cortex, where the cardiac stimuli are decoded.

Afferent impulses also travel in the cholinergic fibers of the vagus nerve, many of which arise from the inferior cardiac wall. The signs and symptoms of nausea, bradycardia, and hypotension, which appear to be more prevalent in patients with inferior wall MI, are believed to be related to the larger number of vagal afferent fibers located in the inferior cardiac wall.⁷

Like other visceral sensations, myocardial pain is poorly and variably localized. In addition, sensations originating in other intrathoracic structures (particularly the esophagus) can cause pain that is indistinguishable from cardiac pain.

Conditions That May Present With Symptoms Suggestive of Myocardial Infarction

There are many other clinical conditions that can present with symptoms suggestive of acute MI, which can be broadly divided into cardiac and noncardiac disorders. The noncardiac causes of chest pain are further divided into gastroesophageal diseases and nongastroesophageal diseases, whereas the cardiac causes are grouped into ischemic and nonischemic conditions. Figure 35-3 illustrates the most common of these conditions but is not all inclusive.

Given the diversity of the conditions presenting with chest pain, and the extent of the diagnostic testing that would be required, it is difficult to determine the relative frequency of each of these conditions occurring in the setting of chest pain.

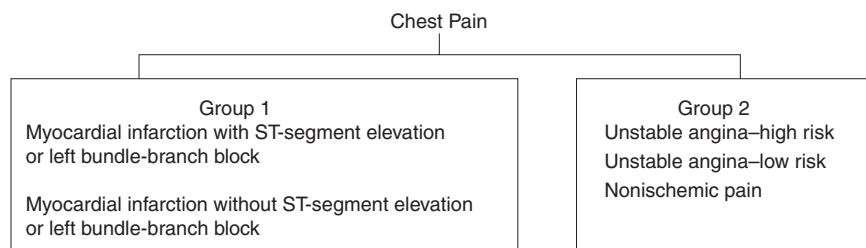


Figure 35-2 Categorization of Patients With Acute Chest Pain in Studies Ascertaining Test Properties of History, Physical Examination, and Electrocardiogram

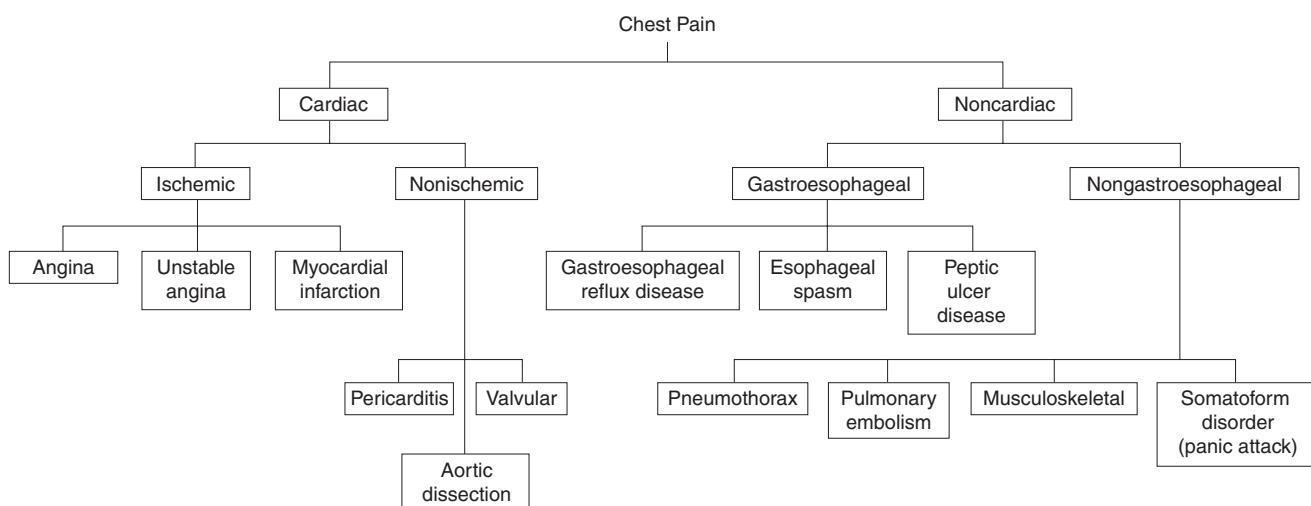


Figure 35-3 Cardiac and Noncardiac Conditions Presenting With Chest Pain

Pozen et al,⁸ in an evaluation of 1032 patients presenting to the emergency department with a chief symptom of chest pain, including follow-up ECG and cardiac enzyme tests for both hospitalized and nonhospitalized patients, reported an overall incidence of acute ischemia of 29% (ischemia included new-onset or unstable angina and MI). In an attempt to determine the etiology of noncardiac chest pain, Panju et al⁹ conducted further cardiac and gastrointestinal (GI) investigations in 100 patients discharged from a cardiac care unit (CCU) with chest pain not yet diagnosed (8.1% of the CCU admissions for chest pain). More than 75% of these patients had evidence of esophageal disorders by objective testing, including 24-hour intraesophageal pH monitoring, upper GI tract endoscopy with biopsy, esophageal motility studies, or upper GI tract barium series. These results are generalizable to patients discharged from the CCU with chest pain not yet diagnosed, a distinct subset of the patients who have noncardiac chest pain and present to the emergency department.

METHODS

Inclusion Criteria of Tests for Precision and Accuracy

Given the limited number of studies that have focused on the precision of the medical history, physical examination, and ECG in the diagnosis of MI, we developed a broad set of inclusion criteria. We included studies that consisted of an assessment of the interobserver or intraobserver variation, of features of the medical history, physical examination, and ECG among patients with chest pain or a diagnosis of MI.

For the accuracy of the medical history, physical examination, and ECG, we included studies that met the following criteria: patients: those with chest pain thought to be ischemic in nature; test: history, physical examination, or ECG described in adequate detail; outcome: MI or no infarction using the definition described above; sample size: studies with a sample size of at least 200 patients.

Search Strategy

For both precision and accuracy of the medical history, physical examination, and ECG, we performed an English-language MEDLINE search from 1980 to 1997, using the following Medical Subject Heading terms and search strategy: (1) “medical history taking or physical examination” and “myocardial infarction or chest pain” and (2) “reproducibility of results or observer variation” and “myocardial infarction” or “chest pain.” A textword search was also performed, using “interobserver,” “intraobserver,” “accuracy,” “precision,” “reliability,” “sensitivity,” “specificity,” and “myocardial infarction” or “chest pain.” Additional search strategies for accuracy included the term “myocardial infarction, diagnosis” (subheading). For all strategies, references from appropriate articles were reviewed to provide additional references for this article. Of the 14 references used to assess the precision and accuracy of the history, physical examination, and ECG in the diagnosis of acute MI, 12 were obtained from the MEDLINE search strategy and 2 from the review of reference lists.

Selection of Articles

One author (B.R.H.) initially screened the titles and abstracts. If she thought the articles might be relevant, she and another author (A.A.P.) reviewed the articles in detail and determined their eligibility.

Methodologic Quality Assessments

We evaluated the methodologic quality of articles addressing the accuracy of medical history, physical examination, or ECG using criteria developed for this series (see Table 1-7).¹⁰ A grade A designation meant an independent, blind comparison of sign or symptom with a gold standard among 500 or more consecutive patients suspected of having the target condition; grade B meant an independent, blind comparison of sign or symptom with a gold standard among fewer than 500 consecutive patients suspected of having the target condition;

grade C meant an independent, blind comparison of sign or symptom with a standard of uncertain validity or independent, blind comparison of sign or symptom with a gold standard among nonconsecutive patients suspected of having the target disorder.

Analysis

To calculate likelihood ratios (LRs) for features of the medical history, physical examination, and ECG, we considered studies suitable for combination if the sensitivity and specificity met one of the following criteria: χ^2 test of sensitivity and specificity excluding statistically significant heterogeneity ($P > .05$) or range of sensitivity and specificity across studies of 15% or less. We pooled studies satisfying at least 1 criterion and calculated LRs by simple combination of results across studies. The 95% confidence intervals (CIs) were calculated according to the method of Simel et al.¹¹

RESULTS

Precision of the Medical History and Physical Examination

Precision refers to the degree of variation between observers (interobserver variation) or within observers (intraobserver variation) regarding a particular clinical finding. Hickan et al¹² studied the precision of an important aspect of the history, namely, that of chest pain. They assessed the interobserver agreement in chest pain histories obtained by general internists, nurse practitioners, and self-administered questionnaires for 197 inpatients and 112 outpatients with chest pain. As outlined in Table 35-2, the 2 internists, who each independently interviewed 47 of 197 inpatients, showed high agreement for 7 of the 10 items, including location and description of the pain, as well as aggravating and relieving factors. Agreement was slightly lower between internist and questionnaire and between the nurse practitioners and internist, with the lowest level of agreement between nurse and questionnaire. Features of the chest pain associated with a

lower probability of MI, namely, pleuritic, positional, and sharp chest pain, typically showed a lower level of agreement for all comparisons.

The precision of the medical history obtained also depends on the reliability of the sources themselves. Kee et al¹³ assessed the reliability of a reported family history of MI from patients who had recently survived MI with that of other documented sources, including hospital charts and death certificates. They reported a moderate level of agreement, with a κ of 0.65.

Few studies have evaluated the precision of features of the physical examination in the assessment of patients with suspected MI. One study did evaluate the interobserver agreement among 3 clinicians in the assessment of physical symptoms and signs of heart failure in 102 MI patients.¹⁴ As shown in Table 35-3, agreement was high for dyspnea, as well as for the displaced apex beat. However, the level of agreement for the other physical symptoms and signs of heart failure, particularly the assessment of pulmonary rales and hepatomegaly, was considerably lower.

Precision of the Electrocardiogram Interpretation

Unfortunately, most studies that assessed the precision of ECG interpretation reported the percentage agreement between clinicians, without taking into account chance agreement through the use of κ or other statistical measures.¹⁵ Precise interpretations are important because they are made at the bedside and set off immediate management strategies. There are several factors that may influence the interpretation of the ECG, including the clinical observation of the patient and clinical data (expectation bias), as well as the training and experience of the individual reading the ECG. Although they must be interpreted with caution, the results of earlier studies suggest appreciable variability in precision in the interpretation of ECGs.

In one of the earlier studies,¹⁶ 10 clinicians with experience in cardiology read 100 ECGs on 2 separate occasions and classified the tracings as normal, abnormal, or infarction. The 3 clinicians

Table 35-2 Interobserver Agreement in Recording Chest Pain Histories^a

Attribute	Inpatients (n = 197)		Outpatients (n = 112)	
	Two Internists, κ	Internists and Questionnaire, κ	Nurse and Internists, κ	Nurse and Questionnaire, κ
Pain radiates to left arm	0.89	0.58	0.43	0.41
Pain relieved by nitroglycerin	0.79	0.51	0.94	0.77
History of myocardial infarction	0.78	0.81	0.70	0.81
Pain in substernal location	0.74	0.50	0.38	0.19
Pain brought on by exertion	0.63	0.51	0.42	0.22
Pain described as "pressure"	0.57	0.37	0.49	0.50
Patient must stop activities when pain occurs	0.50	0.47	0.44	0.40
Pain brought on by cough or deep breath	0.44	0.30	0.55	0.62
Pain described as "sharp"	0.30	0.26	0.33	0.31
Pain brought on by moving arms or torso	0.27	0.44	0.52	0.54

^aAdapted, with permission, from Hickan et al.¹²

agreed completely in only one-third of the ECGs. After a second reading, the clinicians disagreed with 1 of 8 of their original reports. Gjorup et al¹⁷ had 16 residents in internal medicine read 107 ECGs of suspected MI patients and assess whether signs indicative of acute infarction were present. There was disagreement in approximately 70% of the cases.

Brush et al¹⁸ reported much higher agreement in a study in which 2 clinicians classified 50 ECGs according to evidence of infarction, ischemia or strain, left ventricle hypertrophy,

LBBB, or paced rhythm. They obtained agreement in 45 of the 50 cases ($\kappa = 0.69$).

The precision in the interpretation of ECGs appears to increase with experience. Eight cardiologists interpreted ECGs of 1220 clinically validated cases of various cardiac disorders, including anterior, inferior, or combined MI, as well as right, left, or biventricular hypertrophy.¹⁹ The interobserver agreement among cardiologists was reasonably high, with an average κ of 0.67. For the 125 selected ECGs that were read twice by each cardiologist, different diagnoses were given for 10% to 23% of the ECGs (intraobserver reproducibility, 77%-90%).

Sgarbossa et al²⁰ assessed the precision of features of the ECG that may aid in the diagnosis of acute MI in the presence of LBBB. In this study, 4 investigators read 2600 ECGs and achieved a κ of more than 0.85 for QRS-complex and T-wave polarities, with a high degree of correlation among the investigators for interpretation of ST-segment deviation (Pearson product moment correlation coefficient, > 0.9).

Studies Used to Determine Accuracy of the Medical History, Physical Examination, and Electrocardiogram

Table 35-4 summarizes features of the 14 studies^{8,21-33} used to determine the accuracy of the medical history, physical

Table 35-3 Interobserver Agreement in Assessment of Physical Symptoms and Signs of Heart Failure in Patients With Myocardial Infarction^a

Physical Sign	Range, κ
Dyspnea	0.62-0.75
Displaced apex beat	0.53-0.73
S ₃ gallop	0.14-0.37
Rales	0.12-0.31
Neck vein distention	0.31-0.51
Hepatomegaly	0-0.16
Dependent edema	0.27-0.64

^aAdapted, with permission, from Gadsbøll et al.¹⁴

Table 35-4 Features of Studies Used to Determine Accuracy of the Medical History, Physical Examination, and Electrocardiogram

Source, y	Methodologic Quality ^a	Inclusion Criteria	Incidence of MI, %	No. of Patients (% Women)	Age, y	Country
Rude et al, ²¹ 1983	A	Consecutive patients admitted to CCU with suspected MI	50	3697 (38)	Mean = 61	United States
Yusuf et al, ²² 1984	B	Consecutive patients admitted to CCU with suspected MI	85	475 (15)	Mean = 56	United Kingdom
Pozen et al, ⁸ 1984	A	Consecutive patients presenting to ED with chest pain	NR	2801 (NR)	Men \geq 30 Women \geq 40	United States
Lee et al, ²³ 1985	A	Consecutive patients presenting to ED with chest pain	17	596 (52)	\geq 25	United States
Tierney et al, ²⁴ 1986	B	Consecutive patients presenting to ED with chest pain	12	492 (NR)	Men \geq 30 Women \geq 40	United States
Herlihy et al, ²⁵ 1987	B	Consecutive patients admitted to CCU with suspected MI	44	265 (NR)	NR	United States
Klaeboe et al, ²⁶ 1987	B	Consecutive patients admitted to CCU with suspected MI	59	237 (36)	Range = 29-90	Norway
Rouan et al, ²⁷ 1989	A	Consecutive patients presenting to ED with chest pain	14	7115 (50)	\geq 30	United States
Solomon et al, ²⁸ 1989	A	Consecutive patients presenting to ED with chest pain	14	7734 (50)	\geq 30	United States
Berger et al, ²⁹ 1990	B	Consecutive patients admitted to hospital with chest pain	36	278 (31)	57	Switzerland
Weaver et al, ³⁰ 1990	C	Patients with chest pain brought to ED by paramedics	18	2472 (NR)	< 75	United States
Jonsbu et al, ³¹ 1991	B	Consecutive patients admitted to hospital with suspected MI	36	200 (NR)	NR	Norway
Karlson et al, ³² 1991	A	Consecutive patients admitted to hospital with suspected MI	20	4690 (NR)	NR	Sweden
Kudenchuk et al, ³³ 1991	C	Patients brought to ED by paramedics	33	1189 (34)	\leq 74	United States

Abbreviation: CCU, cardiac care unit; ED, emergency department; MI, myocardial infarction; NR, not reported.

^aSee "Methodologic Quality Assessments" subsection of the text for an explanation of these grades.

examination, and ECG in the diagnosis of acute MI. Five of the studies included consecutive patients presenting to the emergency department with chest pain,^{8,23,24,27,28} 7 included patients admitted to the hospital or CCU for suspected MI,^{21,22,25,26,29,31,32} and 2 included patients with chest pain who were brought to the emergency department by paramedics.^{30,33}

The studies examined a variety of features of the clinical examination and ECG. For the sake of relevance and clarity, we highlight the results of those variables in which an LR of 2.0 or more, or an LR of 0.5 or less, was obtained. These studies provide the best available evidence for identifying those features that aid in the diagnosis of MI.

Accuracy of the Medical History and Physical Examination

Nine of the studies outlined in Table 35-4 reported the relation between features of the clinical examination of patients presenting to the emergency department with chest pain, as determined by physicians, with that of the final diagnosis of MI. In all studies, the gold standard for the diagnosis of MI was based on cardiac enzyme and ECG changes, except for the study by Weaver et al,³⁰ in which the discharge diagnosis was used to define acute MI. Although features of the clinical examination are extremely insensitive in diagnosing MI, they are reasonably specific and their presence is more likely to occur in patients with MI.

Although patients can present with MI and have no chest pain, chest pain always prompts a consideration that the patient is having myocardial ischemia. Nonetheless, multivariate models show that the independent value of chest pain or pain in the left arm, once all factors are considered, has an odds ratio (OR) of only 2.7.⁸ Confining chest pain in the model to "chest pain as the most important symptom" has an even lower OR for MI (OR, 2.0).⁸

As noted in Table 35-5, chest pain radiation was the clinical feature that increased the probability of MI the most, with a wider extension of pain associated with the highest likelihood of MI. In particular, chest pain radiating to the left arm was twice as likely to occur in patients with, as opposed to those without, MI, whereas radiation to the right shoulder was about 2 times as likely, and radiation to both the left and right arm was 9 times as likely to occur in such patients. Chest pain radiating to the right arm alone has been reported to be an extremely specific, but insensitive, marker of MI (LR, 7.3; 95% CI, 3.9-14).²⁹ However, as reflected by the width of the CI, these results were based on a small number of subjects (6 of the 100 patients with MI) and therefore require confirmation.

Other items of the history that aided in the diagnosis of MI included history of MI (LR ≤ 3.0) or diaphoresis (LR, 2.0).

A number of features from the history and clinical examination thought to be useful in determining the presence of MI were of little value in establishing such a diagnosis. Features of the history, including age older than 60 years, male sex, history of angina or coronary artery disease, history of nitroglycerin use, duration of chest pain greater than 60 minutes, constant or episodic chest pain, and chest pain of sudden onset, were all associated with LRs of less than 2.

Adjectives used to describe the quality of the chest pain, including that of pressure, aching, and squeezing, were also associated with LRs of less than 2. Therefore, none of these features carry information independently useful in establishing an MI diagnosis.

The 3 components of the physical examination associated with LRs higher than 2 included the presence of a third heart sound (LR, 3.2), hypotension (LR, 3.1), and pulmonary crackles on auscultation (LR, 2.1). Dyspnea was not found to be an important component of the clinical examination. Other features frequently described in the assessment of the patient with chest pain, including bradycardia and tachycardia, were not evaluated.

Cardiac risk factors, including hypertension, smoking, obesity, hypercholesterolemia, diabetes, and a family history of cardiovascular disease, are frequently included in the medical history of a patient presenting with chest pain. However, current evidence provides little support for the diagnostic value of a history of these risk factors. In 3 large studies of patients presenting to the emergency department with chest pain, none of the classic cardiac risk factors emerged as independent predictors of acute MI.^{8,34,35}

Table 35-6 presents clinical features that decrease the probability of MI. Chest pain described as pleuritic, sharp, stabbing, or positional decreased the likelihood of MI significantly. In addition, chest pain reproduced by palpation on

Table 35-5 Clinical Features That Increase the Probability of a Myocardial Infarction in Patients Presenting With Acute Chest Pain

Clinical Feature	LR (95% CI)	Reference
Chest pain radiation	Both arms with pain	9.7 (4.6-20)
	Left arm pain	2.2 (1.6-3.1)
	Right shoulder pain	2.9 (1.4-6.0)
Third heart sound on auscultation	3.2 (1.6-6.5)	24
Hypotension (systolic blood pressure ≤ 80 mm Hg)	3.1 (1.8-5.2)	30
Pulmonary crackles on auscultation	2.1 (1.4-3.1)	24
Diaphoresis	2.0 (1.9-2.2)	24, 28, 31
Nausea or vomiting	1.9 (1.7-2.3)	24, 25, 29, 31
History of MI	1.5-3.0 ^a	8, 24

Abbreviations: CI, confidence interval; LR, likelihood ratio; MI, myocardial infarction.

^aIn heterogeneous studies the LRs are reported as ranges.

Table 35-6 Clinical Features That Decrease the Probability of a Myocardial Infarction in Patients Presenting With Acute Chest Pain

Clinical Feature	LR (95% CI)	Reference
Pleuritic chest pain	0.2 (0.2-0.3)	23, 24, 28
Chest pain sharp or stabbing	0.3 (0.2-0.5)	23, 24
Positional chest pain	0.3 (0.2-0.4)	23, 28
Chest pain reported by palpation	0.2-0.4 ^a	23, 24, 28

Abbreviations: CI, confidence interval; LR, likelihood ratio.

^aIn heterogeneous studies the LRs are reported as ranges.

physical examination was associated with a low LR, ranging from 0.2 to 0.4.

Accuracy of the Electrocardiogram

Eight studies addressed the accuracy of the ECG in diagnosing MI. The results reported in this article are for interpretation of the ECGs by clinicians and not by computer algorithms. Interpretation of the ECG was by an independent physician blinded to the clinical data in 5 of the studies,^{8,21,22,32,33} by the emergency department physician alone in 2 others,^{23,27} and by the emergency department physician with a review by an independent physician blinded to the clinical data in 1.²⁴ In all studies, the gold standard for the diagnosis of MI was based on cardiac enzyme levels, except for the study by Kudenchuk et al,³³ in which the hospital discharge diagnosis was used to define MI.

Several features of the ECG have been used to assist in the diagnosis of acute MI. The most common characteristics include the presence of Q waves, ST-segment elevation or depression, and T-wave inversion. As noted in Table 35-7, there was a considerable degree of variability among the studies for some of these features. New ST-segment elevation was the most powerful feature in increasing the probability of MI, with the LRs ranging from 5.7 to 54. The presence of a new Q wave was also much more likely to occur in patients with, as opposed to those without, MI, with LRs ranging from 5.3 to 25, although the usefulness of this finding was reduced when patients with old Q waves were included.

ST-segment depression, whether new or known to have been present previously, and new T-wave peaking or inversion were all approximately 3 times as likely to occur in patients with, as opposed to those without, MI. In addition, conduction defects, particularly those reported to be new, also increased the probability of MI.

A normal ECG result decreased the probability of MI the most and was associated with LRs of 0.1 to 0.3.^{19,20,26,31}

Table 35-7 Features of the Electrocardiogram That Increase the Probability of a Myocardial Infarction in Patients Presenting With Acute Chest Pain

Feature of the Electrocardiogram	LR (95% CI)	Reference
Any ST-segment elevation	11 (7.1-18)	24
New ST-segment elevation \geq 1 mm	5.7-54 ^a	21-24, 32, 33
New conduction defect	6.3 (2.5-16)	24
New Q wave	5.3-25 ^a	21, 24, 32, 33
Any Q wave	3.9 (2.7-5.7)	24
Any ST-segment depression	3.2 (2.5-4.1)	24
T-wave peaking or inversion \geq 1 mm	3.1 ^b	8
New ST-segment depression	3.0-5.2 ^a	21, 24, 32
Any conduction defect	2.7 (1.4-5.4)	24
New T-wave inversion	2.4-2.8 ^a	24, 32, 33

Abbreviations: CI, confidence interval; LR, likelihood ratio.

^aIn heterogeneous studies, the LRs are reported as ranges.

^bData not available to calculate CIs.

The Role of Combined Findings and Clinical Prediction Rules for Myocardial Infarction

Clinicians are frequently presented with multiple clinical examination items, each of which can be considered a separate diagnostic test for establishing the diagnosis of MI. The problem in situations such as this is in knowing how to combine the LRs from these multiple tests to obtain an accurate estimate of the posttest probability of MI. The simple serial multiplication of LRs that has been proposed assumes that the tests are conditionally independent¹⁵; that is, that the patient's results on one test bear no relationship to the results on any of the other tests. As demonstrated by Holleman and Simel,³⁶ violation of the conditional independence assumption can yield inaccurate posttest probabilities of disease. Unfortunately, the precision and accuracy of serial combinations of findings were not reported in the studies included in this review. However, the combination of clinical findings considered as a group is assessed in clinical prediction rules.

By combining findings from patients' medical history, physical examination, and ECG, investigators have developed probability-based decision aids, as well as computer-based protocols and guidelines, that categorize patients with chest pain into risk groups according to their probability of MI.^{34,35,37} These tools have been devised to improve physician recognition and triage of patients with acute ischemic events.^{8,38} Although these measures have helped clinicians make appropriate decisions, not all studies of probability-based risk assessment tools have demonstrated improvement in emergency department triage or reduction in resource use.³⁹ These clinical prediction rules conform to the methodologic standards of clinical prediction rules initially proposed by Wasson et al⁴⁰ and recently revised,⁴¹ except for the validation of the rule by Tierney et al,³⁴ which was performed on a subset, rather than on a prospective sample of the population.

Tierney et al³⁴ developed an instrument for the prediction of MI. According to multivariate analysis of 540 emergency department patients with chest pain, 4 variables with independent predictive value for infarction were identified. These included diaphoresis with chest pain, history of MI, ECG changes of a new Q wave, and ST-segment elevation either new or old.

Goldman et al^{35,37} also developed a protocol to predict MI in emergency department patients with chest pain. The instrument was based on the medical history, physical examination, and ECG of more than 6000 patients presenting at an emergency department with a chief complaint of chest pain. Variables in Goldman's algorithm include patient older than 40 years, history of angina or MI, chest pain that began less than 48 hours before arrival at the emergency department, longest pain episode 1 hour or more, pain worse than usual angina or the same as earlier MI, and radiation of pain to neck, left shoulder, or left arm as predictors of infarction. Features of the chest pain, including radiation to the back, abdomen, or legs; stabbing pain; and pain reproduced by palpation included in the algorithm decrease the probability of infarction. The ECG changes predictive of an acute MI included new ST-segment elevation or Q waves in 2 or more

leads and new ST-T-segment changes of ischemia or strain. According to the algorithm, patients can be assigned to one of 14 subgroups, with a probability of acute MI ranging from 1% to 77%.

These prediction rules included several of the common variables identified in univariate analysis and included in this review, namely, the location and extent of the chest pain, chest pain with diaphoresis, and ECG changes, including new Q-wave and ST-segment elevation. However, in situations in which the independence of features of the medical history and clinical examination has not been tested, as in these studies, clinicians may be misled when combining these multiple clinical findings. In these situations they should look to clinical prediction rules to help integrate and interpret the results.

Pretest Probability in the Diagnosis of Myocardial Infarction

To determine the posttest probability, or likelihood, of disease according to the clinical features and their associated LRs, one must take into account the pretest probability, or likelihood, of that condition. Although much focus has been placed on the combination of multiple clinical variables and the development of prediction rules for MI, as described above, there has been little emphasis on establishing the pretest probability of MI according to standard clinical assessment. If an estimate of the pretest probability of MI is available, a diagnostic test, based on its sensitivity, specificity, and LR, can be used to establish a new estimate of disease likelihood. A classic and widely used example of this concept was proposed by Diamond and Forrester.⁴² Estimates of the pretest probability of coronary artery disease according to age, sex, and chest pain description have been published and are easily used in the clinical setting. A more comprehensive attempt to consider all clinical characteristics has also been undertaken.⁴³

The predictive value of the medical history, physical examination, and ECG depends on the pretest probability of MI. Even with a normal ECG result, for example, a high pretest probability of MI would result in a high posttest probability of this condition being present. Proper use of these findings must therefore incorporate the pretest probability of MI.

COMMENT

The diagnosis of MI in the setting of chest pain is a complex task. Clinicians categorize patients with chest pain into 3 groups according to current therapeutic interventions, whereas in the literature patients with chest pain are typically categorized into the presence or absence of MI. According to this latter categorization, we have assessed the features of the medical history, physical examination, and ECG, which aid in increasing or decreasing the likelihood of acute MI. We have also addressed the use of clinical prediction rules, which use a number of clinical variables, to aid in the diagnosis of MI, as well as the need to take into account pretest probability

of disease when assessing the predictive value of individual variables.

Referring to the scenarios presented at the beginning of this article, the first 3 have features that increase the likelihood of acute MI. Patient 1 has chest pain, diaphoresis, and ST-segment elevation. Patient 2 has diaphoresis, hypotension, and history of an MI. Patient 3 has nausea and ST-segment elevation. In contrast, patient 4 has features that decrease the likelihood of MI, namely, a normal ECG result and chest pain that is both positional and reproducible by palpation.

Clinicians interested in distinguishing patients with acute MI from those with unstable angina and nonanginal chest pain can use either Goldman's algorithm or the individual clinical features that we summarize in Tables 35-5, 35-6, and 35-7. However, the distinction between MI and non-MI chest pain may not be the most relevant initial clinical decision; it is more important to decide on appropriate immediate therapy.

THE BOTTOM LINE

The presence of any of the following clinical findings increases the likelihood of MI: patients presenting with chest pain radiating to the left arm, radiating to the right shoulder, or radiating to both left and right arms; and patients presenting with chest pain diaphoresis, a third heart sound, or with hypotension.

The presence of any of the following clinical findings decreases the likelihood of MI: patients presenting with chest pain that is described as pleuritic, sharp or stabbing, positional, or reproduced by palpation.

Features of ECG that increase the likelihood of MI include the following: new ST-segment elevation, new Q waves, any ST-segment elevation, and new conduction defect. A normal ECG result is a powerful feature in ruling out MI.

Finally, as noted previously, these findings may not be relevant for distinguishing between patients with acute ischemic syndromes requiring CCU admission from those with less dangerous ischemia or nonischemic pain. Further research is required in this regard.

Author Affiliations at the Time of the Original Publication

Departments of Medicine (Drs Panju and Guyatt), Clinical Epidemiology and Biostatistics (Dr Guyatt), and McMaster Medical Programme (Dr Hemmelgarn), McMaster University, Hamilton, Ontario, Canada; and the Center for Health Services Research in Primary Care, Durham Veterans Affairs Medical Center, Duke University Medical Center, Durham, North Carolina (Dr Simel). Dr Hemmelgarn is currently a resident in internal medicine at the University of Calgary, Alberta, Canada.

Acknowledgments

We are indebted to Eric C. Westman, MD, Michael Cuffe, MD, Salim Yusuf, MD, and Ernest Fallen, MD, for their review and contribution to the manuscript, as well as to John Attia, MD, Arie Levinson, MD, and James Velianou, MD, for their suggestions on the final manuscript.

REFERENCES

1. Mathews MB, Julian DG. Angina pectoris: definition and description. In: Julian DG, ed. *Angina Pectoris*. 2nd ed. New York, NY: Churchill Livingstone; 1985:2.
2. Campeau L. Grading of angina pectoris. *Circulation*. 1976;54(3):522-523.
3. Braunwald E, Jones RH, Mark DB, et al. Diagnosing and managing unstable angina. *Circulation*. 1994;90(1):613-622.
4. Tunstall-Pedoe H, Kuulasmaa K, Amouyel P, Arveiler D, Rajakangas AM, Pajak A. Myocardial infarction and coronary deaths in the World Health Organization Monica Project. *Circulation*. 1994;90(1):583-612.
5. Kannel WB, Abbott RD. Incidence and prognosis of unrecognized myocardial infarction. *N Engl J Med*. 1984;311(18):1144-1147.
6. Uretsky BF, Farquhar DS, Berezin AF, Hood WB. Symptomatic myocardial infarction without chest pain: prevalence and clinical course. *Am J Cardiol*. 1977;40(4):498-503.
7. Ness TJ, Gebhart GF. Visceral pain: a review of experimental studies. *Pain*. 1990;41(2):167-234.
8. Pozen MW, D'Agostino RB, Selker HP, Sytkowski PA, Hood WB. A predictive instrument to improve coronary care unit admission practices in acute ischemic heart disease. *N Engl J Med*. 1984;310(20):1273-1278.
9. Panju A, Farkouh ME, Sackett DL, et al. Outcome of patients discharged from a coronary care unit with a diagnosis of "chest pain not yet diagnosed." *CMAJ*. 1996;155(5):541-547.
10. Holleman DR, Simel DL. Does the clinical examination predict airflow limitation? *JAMA*. 1995;273(4):313-319.
11. Simel DL, Samsa GP, Matchar DB. Likelihood ratios with confidence: sample size estimation for diagnostic test studies. *J Clin Epidemiol*. 1991;44(8):763-770.
12. Hickam DH, Sox HC, Sox CH. Systematic bias in recording the history in patients with chest pain. *J Chronic Dis*. 1985;38(1):91-100.
13. Kee F, Tiret L, Robo JY, et al. Reliability of reported family history of myocardial infarction. *BMJ*. 1993;307(6918):1528-1530.
14. Gadsbøll N, Hoilund-Carlsen PF, Nielsen GG, et al. Symptoms and signs of heart failure in patients with myocardial infarction: reproducibility and relationship to chest x-ray, radionuclide ventriculography and right heart catheterization. *Eur Heart J*. 1989;10(11):1017-1028.
15. Sackett DL, Haynes RB, Guyatt GH, Tugwell P. *Clinical Epidemiology: A Basic Science for Clinical Medicine*. 2nd ed. Boston, MA: Little Brown & Co; 1991.
16. Davies LG. Observer variation in reports of electrocardiograms. *Br Heart J*. 1958;20(2):153-161.
17. Gjorup T, Kelbaek H, Nielsen D, Kreiner S, Godtfredsen J. Interpretation of the electrocardiogram in suspected myocardial infarction: a randomized controlled study of the effect of a training programme to reduce interobserver variation. *J Intern Med*. 1992;231(4):407-412.
18. Brush JE, Brand DA, Acampora D, Chalmer B, Wackers FJ. Use of the initial electrocardiogram to predict in-hospital complications of acute myocardial infarction. *N Engl J Med*. 1985;312(18):1137-1141.
19. Willems JL, Abreu-Lima C, Arnaud P, et al. The diagnostic performance of computer programs for the interpretation of electrocardiograms. *N Engl J Med*. 1991;325(25):1767-1773.
20. Sgarbossa EB, Pinski SI, Barbagelata A, et al. Electrocardiographic diagnosis of evolving acute myocardial infarction in the presence of left bundle-branch block. *N Engl J Med*. 1996;334(8):481-487.
21. Rude RE, Poole WK, Muller JE, et al. Electrocardiographic and clinical criteria for recognition of acute myocardial infarction based on analysis of 3,697 patients. *Am J Cardiol*. 1983;52(8):936-942.
22. Yusuf S, Pearson M, Parish S, Ramsdale D, Rossi P, Sleight P. The entry ECG in the early diagnosis and prognostic stratification of patients with suspected acute myocardial infarction. *Eur Heart J*. 1984;5(9):690-696.
23. Lee TH, Cook EF, Weisberg M, Sargent RK, Wilson C, Goldman L. Acute chest pain in the emergency room. *Arch Intern Med*. 1985;145(1):65-69.
24. Tierney WM, Fitzgerald D, McHenry R, et al. Physicians' estimates of the probability of myocardial infarction in emergency room patients with chest pain. *Med Decis Making*. 1986;6(1):12-17.
25. Herlihy T, McIvor ME, Cummings CC, Siu CO, Alikahn M. Nausea and vomiting during acute myocardial infarction and its relation to infarct size and location. *Am J Cardiol*. 1987;60(1):20-22.
26. Klaeboe G, Otterstad JE, Winsnes T, Espeland N. Predictive value of prodromal symptoms in myocardial infarction. *Acta Med Scand*. 1987;222(1):27-30.
27. Rouan GW, Lee TH, Cook EF, Brand DA, Weisberg MC, Goldman L. Clinical characteristics and outcome of acute myocardial infarction in patients with initially normal or nonspecific electrocardiograms. *Am J Cardiol*. 1989;64(18):1087-1092.
28. Solomon CG, Lee TH, Cook EF, et al. Comparison of clinical presentation of acute myocardial infarction in patients older than 65 years of age to younger patients: the Multicenter Chest Pain Study experience. *Am J Cardiol*. 1989;63(12):772-776.
29. Berger JP, Buclin R, Haller E, Van Melle G, Yersin B. Right arm involvement and pain extension can help to differentiate coronary diseases from chest pain of other origin: a prospective emergency ward study of 278 consecutive patients admitted for chest pain. *J Intern Med*. 1990;227(3):165-172.
30. Weaver WD, Eisenberg S, Martin JS, et al. Myocardial Infarction Triage and Intervention Project, phase I: patient characteristics and feasibility of prehospital initiation of thrombolytic therapy. *J Am Coll Cardiol*. 1990;15(5):925-931.
31. Jonsbu J, Rollag A, Aase O, et al. Rapid and correct diagnosis of myocardial infarction: standardized case history and clinical examination provide important information for correct referral to monitored beds. *J Intern Med*. 1991;229(2):143-149.
32. Karlson BW, Herlitz J, Wiklund O, Richter A, Hjalmarson A. Early prediction of acute myocardial infarction from clinical history, examination and electrocardiogram in the emergency room. *Am J Cardiol*. 1991;68(2):171-175.
33. Kudenchuk PJ, Ho MT, Weaver D, et al. Accuracy of computer-interpreted electrocardiography in selecting patients for thrombolytic therapy. *J Am Coll Cardiol*. 1991;17(7):1486-1491.
34. Tierney WM, Roth BJ, Psaty B, et al. Predictors of myocardial infarction in emergency room patients. *Crit Care Med*. 1985;13(7):526-531.
35. Goldman L, Cook EF, Brand DA, et al. A computer protocol to predict myocardial infarction in emergency department patients with chest pain. *N Engl J Med*. 1988;318(13):797-803.
36. Holleman DR, Simel DL. Quantitative assessments from the clinical examination: how should clinicians integrate the numerous results? *J Gen Intern Med*. 1997;12(3):165-171.
37. Goldman L, Weinberg M, Weisberg M, et al. A computer-derived protocol to aid in the diagnosis of emergency room patients with acute chest pain. *N Engl J Med*. 1982;307(10):588-596.
38. Sarasin F, Reymond J, Griffith J, et al. Impact of the acute cardiac ischemia time-insensitive predictive instrument (ACI-TIPI) on the speed of triage decision making for emergency department patients presenting with chest pain. *J Gen Intern Med*. 1994;9(4):187-194.
39. Pearson S, Goldman L, Garcia T, Wok E, Lee T. Physician response to a prediction rule for triage of emergency department patients with chest pain. *J Gen Intern Med*. 1994;9(5):241-247.
40. Wasson JH, Sox HC, Neff RK, Goldman L. Clinical prediction rules: application and methodological standards. *N Engl J Med*. 1985;313(13):793-799.
41. Laupacis A, Sekar N, Stiell IG. Clinical prediction rules: a review and suggested modifications of methodological standards. *JAMA*. 1997;277(6):488-494.
42. Diamond GA, Forrester JS. Analysis of probability as an aid in the clinical diagnosis of coronary artery disease. *N Engl J Med*. 1979;300(24):1350-1358.
43. Pryor DB, Harrell FE Jr, Lee KL, Calif RM, Rosati RA. Estimating the likelihood of significant coronary artery disease. *Am J Med*. 1983;75(5):771-780.

Prepared by David L. Simel, MD, MHS

Reviewed by Stephen W. Goodacre, MD, and L. Kristin Newby, MD

CLINICAL SCENARIO

A 62-year-old woman experienced chest discomfort while walking from the parking garage to your hospital. She decided to stop in the emergency department for evaluation. The discomfort has been present for about 10 to 12 minutes and is creating a dull ache in her left shoulder and arm. As you interview her, she is diaphoretic and experiencing the chest discomfort. Her blood pressure is 145/95 mm Hg. The lungs are clear to auscultation, whereas the cardiac examination reveals an S₄ systolic sound but no murmur. The pulses are equal in all of her extremities. An electrocardiogram (ECG) result seems normal, but your hospital provides neither computerized ECG reports nor computerized estimates of the probability of a myocardial infarction (MI). She experiences relief after a sublingual nitroglycerin tablet. You find that she was recently diagnosed with diabetes and systolic hypertension, and she has been trying to stop smoking. There was no nausea with the discomfort, although she observes frequent epigastric discomfort that responds to antacids. She takes cimetidine, which helps with her discomfort.

UPDATED SUMMARY ON MYOCARDIAL INFARCTION

Original Review

Panju AA, Hemmelgarn BR, Guyatt GH, Simel DL. The rational clinical examination: is this patient having a myocardial infarction? *JAMA*. 1998;280(14):1256-1263.

UPDATED LITERATURE SEARCH

Our literature search combined the search terms “myocardial infarction/di” with the parent search strategy for The Rational Clinical Examination series, “meta-analysis,” or “roc curve,” limited to English-language publications in the MEDLINE database from 1997 to October 2004. The search strategy yielded 169 articles, although only 1 prospectively evaluated the sensitivity and specificity of the clinical examination for acute cardiac ischemia (ACI). Acute cardiac ischemia includes patients with MI or unstable angina pectoris.

NEW FINDINGS

- The reference standard for MI now includes cardiac troponin levels.
- The new reference standard requires reappraisal of the role of clinical findings.
- After clinical symptoms are used to identify patients with possible ischemia, the ECG and troponin results take precedence in making the diagnosis.
- Radiation of chest pain to the shoulder or right arm has been validated as reflecting a more diffuse pain pattern that increases the likelihood of an MI among patients admitted to the hospital. However, the value of individual clinical symptoms or signs in the decision to admit or discharge the patient has not been fully evaluated with troponin-based case definitions.
- The presence of diabetes, hypertension, or hyperlipidemia should not affect the clinician’s probability estimate that an episode of chest pain represents an ACI.

Details of the Update

We found a systematic review of the diagnosis of ACI, published in 2001, that formed the basis for evidence-based guidelines.¹ We reviewed this article, articles in the reference lists of a general systematic review on MI,² and a recent nonsystematic review of ACI diagnosis.³ These 3 reviews addressed the reference standard for acute MI, the Goldman chest pain protocol, Acute Cardiac Ischemia Time-Insensitive Predictive Instrument (ACI-TIPI), and computerized decision aids for diagnosing MI. Each of these diagnostic approaches uses combinations of findings (including the clinical examination) to diagnose acute MI. From reviewing the reference lists in the review articles, it became apparent that information on the clinical examination might be buried in articles that did not lead to Medical Subject Headings indexing of clinical examination terms. To explore the possibility that we might be missing articles, we entered the original Rational Clinical Examination article into Citation Index (ISI Web of Knowledge, Science Citation Index Expanded). Thirty articles cited the original Rational Clinical Examination article on MI, including 3 that contained new information on the sensitivity and specificity of clinical evaluation items.

IMPROVEMENTS IN THE DATA PRESENTED IN THE ORIGINAL PUBLICATION

For chest pain radiation patterns shown in Table 35-5, we reassessed the values. We found 1 minor calculation error in the original Rational Clinical Examination article for the likelihood ratio (LR) when chest pain radiates to both arms. We also found that the data from the studies referenced in

Table 35-8 Likelihood Ratios of Chest Pain Radiation Patterns for Myocardial Infarction

Pain radiation	LR+ (95% CI)	LR- (95% CI)
Both arms with pain ⁴	9.7 (4.6-20)	0.64 (0.54-0.74)
Right arm pain ⁴	7.3 (3.9-14)	0.62 (0.52-0.73)
Left arm pain ⁴	2.2 (1.6-301)	0.60 (0.48-0.75)
Right shoulder pain (n = 2) ^{4,5}	2.2 (1.4-3.4) ^a	0.90 (0.82-0.97) ^a

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

^aValues represent summary likelihood ratios.

Table 35-9 Criteria for Acute, Evolving, or Recent Myocardial Infarction^a

Either of the following criteria satisfies the diagnosis for an acute, evolving, or recent MI:

1. Typical increase and gradual decrease (troponin) or more rapid increase and decrease (creatinine kinase-MB isoenzyme) of biochemical markers of myocardial necrosis, with at least 1 of the following:
 - a. Ischemic symptoms
 - b. Development of pathologic Q waves on the ECG
 - c. ECG changes indicative of ischemia (ST-segment elevation or depression) *or*
 - d. coronary artery intervention (eg, coronary angioplasty)
2. Pathologic findings of an acute MI

Abbreviations: ECG, electrocardiogram; MI, myocardial infarction.

^aFrom Alpert et al.⁶

Table 35-10 Effect of Change in Case Definition on Sensitivity and Specificity^a

MI (WHO Criteria 1990)	No MI	LR+ (95% CI)	LR- (95% CI)
Original Data Before Cardiac Troponins Were Available^a			
Nausea	40	30	2.4 (1.6-3.6)
No nausea	60	150	0.72 (0.6-0.84)
Scenario 1^b			
All Newly Reclassified Patients Have Nausea			
Nausea	62	← 8	10 (5.0-20)
No nausea	60	150	0.52 (0.43-0.62)
Scenario 2^b			
No Newly Reclassified Patients Have Nausea			
Nausea	40	30	1.7 (1.1-2.6)
No nausea	82	← 128	0.83 (0.72-0.90)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio; MI, myocardial infarction; WHO, World Health Organization.

^aThe data are rounded for display purposes.

^bArrows indicate the effect of reclassifying 7.7% of patients without MI to having an MI.

Table 35-5 were sufficient for calculating the negative likelihood ratios. These new findings appear in Table 35-8.

CHANGES IN THE REFERENCE STANDARD

The case definition for MI changed with the validation of new biomarkers.⁶ The change in cardiac troponin now has primacy over other biochemical markers (Table 35-9).

The new definition means that more patients will now be diagnosed with MI. A study of patients admitted to a cardiac care unit (CCU) showed that the new definitions of MI with troponin would lead to an absolute increase of 7.7% in the incidence (assuming threshold values of 8.0 mg/mL for creatine kinase-MB isoenzyme (CK-MB) and 0.3 mg/mL for troponin).⁷ However, the new incidence depends on the threshold values for troponin and CK-MB, and the rates could vary considerably, depending on the study population.

The change in MI definition undoubtedly affects the sensitivity and specificity for clinical symptoms and signs, but the direction and magnitude of the effect are unpredictable. Although cardiac troponin has better sensitivity and specificity than the CK-MB, the issue for the clinical examination becomes a question of how the symptoms or signs will distribute among those newly classified as having an MI. To demonstrate this, we can look at the effect on the finding of nausea as a presenting symptom. Berger et al⁴ evaluated nausea with the World Health Organization criteria for MI before cardiac troponin levels and found the results shown in Table 35-10. According to an increased prevalence of 7.7% in MI using troponin as part of the case definition, Table 35-10 shows the extremes of the effect on the LR of nausea for MI, depending on the distribution of newly defined cases with troponins. At one extreme, we show what happens if all the “new” patients with MI had nausea (scenario 1), whereas the opposite extreme occurs when none of the new patients experienced nausea (scenario 2); the truth will be somewhere between these extremes.

In scenario 1, the results lead to an improved utility of nausea for correctly identifying patients. In scenario 2, in which all new cases are from patients who lacked nausea as a presenting clinical feature, the value of the finding is worse compared with the older case definition. The results could change between populations with different baseline incident rates and different threshold values for biomarkers. Thus, older data on the clinical examination applied to current diagnostic standards may lead to either underestimates or overestimates of the utility of symptoms or signs. The most reliable estimates for the clinical examination will come from studies that compare the results to current definitions of MI.

RESULTS OF LITERATURE REVIEW

The patients in the studies shown in Table 35-11 had normal or nondiagnostic ECG results. Patients with known coronary heart disease who had prolonged or recurrent pain typical of angina, those with suspected pulmonary emboli, or those with comorbid illness requiring admission were excluded. Thus, the patients remaining for inclusion in the study were typical of those presenting with chest pain who might have acute coronary syndromes, but for whom the diagnosis is uncertain.

The finding that radiation of pain to the right arm or both arms has diagnostic value may seem counterintuitive to physicians who consider only left arm pain as related to myocardial ischemia. However, in the original study that reported the significance of right arm discomfort, 45 of 51 patients with pain in the right arm also had pain in the left arm.⁴ The authors speculated that the presence of right arm pain represents part of a larger extension of pain with an MI, rather than something intrinsic to the radiation of pain with MI. A count of patients who had both right arm and left arm pain in the studies by Goodacre et al^{8,9} was not provided, but the importance of the finding of pain in the right arm was confirmed and was present even after adjusting for other symptoms.

Chest discomfort with indigestion/burning quality independently increased the likelihood of an MI (positive LR, 2.3; 95% confidence interval (CI), 1.5-3.5).⁹ Because the presence of indigestion/burning might have been used to discharge patients from the emergency department, this created verification bias that could have distorted the LR. Therefore, the value of this symptom is of uncertain significance when assessing for MI.

For patients with chest pain, the response to nitroglycerin does not distinguish those who will prove to have an MI from those who will not. The diagnostic odds ratio combined from 2 studies was not significantly different from 1 (ie, diagnostic odds ratio, 0.66; 95% CI, 0.38-1.2).^{8,10}

Multivariate Findings for ACI Syndromes

One study evaluated a variety of clinical symptoms among a group of patients who proved to have a 30% incidence of MI.¹¹ The model included clinical variables without ECG data (Box 35-1). The model has *not* been studied as a tool for emergency department triage of patients with chest pain. A patient's increasing age, the presence of diaphoresis with the chest discomfort, nausea, and left arm radiation were the most important variables for increasing the probability of an MI. The variables that decreased the probability the most were the presence of pleuritic chest discomfort or episodic pain.

A second study evaluated the clinical findings after an ECG had been obtained (Box 35-2).⁹ This study included only patients with normal or nondiagnostic ECG results after excluding those with known coronary heart disease who had prolonged or recurrent pain typical of angina, those with suspected pulmonary emboli, or those with comorbid illness requiring admission.

There are similarities in the variables between the multivariate models of these 2 studies,^{9,11} although the populations were quite different. The generalizability of these models, both of which address important study populations, should be verified in new studies. The first multivariate model makes sense for patient education and is compatible with what patients should generally understand—chest pain associated with sweating or diaphoresis may be a harbinger of an MI, especially in a smoker or when the pain goes to the left arm.

The second multivariate model should be useful to clinicians because the model applies only to those who would not be readily admitted for additional testing to rule out an MI.

Several chest pain protocols or decision models have been studied to assess their performance in identifying patients with cardiac ischemia (Table 35-12).³ These approaches can

Table 35-11 Univariate Findings for Acute Myocardial Infarction in Patients With Undifferentiated Chest Pain Admitted for Suspected Acute Coronary Syndrome^a

Symptom	LR+ (95% CI)	LR- (95% CI)
Radiation to the shoulder OR both arms ^b	4.1 (2.5-6.5)	0.68 (0.52-0.89)
Radiation to right arm ^b	3.8 (2.2-6.6)	0.86 (0.77-0.96)
Vomiting ^b	3.5 (2.0-6.2)	0.87 (0.79-0.97)
Ex-smoker ^b	2.5 (1.6-4.0)	0.85 (0.76-0.96)
Male sex ^b	1.5 (1.3-1.6)	0.24 (0.12-0.48)
Current smoker ^b	1.4 (1.0-1.8)	0.83 (0.68-1.0)
Radiation to left arm ^b	1.3 (0.93-1.8)	0.90 (0.76-1.1)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

^aAll patients had normal or nondiagnostic electrocardiogram, no established coronary heart disease, and prolonged or recurrent chest pain typical of their usual discomfort.

Table 35-12 Likelihood Ratios of Chest Pain Protocols for Acute Cardiac Ischemia^c

Test (No. of Studies)	Diagnosis ^d	Sensitivity (Range)	Specificity (Range)	LR+ ^e	LR- ^e
ACI-TIPI ^f (4)	ACI ^d	0.86-0.95	0.78-0.92	3.9-12	0.05-0.18
Goldman protocol (3)	AMI	0.88-0.91	0.70-0.74	2.9-3.6	0.12-0.17
Computer-based decision aids (6)	AMI	0.52-0.98	0.58-0.96	1.2-24	0.02-0.83

Abbreviations: ACI, acute cardiac ischemia; ACI-TIPI, Acute Cardiac Ischemia Time-Inensitive Predictive Instrument; AMI, acute myocardial infarction; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

^cAll diagnoses were based on World Health Organization criteria before adoption of the cardiac troponin assay.

^dLikelihood ratio ranges are estimated from ranges for sensitivity and specificity using data provided by authors.

^ePatient age, sex, and chest pain or left arm pain were the primary symptoms, plus a computerized analysis of the electrocardiogram (Q-wave presence and assessment of the ST and T waves).

^fACI includes AMI and unstable angina.

Box 35-1 All Patients With Chest Pain Using Data Obtained Without Knowledge of the Electrocardiogram Results

$$\text{MI score} = -92 + 1.0 \times (\text{age}) + 17 \times (\text{diaphoresis}) + 14 \times (\text{nausea}) + 11 \times (\text{smokes}) + 11 \times (\text{left arm pain}) + 8 \times (\text{male}) - 44 \times (\text{pleuritic pain}) - 30 \times (\text{episodic pain}) - 15 \times (\text{sharp pain}) - 15 \times (\text{previous angina}) - 12 \times (\text{previous MI})$$

(If symptom present, substitute 1; if symptom absent, substitute 0.)

$$\text{MI probability} = [\exp^{(\text{score}/15)}]/[1 + \exp^{(\text{score}/15)}]$$

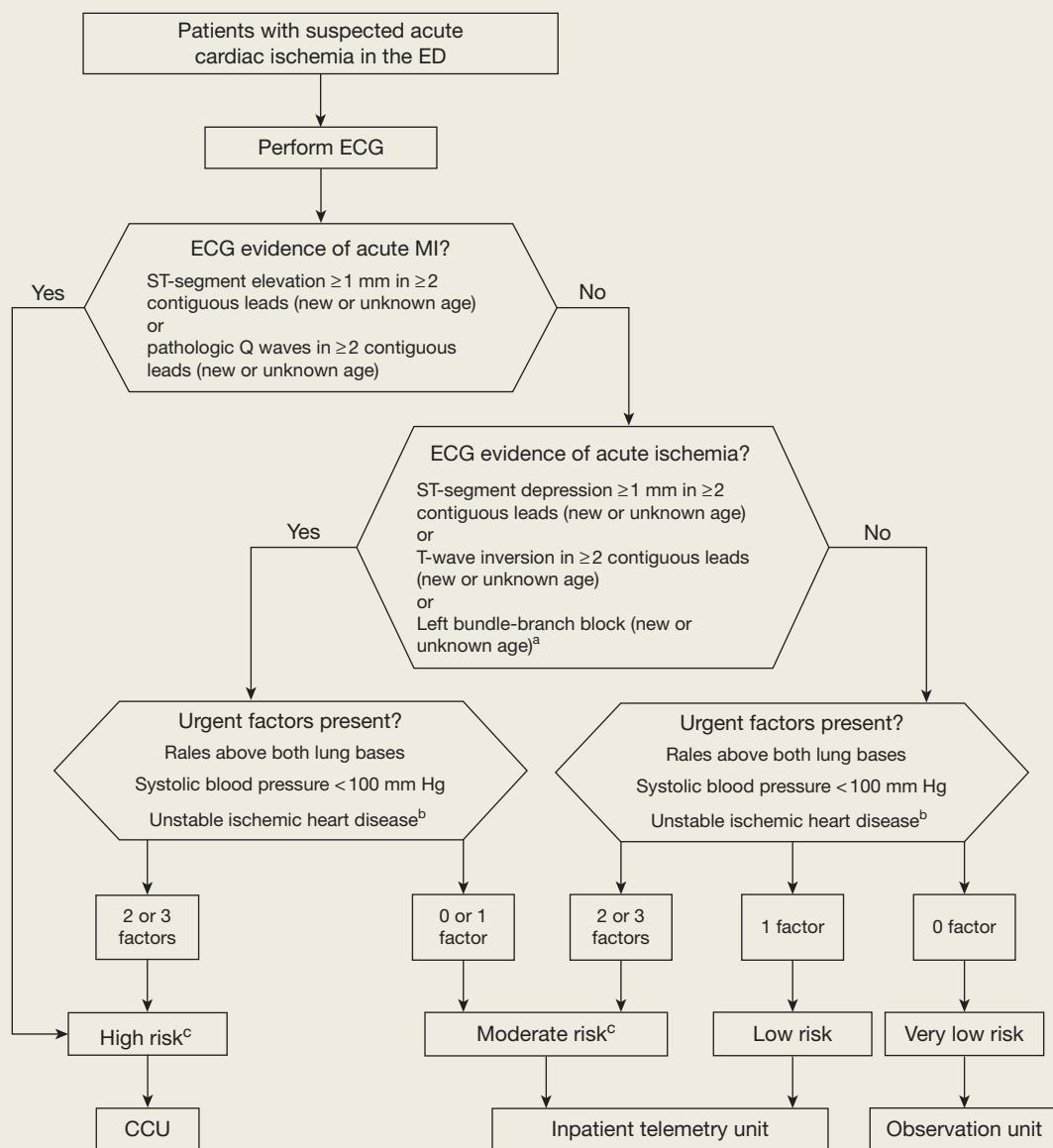
Box 35-2 Patients With Undifferentiated Chest Discomfort After a Normal or Nondiagnostic Electrocardiogram Result^a

$$\text{MI score} = 116 + 1.0 \times (\text{age}) + 23 \times (\text{male}) + 21 \times (\text{right arm pain}) + 18 \times (\text{ex-smoker}) + 11 \times (\text{left arm pain}) + 15 \times (\text{vomiting}) + 15 \times (\text{smokes}) + 10 \times (\text{burning pain})$$

(If symptom present, substitute 1; if symptom absent, substitute 0.)

$$\text{MI probability} = [\exp^{(\text{score}/11)}]/[1 + \exp^{(\text{score}/11)}]$$

^aModel provided by Dr. Steve Goodacre from data originally reported in Goodacre et al, 2003.⁹

**Figure 35-4** Goldman Chest Pain Decision RuleFrom Reilly et al.¹²

Abbreviations: CCU, cardiac care unit; ECG, electrocardiogram; ED, emergency department; MI, myocardial infarction.

^aModification to Goldman's prediction rule: left bundle-branch block not known to be old was also considered evidence of ischemia on ECG. ^bUnstable ischemic heart disease was defined as a worsening of previously stable angina, new onset of postinfarction angina or angina after a coronary revascularization procedure, or pain that was the same as that associated with a previous MI.

^cCardiology consultation in the ED (for possible admission to the CCU) was recommended for patients stratified as high risk, which included patients who had experienced a major complication in the ED (eg, cardiogenic shock). Modification to Goldman's prediction rule: cardiology consultation for possible CCU admission was also recommended for 2 subgroups of patients: (a) patients stratified as moderate risk by the original prediction rule because they had acute pulmonary edema or ongoing angina despite maximal medical therapy in the ED, and (b) patients presenting with unstable angina within 2 weeks of acute MI or within 6 months of coronary revascularization. Patients stratified as moderate risk who also had a high probability of significant coronary artery disease (using the Diamond and Forrester criteria¹³) were recommended for cardiology consultation.

be evaluated to see whether they appropriately identify patients with an MI or to see whether they have a favorable effect on the accuracy of patient management decisions. The protocols have not been extensively evaluated with current biomarkers for MI.

The Goldman chest pain protocol has been evaluated for safety and efficiency for triage decisions in a before-after study design. For avoiding major cardiac complications, the protocol allowed a safely increased admission rate to less-intensive

unmonitored beds of patients with possible ACI vs admission to monitored or CCU beds.¹² The protocol uses no symptoms but instead relies on the ECG, 2 physical examination findings, and 3 items from the clinical history (Figure 35-4).

EVIDENCE FROM GUIDELINES

The diagnosis and management guidelines for ACI were updated in 2004.¹⁴ A separate update on unstable angina, non-

ST-segment elevation MI was released in 2002 and revised in 2007.¹⁵ The guidelines emphasize the importance of the ECG, an approach to early risk stratification (as opposed to focusing only on whether or not the patient has had an MI), and they emphasize that single clinical findings should not drive decision making and risk assessment because the diagnosis is based on a variety of findings. In providing general guidance, the authors recommend assessing symptoms, history of coronary heart disease, age, sex, and the number of traditional coronary heart disease risk factors. An increasing number of traditional coronary heart disease risk factors in a patient affects the prognosis of those who prove to have cardiac ischemia, but the number of risk factors itself does not correlate well with the likelihood of acute ischemia in an individual episode of chest pain.

CLINICAL SCENARIO—RESOLUTION

Most physicians will recognize that this patient could be having ACI. Many will assume this according to her recent diagnosis of diabetes and hypertension. However, these variables are not diagnostically important in assessing this episode. Instead, you should focus on the current symptoms, her age, and smoking status.

The immediate goal of the bedside clinician is the prompt assessment of the likelihood of ACI and risk stratification if the chest pain seems of cardiac origin.

Had your hospital provided ACI-TIPI estimates, the patient's age and chest or left arm pain as the primary symptom would have contributed to her probability of MI. However, her sex and the normal ECG result would have lowered the probability. The Goldman chest pain protocol suggests that she is at low risk of a major cardiac complication, but you are concerned that the risk of an MI is high.

The history of gastrointestinal symptoms might suggest that she is simply having peptic discomfort. Her response to nitroglycerin does not allow you to sort out cardiac from noncardiac chest discomfort.

Her age, diaphoresis with the pain, left arm discomfort, and her current smoking status are all important variables that increase the likelihood of an ACI event. Although a man with the same symptoms would have a higher risk of an MI, her sex does nothing to protect her from ischemia, given the current symptoms. A probability estimate is not necessary to make a decision that prompt management of ischemia and an effort to rule out an MI are necessary.

After you make your decision to rule out an MI, you decide to evaluate the effect that the normal ECG result had on the importance of the clinical findings. First, you insert the values for her clinical findings into the decision model developed by Wang et al.¹¹ The model shows that she has a 65% probability of an MI, supporting your decision to obtain serial cardiac troponin levels and ECG results. Her smoking status and association of the pain with diaphoresis were important variables because the absence of either of those would have decreased the probability of an MI to 23%, which is about the baseline risk for all patients presenting to the emergency department with possible ACI. However, the normal ECG result has a big effect on the clinical findings.

The model by Goodacre et al⁹ was evaluated in just such a population of patients with normal or nondiagnostic ECG results. With that model, the probability of an MI is about 7%. The presence of an MI is considerably lower with the normal ECG result, but the 7% prediction would still lead most physicians to obtain the serial biomarkers and ECGs.

REFERENCES FOR THE UPDATE

1. Agency for Healthcare Research and Quality. *Evaluation of Technologies for Identifying Acute Cardiac Ischemia in Emergency Departments*. Rockville, MD: Agency for Healthcare Research and Quality; 2000. Summary, Evidence Report/Technology Assessment: Number 26. AHRQ Publication No. 00-E031, September 2000. <http://www.ahrq.gov/clinic/epcsums/cardsum.htm>. Accessed June 5, 2008. In summary form in *Ann Emerg Med*. 2001;37(5):453-460.
2. Boersma E, Mercado N, Poldermans D, Gardien M, Vos J, Simoons ML. Acute myocardial infarction. *Lancet*. 2003;361(9360):847-858.
3. Pope JH, Selker HP. Diagnosis of acute cardiac ischemia. *Emerg Clin North Amer*. 2003;21:27-59.
4. Berger JP, Buclin T, Haller E, Melle GV, Yerson B. Right arm involvement and pain extension can help to differentiate coronary diseases from chest pain of other origin: a prospective emergency ward study of 278 consecutive patients admitted for chest pain. *J Intern Med*. 1990;227(3):165-172.
5. Tierney WM, Fitzgerald D, McHenry R, et al. Physicians' estimates of the probability of myocardial infarction in emergency room patients with chest pain. *Med Decis Making*. 1986;6(1):12-17.
6. Alpert JS, Thygesen K, Antman E, Bassand JP. Myocardial infarction redefined—a consensus document of the Joint European Society of Cardiology/American College of Cardiology Committee for the redefinition of myocardial infarction. *J Am Coll Cardiol*. 2000;36(3):959-969.
7. Kontos MC, Fritz LM, Anderson FP, Tatum JL, Ornato JP, Jesse RL. Impact of the troponin standard on the prevalence of acute myocardial infarction. *Am Heart J*. 2003;146(3):446-452.
8. Goodacre S, Locker T, Morris F, Campbell S. How useful are clinical features in the diagnosis of acute, undifferentiated chest pain? *Acad Emerg Med*. 2002;9(3):203-208.^a
9. Goodacre SW, Angelini K, Arnold J, Revill S, Morris F. Clinical predictors of acute coronary syndromes in patients with undifferentiated chest pain. *Q J Med*. 2003;96(12):893-898.^a
10. Henrikson CA, Howell EE, Bush DE, et al. Chest pain relief by nitroglycerin does not predict active coronary artery disease. *Ann Intern Med*. 2003;139(12):979-986.
11. Wang SJ, Ohno-Machado L, Fraser HSF, Kennedy RL. Using patient-reportable clinical history factors to predict myocardial infarction. *Comput Biol Med*. 2001;31(1):1-13.^a
12. Reilly BM, Evans AT, Schaider JJ, et al. Impact of a clinical decision rule on hospital triage of patients with suspected acute cardiac ischemia in the emergency department. *JAMA*. 2002;288(3):342-350.
13. Diamond GA, Forrester JS. Analysis of probability as an aid in the clinical diagnosis of coronary artery disease. *N Engl J Med*. 1979;300(24):1350-1358.
14. Antman EM, Anbe DT, Armstrong PW, et al, for the American College of Cardiology; American Heart Association Task Force on Practice Guidelines; Canadian Cardiovascular Society. ACC/AHA guidelines for the management of patients with ST-elevation myocardial infarction: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines (Committee to Revise the 1999 Guidelines for the Management of Patients with Acute Myocardial Infarction). *Circulation*. 2004;110(9):282-292.
15. Anderson JL, Adams CD, Antman EM, et al. ACC/AHA 2007 guidelines for the management of patients with unstable angina/non-ST-elevation myocardial infarction: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines (Writing Committee to Revise the 2002 Guidelines for the Management of Patients With Unstable Angina/Non-ST-Elevation Myocardial Infarction) [published correction in *J Am Coll Cardiol*. 2008;51(9):974]. *J Am Coll Cardiol*. 2007;50(7):e1-e157.

^aFor the Evidence to Support the Update for this topic, see <http://www.JAMAevidence.com>.

ACUTE MYOCARDIAL INFARCTION—MAKE THE DIAGNOSIS

PRIOR PROBABILITY

Approximately 25% of patients with symptoms suggesting ACI will prove to have an MI.

POPULATION FOR WHOM ACUTE MYOCARDIAL INFARCTION SHOULD BE CONSIDERED

Focus primarily on the symptoms associated with the presenting complaints, rather than the risk factors.

- Chest pain
- Shortness of breath
- Cardiac arrest
- Dizziness/weakness/syncope
- Abdominal pain

DETECTING THE LIKELIHOOD OF ACUTE MYOCARDIAL INFARCTION

The ECG is by far the most useful finding available at the patient's bedside. For patients with an abnormal ECG results suggesting acute MI (ST-segment elevation or Q waves, new conduction defects, diagnostic T-wave abnormalities), the symptoms and signs of MI become less important (Table 35-13). For patients with chest discomfort and normal or nondiagnostic

ECG results, some of the symptoms are diagnostically useful. Perhaps the most important finding for clinicians is the realization that a few of the important risk factors for coronary heart disease do not help in the acute setting for identifying patients with chest pain who are having an acute MI. The presence of diabetes, hypertension, and hyperlipidemia does identify patients at higher risk of coronary heart disease, but the presenting symptoms are more important for determining whether the current episode represents ACI.

The availability of the ACI-TIPI probability estimate requires integration of the computerized implementation protocol into an ECG reading. Consequently, physicians may not have access to the results. In the absence of the estimates, the multivariate models and the values in the table are the best estimates for identifying patients most likely to have ACI. However, it is crucial that clinicians understand that these variables have not been used to determine whether patients should be discharged from emergency care, observed, or admitted to rule out an MI. No single variable, in and of itself, has had consistently useful utility for ruling out an MI. The guidelines recommend a multivariate approach (<http://www.acc.org/qualityandscience/clinical/guidelines/unstable/incorporated/table5.htm>; accessed June 4, 2008). Once an ECG is obtained, the ACI-TIPI probability estimate given to the clinician is the approach with the best-demonstrated effect on clinical decisions and outcomes. Clinicians might use the multivariate model by Goodacre et al⁹ to quantify their overall estimates for patients with nondiagnostic ECG results and for whom the decision to rule out cardiac ischemia is less certain. Although these predictive models may have good measurement characteristics that could help clinicians, the requirement for programmable devices impedes widespread implementation.

Table 35-13 Multivariate and Univariate Predictors of Myocardial Infarction

MI		
	LR+ (95% CI) or Range	LR- (95% CI) or Range
Multivariate Predictors		
ACI-TIPI with clinical decision (n = 4 studies) ³	3.9-12	0.05-0.18
Wang logistic model (n = 1 study) ¹¹	This model provides a probability estimate for MI for patients with chest discomfort independent of the ECG result and coronary history.	
Goodacre et al ⁸ logistic model (n = 1 study)	This model provides a probability estimate for MI for patients with chest discomfort and a normal or non-diagnostic ECG result, no history of coronary heart disease with similar pain, and low suspicion of pulmonary embolus. The model should not be applied to other patient populations.	
Univariate Predictors		
Univariate findings for acute MI in patients with normal or nondiagnostic ECG results without known coronary heart disease with prolonged or recurrent chest pain typical of their angina		
Pain radiation to the shoulder OR both arms ⁸	4.1 (2.5-6.5)	0.68 (0.52-0.89)
Pain radiation to the right arm ⁹	3.8 (2.2-6.6)	0.86 (0.77-0.96)
Vomiting ⁹	3.5 (2.0-6.2)	0.87 (0.79-0.97)
Ex-smoker ⁹	2.5 (1.6-4.0)	0.85 (0.76-0.96)

Abbreviations: ACI-TIPI, Acute Cardiac Ischemia Time-Insensitive Predictive Instrument; CI, confidence interval; ECG, electrocardiogram; LR+, positive likelihood ratio; LR-, negative likelihood ratio; MI, myocardial infarction.

REFERENCE STANDARD TESTS

Either of the following criteria satisfies the diagnosis for an acute, evolving, or recent MI:

1. Typical increase and gradual decrease (troponin) or more rapid increase and decrease (CK-MB) of biochemical markers of myocardial necrosis with at least 1 of the following:
 - a. Ischemic symptoms
 - b. Development of pathologic Q waves on the ECG
 - c. ECG changes indicative of ischemia (ST-segment elevation or depression)
 - d. Coronary artery intervention (eg, coronary angioplasty)
2. Pathologic findings of an acute MI.

EVIDENCE TO SUPPORT THE UPDATE:

35

Myocardial Infarction

TITLE How Useful Are Clinical Features in the Diagnosis of Acute, Undifferentiated Chest Pain?

AUTHORS Goodacre S, Locker T, Morris F, Campbell S.

CITATION *Acad Emerg Med*. 2002;9(3):203-208.

QUESTION Do clinical features in clinically stable patients with nondiagnostic electrocardiograms (ECGs) identify those with acute myocardial infarction (AMI)?

DESIGN Prospective, consecutive patients meeting study criteria, with data collected independently of outcome.

SETTING British emergency department.

PATIENTS During a 16-month period, data were collected prospectively on a chest pain observation unit in a large, urban teaching hospital. Patients were excluded if they had ECG evidence of acute cardiac ischemia (ACI), known coronary heart disease with prolonged or recurrent chest pain typical of their angina, comorbid conditions or an alternate problem that required admission (eg, heart failure, pulmonary embolus), or minimal risk of coronary heart disease (eg, age < 25 years, chest discomfort related to trauma, chest wall pain reproduced by palpation in patients with no or few risk factors for coronary heart disease; Steve Goodacre, PhD, University of Sheffield, Sheffield, UK, written communication, November 2004). Among all emergency department patients with chest pain, only 25% were included in the study.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

A nurse with expertise in chest pain was trained to record symptoms.

MAIN OUTCOME MEASURES

An AMI was defined by World Health Organization (WHO) criteria. An acute coronary syndrome (ACS) was defined as a myocardial infarction (MI) at presentation *or* an increased concentration of cardiac troponin by 3 days, an early positive

exercise treadmill test result during the next 6 months, cardiac death, arrhythmia, or coronary revascularization within 6 months.

MAIN RESULTS

Of the 893 assessed patients, 57 met the study criteria for an ACS (9.1%); 34 patients had an AMI (3.8%), 15 had increased troponin levels without meeting the older WHO case definition for MI, and 78 additional patients had a subsequent early positive treadmill test result or arrhythmia. Overall, 88% of those classified as having ACS actually had an MI using current standards.

For AMI, the unadjusted odds ratios (ORs) with highest statistical significance were pain radiation to both arms (7.7), radiation to the shoulder (6.0), and exertional pain (3.1). For ACS, the unadjusted ORs with highest statistical significance were pain radiation to both arms (6.0), radiation to the shoulder (3.4), and exertional pain (2.5). All variables with diagnostic ORs with $P < .2$ were entered into a multivariate model for MI or ACS.

For diagnosing MI, the following variables were not useful: pain radiating to the throat, sharp/stabbing pain, crushing/gripping pain, heavy/pressing pain, pain duration, diaphoresis, and relief after taking nitroglycerin. Two variables, burning/indigestion pain (OR, 4.0) and nausea/vomiting (OR,

Table 35-14 Likelihood Ratios for Pain Pattern for Acute Myocardial Infarction or Acute Coronary Syndrome

Symptom	Diagnosis	LR+ (95% CI)	LR- (95% CI)
Pain radiation to shoulder or both arms	AMI	4.1 (2.5-6.5)	0.68 (0.52-0.89)
Exertional pain	AMI	2.3 (1.4-3.8)	0.76 (0.59-0.98)
Chest wall tenderness ^a	AMI	0.30 (0.08-1.1)	1.3 (1.1-1.4)
Exertional pain	ACS	2.1 (1.4-23)	0.82 (0.71-0.95)
Pain radiation to shoulder, left arm, or both arms	ACS	1.6 (1.2-2.0)	0.68 (0.53-0.87)

Abbreviations: ACS, acute coronary syndrome; AMI, acute myocardial infarction; CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

^aNote that chest wall tenderness makes AMI less likely, whereas the absence of chest wall tenderness makes AMI more likely.

1.8), appeared useful, with their diagnostic OR significantly different from 1, but the variables had no independent value in a multivariable model for predicting MI.

For diagnosing an ACS, the only significant variables in a multivariable model were pain radiating to the shoulder, left arm, or to both arms and exertional pain (Table 35-14).

CONCLUSIONS

LEVEL OF EVIDENCE

Level 3.

STRENGTHS A large number of variables were analyzed compared with the WHO standard for diagnosing an MI. The data allow some insight into how the performance of symptoms might change when troponin is considered as part of the case definition.

LIMITATIONS The data were timely when collected, but the case definition for AMI changed during the course of the study. The incident rate of MI would have increased 30% had troponin been included in the case definition for AMI. Most of the patients with ACS did have an AMI according to new standards, but 12% had something other than an AMI at presentation. Verification bias exists in that only patients admitted to the chest pain unit were included.

The study subjects were patients whose diagnosis was not obvious initially, leading to admission to the chest pain unit. Thus, the population selected is the most appropriate for answering the study's questions because the individual clinical symptoms were not used to identify patients for enrollment.

The change in case definition during the conduct of this study affords us the ability to make some inferences about the effect on the utility of clinical symptoms. The 2 variables that were independently important in a multivariate analysis appear less important when troponins are included in the case definition. Pain radiation to the shoulder or both arms has a diagnostic OR that decreases from 6.0 to 2.4 with current WHO diagnostic standards for MI; the diagnostic OR for exertional pain decreases from 3.1 to 2.5. The results need confirmation in a group of patients diagnosed with current standards.

Because of the changing definition, the data in these authors' second cohort provide more valid estimates of the likelihood ratios.¹

Reviewed by David L. Simel, MD, MHS

REFERENCE FOR THE EVIDENCE

- Goodacre SW, Angelini K, Arnold J, Revill S, Morris F. Clinical predictors of acute coronary syndromes in patients with undifferentiated chest pain. *Q J Med.* 2003;96(12):893-898.

TITLE Clinical Predictors of Acute Coronary Syndromes in Patients With Undifferentiated Chest Pain.

AUTHORS Goodacre SW, Angelini K, Arnold J, Revill S, Morris F.

CITATION *Q J Med.* 2003;96(12):893-898.

QUESTION Do any clinical predictors help identify patients with undifferentiated chest pain who are having an acute coronary syndrome (ACS)?

DESIGN Prospective, consecutive patients meeting study criteria, with data collected independently of outcome.

SETTING British emergency department.

PATIENTS During a 15-month period, data were collected as part of a randomized trial comparing a chest pain unit to usual care for patients with chest pain. Patients were excluded if they had electrocardiogram (ECG) evidence of an ACS, known coronary heart disease with prolonged or recurrent chest pain typical of their angina, comorbid conditions or an alternate problem that required admission (eg, heart failure, pulmonary embolus), or an obvious noncardiac cause of chest discomfort (eg, chest wall pain reproduced by palpation in patients with no or few risk factors for coronary heart disease).

Of the 6957 patients potentially eligible, 764 (11%) had an abnormal ECG result suggesting acute ischemia, 2402 (34.5%) had known coronary heart disease with prolonged or recurrent angina, 869 (12%) had comorbidities or other pathology requiring admission, and 1291 had obvious noncardiac chest pain (19%). This left 1631 (23%) with undifferentiated chest pain; 972 agreed to participate in the trial.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

A nurse with expertise in chest pain was trained to record symptoms from a list of a priori determined variables.

MAIN OUTCOME MEASURES

An ACS was defined by a troponin T level increased at 2-day follow-up, or a variety of prespecified events during the subsequent 30 days: cardiac death, nonfatal myocardial infarction (MI) using the current troponin-based standards,¹ new heart failure, life-threatening arrhythmia, or coronary revascularization.

MAIN RESULTS

An ACS was found in 77 patients (7.9%). Of those, 70 patients qualified because of an increased troponin level suggesting ischemia.

Table 35-15 Likelihood Ratios for Symptoms Found as Useful in a Multivariable Model

Symptom	LR+ (95% CI)	LR- (95% CI)
Radiation to right arm	3.8 (2.2-6.6)	0.86 (0.77-0.96)
Vomiting	3.5 (2.0-6.2)	0.87 (0.79-0.97)
Ex-smoker	2.5 (1.6-4.0)	0.85 (0.76-0.96)
Indigestion/burning pain	2.3 (1.5-3.5)	0.85 (0.74-0.96)
Male sex	1.5 (1.3-1.6)	0.24 (0.12-0.48)
Current smoker	1.4 (1.0-1.8)	0.83 (0.68-1.0)
Radiation to left arm	1.3 (0.93-1.8)	0.90 (0.76-1.1)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

The variables in Table 35-15 represent the unadjusted likelihood ratios for those found significant in a multivariate model. Variables that were significant by themselves but had no independent value in a multivariate model (diagnostic odds ratio not significantly different from 1) included pain radiation to the neck or jaw, aching/dull/heavy quality to pain, gripping/crushing quality to pain, right-sided chest pain, left-sided chest pain, chest wall tenderness, and diaphoresis. Diabetes, hypertension, hyperlipidemia, and a family history of coronary heart disease all had $P > .50$ and were not tested in the multivariate model.

A multivariate model was developed using the independent predictors:

$$\text{MI score} = 116 + 1.0 \times (\text{age}) + 23 \times (\text{male}) + 21 \times (\text{right arm pain}) + 18 \times (\text{ex-smoker}) + 11 \times (\text{left arm pain}) + 15 \times (\text{vomit}) + 15 \times (\text{smokes}) + 10 \times (\text{burning pain})$$

(Male = 1, female = 0. If symptom present, substitute 1; if negative or unknown, substitute 0.)

$$\text{MI probability} = [\exp^{(\text{score}/11)}] / [1 + \exp^{(\text{score}/11)}]$$

CONCLUSIONS

LEVEL OF EVIDENCE

Level 1.

STRENGTHS These data were prospectively collected from chest pain patients who did not initially have an obvious diagnosis. Current standards for diagnosis with troponin were used, and 91% of the patients with ACSs had ischemia.

LIMITATIONS It is impossible to know whether the findings would work better or worse had the results been reported for all patients with chest discomfort. Most excluded patients were excluded for findings not related to the clinical symptoms (eg, an abnormal ECG result, previous diagnosis of coronary heart disease with prolonged pain).

The authors observed that their population was younger (average age, 50 years) than most populations of patients with chest pain.

These data are important in that they used current standards of diagnosis with cardiac troponin levels. Clinicians must understand the study population before using the results—these patients had an uncertain diagnosis because

those with an abnormal ECG or with prolonged or recurrent chest pain typical of diagnosis before their previous angina were excluded. In addition, patients with obvious noncardiac chest pain or those requiring admission independent of their ACS were excluded. After these exclusions, the remaining patients were those for whom the clinician might be most reliant on the clinical symptoms, representing a common problem for emergency department physicians.

The data for indigestion/burning pain are counterintuitive in suggesting that the finding increases the likelihood for an acute MI. Astute clinicians will recognize that the study population did not include patients discharged from the emergency department who presented with indigestion/burning as the primary symptom, despite less important chest discomfort associated with their indigestion. Thus, the sensitivity and specificity of indigestion/burning pain might be quite different among all patients presenting with chest discomfort. In addition, indigestion/burning might have been a referral filter applied at the patient level in that patients presenting to the emergency department with a burning/indigestion type of pain likely represented those who could have self-medicated without relief or those with exceptionally severe discomfort.

The relative lack of importance for left arm pain radiation in comparison to right arm radiation also seems counterintuitive. Of the total patient population, there was no pain radiation in 38% and radiation to the left arm in 27%; only 6% of patients had right arm pain radiation. Most clinicians consider left arm pain radiation as a feature that suggests chest pain of cardiac origin. Patients may recognize left arm pain radiation as suggestive of an MI, making those experiencing any left arm pain more likely to come to the emergency department even when cardiac ischemia is an unlikely diagnosis (eg, musculoskeletal pain or cervical pain radiating to the left arm). There are 2 other possible explanations for the lesser importance of left arm pain in this population. First, it is possible that left arm pain occurs even more frequently in patients with obvious ACSs associated with ECG changes (these patients were excluded from this study). Second, in other studies, most patients with right arm pain also had bilateral arm pain radiation that would make left arm pain alone appear less important. Once the importance of left arm pain is used to identify patients with possible ACSs, the presence of left arm pain may no longer be independently useful in identifying those with MI vs those without MI.

Acknowledgment

Steven Goodacre kindly provided the results of the multivariate model and information about the clinical exclusion of patients with an obvious noncardiac cause for chest discomfort.

Reviewed by David L. Simel, MD, MHS

REFERENCE FOR THE EVIDENCE

- Alpert JS, Thygesen K, Antman E, Bassand JP. Myocardial infarction redefined—a consensus document of the Joint European Society of Cardiology/American College of Cardiology Committee for the redefinition of myocardial infarction. *J Am Coll Cardiol.* 2000;36(3):959-969.

TITLE Using Patient-Reportable Clinical History Factors to Predict Myocardial Infarction.

AUTHORS Wang SJ, Ohno-Machado L, Fraser HSF, Kennedy RL.

CITATION *Comp Biol Med.* 2001;31(1):1-13.

QUESTION Using only clinical factors, without electrocardiogram (ECG) data, can a logistic model be created that predicts myocardial infarction (MI)?

DESIGN The variables identified in a previous study¹ were collected in 2 patient populations. The details of whether the study included prospective, consecutive patients who had an independently applied reference standard are not provided.

SETTING Two British hospitals. The logistic model was developed on patient data from one hospital and then tested on patients from a second hospital.

PATIENTS The patient population consisted of patients with an MI prevalence of 22% in the first hospital and 31% in the hospital where the model was verified. Presumably, the clinicians suspected that all these patients had acute cardiac ischemia, but details about the patient population are not specified.

MAIN OUTCOME MEASURE

Accuracy (c-index) of the logistic model when evaluated in the validation sample from the second hospital.

MAIN RESULTS

The model had an accuracy of 84% in the validation set.

MI score = $-92 + 1.0 \times (\text{age}) + 17 \times (\text{diaphoresis}) + 14 \times (\text{nausea}) + 11 \times (\text{smokes}) + 11 \times (\text{left arm pain}) + 8 \times (\text{male}) - 44 \times (\text{pleuritic pain}) - 30 \times (\text{episodic pain}) - 15 \times (\text{sharp pain}) - 15 \times (\text{previous angina}) - 12 \times (\text{previous MI})$

(If symptom present, substitute 1; if symptom absent, substitute 0.)

$$\text{MI probability} = [\exp^{(\text{score}/15)}] / [1 + \exp^{(\text{score}/15)}]$$

An expert cardiologist picked the variables anticipated to be important in the logistic model. The variables identified by the cardiologist as important, but that were *not* independently valuable in a multivariable model included diabetes, hyperlipidemia, severe chest pain quality, retrosternal pain, left chest pain location, postural pain, pain that worsened, and pain that was worse than previous angina. The variables identified in variable selection by the computer that were not selected by the cardiologist were a sharp quality to the pain and the presence of nausea.

CONCLUSIONS

LEVEL OF EVIDENCE Level 4.

STRENGTHS These data could be important when they are applied to the appropriate patient population before the ECG is obtained.

LIMITATIONS There is no description of how the study population was obtained and the disease status verified. However, the incidence of MI suggests that the study included all patients with chest pain in the emergency department.

The data support the commonly held notion that chest pain with diaphoresis or nausea, especially when radiating to the left arm or in a smoker, increases the probability that the patient is having a MI.

One of the important findings in the study was the comparison of variables selected by the cardiologist to those remaining in the final model because of their statistical significance. The differences between variables selected by the cardiologist vs the computer highlight findings that might be inappropriately overweighted or underweighted by clinicians.

After age, the findings of diaphoresis, nausea, and left arm pain are the variables that increased the probability of MI the most. The variables that decrease the likelihood the most are pleuritic type pain and episodic pain. These data should be applied to patients who have not had ECGs. Thus, the model could be used at triage of the patient, but it requires validation in an appropriate population with current diagnostic standards.

The finding that a previous MI or angina decreases the probability of a current MI seems counterintuitive. However, if patients with a history of ischemic heart disease are more likely to use emergency services for any given episode of pain, then the proportion of visits for a new MI might be less than in those with no history.

Reviewed by David L. Simel, MD, MHS

REFERENCE FOR THE EVIDENCE

- Kennedy RL, Garrison RF, Burton AM, et al. An artificial neural network system for diagnosis of acute myocardial infarction (AMI) in the accident and emergency department: evaluation and comparison with serum myoglobin measurements. *Comput Methods Programs Biomed.* 1997;52 (2):93-103.

CHAPTER 36

Does This Woman Have Osteoporosis?

Amanda D. Green, MD

Cathleen S. Colón-Emeric, MD, MHSc

Lori Bastian, MD, MPH

Matthew T. Drake, MD, PhD

Kenneth W. Lyles, MD

CLINICAL SCENARIOS

CASE 1 You recommend screening densitometry to a healthy 64-year-old woman. She will have to drive 1 hour to the nearest testing center, and she does not believe that she needs the test. To further assess her risk, you note that she weighs 49 kg (108 lb). What can you tell this patient about her probability of osteoporosis?

CASE 2 A frail, 79-year-old woman is admitted to the hospital with a diverticular bleeding event. On examination, you observe that she has significant kyphosis. When she stands upright against a wall, she cannot touch the back of her head to the wall. You wonder whether she has vertebral fractures.

CASE 3 A 58-year-old woman presents for her annual examination. She experienced physiologic menopause 8 years ago but is asymptomatic and has no other risk factors for osteoporosis. On examination, you note that her rib-pelvis distance is 1 fingerbreadth. She tells you that she has developed a humped back. Should this patient be referred for densitometry?

WHY IS THE CLINICAL EXAMINATION IMPORTANT?

Osteoporosis causes 1.5 million fractures per year in the United States.¹ As the population continues to age, this number is expected to double by 2040.² Half of all postmenopausal women and 15% of white men older than 50 years will have an osteoporosis-related fracture in their lifetime, with 15% of those occurring in the hip. Pain, loss of independence, impaired ambulation, depression, and nursing home admission are common sequelae.³⁻⁸

In 1995, health care spending for osteoporotic fractures in the United States was \$13.8 billion and is estimated to be \$31 billion to \$62 billion by 2020.⁹ The US Preventive Services Task Force recommends that women 65 years of age or older be screened routinely for osteoporosis and women younger than 65 years be screened if they have risk factors.¹⁰ There are no current guidelines on when to screen healthy perimenopausal women, and few to no risk factors identified for men.

The physical examination may assist clinicians in preventing osteoporotic fractures in several ways. First, it may identify patients with low bone mineral density (BMD), in whom routine screening is not currently recommended or has not been completed. It may also identify patients at low risk of osteoporosis, in whom BMD testing is unnecessary. Although it is an imperfect indicator of fracture risk, BMD measurement is widely used both in randomized controlled trials and in clinical practice as the primary criterion for initiating osteoporosis therapies.

Second, the physical examination could identify patients with occult vertebral fracture. Two-thirds of vertebral fractures are clinically silent but are associated with a 2- to 3-fold increased risk of further fractures. Several osteoporosis therapies reduce the risk of further fractures in women with vertebral fractures, and the National Osteoporosis Foundation algorithm suggests that patients found to have vertebral fracture should be treated regardless of their BMD measurement.¹¹ Thus, the objective of this review was to identify clinical examination findings that improve the identification of patients with low BMD or occult vertebral fractures who would benefit from therapy or in whom further screening with BMD testing is unnecessary.

Case Definitions and Pathophysiology

Osteoporosis is a skeletal disorder characterized by compromised bone strength, predisposing a person to an increased risk of fracture. For this review, we used the World Health Organization's definition of osteoporosis, based on BMD that compares a patient's density to normative values for a population of 20- to 40-year-olds in terms of the number of deviations from the mean value. Osteoporotic bones have a density that is more than 2.5 SD below the mean (T score < -2.5). Osteopenic bones have a T score that is between -2.5 and -1 . Normal bones have a BMD T score of -1 or higher.¹²

Vertebral fractures are compression deformities that reduce vertebral body height by 20% or more on imaging studies; most of the articles included in this review used a semiquantitative technique to diagnose vertebral fractures on plain lateral radiographs of the spine. Spinal fractures are classified by the maximal percentage of vertebral body height loss as follows: grade 1, 20% to 24%; grade 2, 25% to 39%; and grade 3, 40% or more.¹³

The prevalence of osteoporosis in large population-based studies allows an estimation of the pretest probability in women of various ages. The prevalence of BMD-defined osteoporosis at the spine, wrist, or hip in white women in the United States by decade is as follows: for aged 50 to 59 years, 15%; 60 to 69 years, 22%; 70 to 79 years, 38%; and 80 years or older, 70%.¹⁴

Table 36-1 Prevalence of Vertebral Deformities in Women Aged 50 Years or Older¹⁷

Age, y	Vertebral Deformity, % ^a	
	≥ Grade 1	≥ Grade 2
50-54	10	4.7
55-59	12	6.6
60-64	12	8.9
65-69	17	12
70-74	30	21
75-79	33	29
80-84	56	49
85-89	49	47
≥ 90	75	75

^aGrade 1 or greater is equal to 20% or more vertebral body height loss; grade 2 or greater is equal to 25% or more vertebral body height loss.

For nonwhite women older than 50 years, the prevalence of BMD-defined osteoporosis in the Third National Health and Nutrition Examination Survey was reported as follows: non-Hispanic black women, 12%; Mexican Americans, 19%; and women in other ethnic groups, 28%.¹⁵ In special populations, the prevalence of osteoporosis can be much higher. For example, in residents of skilled nursing facilities who are older than 75 years, the prevalence of osteoporosis exceeds 50% for all the residents, regardless of race and sex.¹⁶

Occult vertebral fractures are also common and increase with age (Table 36-1). Grade 2 vertebral deformities are found in 6.6% of women aged 55 to 59 years and in 49% of women aged 80 to 84 years.¹⁷ Clinical characteristics or historical items that might increase a clinician's pretest probability of osteoporosis or vertebral fracture include older age, low activity level, family history, hypogonadism (men), and exposure to glucocorticoids and alcohol. The pretest probability threshold for testing BMD depends on the anticipated benefit of treatment for an individual patient and the patient's desire for treatment.

The pathophysiology of osteoporosis is related to physical examination findings in several ways. The loading or mechanical forces on bone tend to increase bone formation and bone mass through osteoblast stimulation. Thus, increasing body weight and muscle strength are inversely related to osteoporosis. Type I collagen is a major constituent of both bone and skin that is reduced with advancing age and low estrogen levels.¹⁸⁻²⁰ Skinfold thickness may therefore reflect skeletal collagen content. Similarly, tooth loss is influenced by mandibular alveolar bone quality and may provide an easily observed marker of bone health in the rest of the skeleton.

The sequelae of clinically occult vertebral fractures can also lead to physical examination findings that may become apparent before a symptomatic fracture occurs. Height loss resulting from vertebral compression fractures can be measured in the clinic over time or with the patient's recalled maximal adult height. Vertebral fractures affect height but not arm span, so arm span–height differentials may identify individuals with occult vertebral fractures.²¹ Thoracic kyphosis can result from anterior compression fractures in the thoracic spine ("dowager's hump"). Kyphosis can be measured on physical examination with a curved ruler such as an architect's rule or by measuring the wall–occiput distance. The wall–occiput distance describes the difference between the wall and the patient's occiput when he or she stands straight with heels and back against the wall. Lumbar fractures also result in decreased rib–pelvis distance that can be measured in fingerbreadths on examination.

How to Elicit the Relevant Signs

Data for several physical examination signs are included in this review. Weight and height are routinely measured in the clinical setting. Aside from clinic notes, height change can be documented from alternate sources (such as a driver's license) or from the patient's memory of height at age 25 years.²²⁻²⁴ Several studies have shown good to excellent correlation between elderly patients' recalled maximal height and

previous health records.²⁵⁻²⁷ A stadiometer (an upright bar marked with a height scale with a sliding notch to designate height) is the most accurate method of height measurement.

Arm span–height differential is determined by subtracting a patient's height in centimeters from the arm span in centimeters measured with arms at a 90-degree angle from the trunk. The arm span is the distance between the tips of the middle fingers while the patient faces forward with the arms fully extended and palms facing forward.

Measurements of thoracic kyphosis can be made indirectly on radiographs but can also be directly measured by applying an architect's semiflexible rule, called a flexicurve, to the patient's back.²⁸ The flexicurve is a device that can be bent in 1 plane only and retains its shape after application to the curvature of the back between the C7 spinous process and S2 spinous process. The outline is traced on paper, and the maximal angle is measured with calipers or a ruler.²⁹ The kyphosis index is the ratio of thoracic curvature to the length of the upper back and is calculated as 100 times the maximum horizontal distance divided by the vertical length of the upper back curve. Flexicurve measurements, although painless, inexpensive, and safe, are time consuming.³⁰

Another measure that quantitates the degree of kyphosis is wall–occiput distance. It is measured while the patient stands straight with his or her back against the wall and heels touching the wall (**Figure 36-1**). While the head faces forward so that an imaginary line connecting the lateral corner of the eye to the superior junction of the auricle of the ear is parallel to the floor, the distance between the occipital prominence and the wall is quantified with a tape measure.³¹ For the purpose of this review, the inability to touch the wall with the back of the head is a positive finding.

Rib–pelvis distance is a measure of lumbar fracture. The patient stands erect with arms outstretched at 90 degrees. The examiner stands behind the patient and inserts his or her fingers into the space between the inferior margin of the ribs and the superior surface of the pelvis in the midaxillary line. The rib–pelvis distance is the closest whole number of fingerbreadths between these structures.³²

Skinfold thickness is measured at the back of the hand with calipers.¹⁸⁻²⁰ The back of the hand is a convenient site for measurement in the clinic.¹⁹ The fourth metacarpal longitudinal fold site was used in the studies of skinfold thickness included in this review.

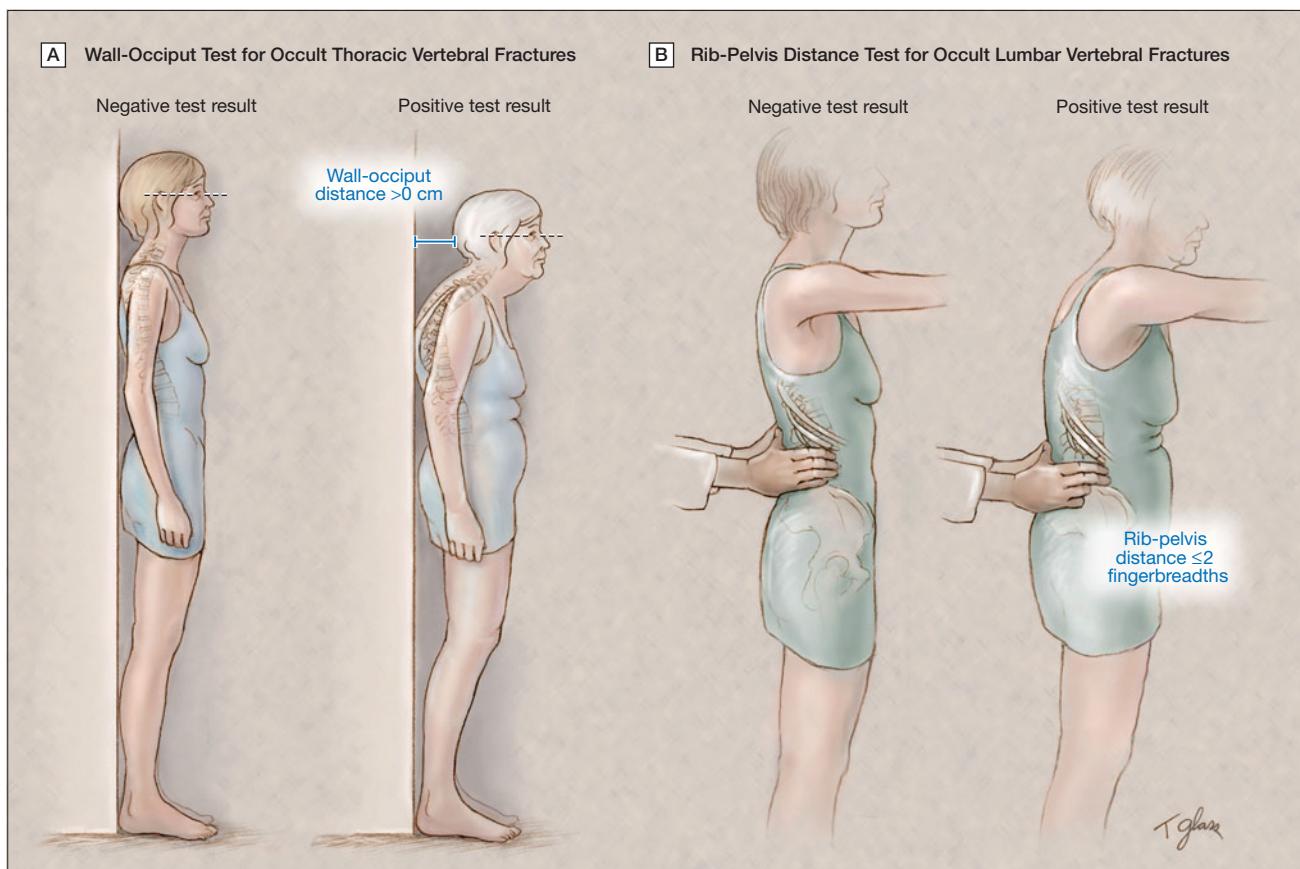


Figure 36-1 Physical Examination Tests for Detection of Occult Vertebral Fractures

A, Wall–occiput test is used to detect occult thoracic vertebral fractures. A positive test result in this review is defined as being unable to touch the wall with the occiput when standing with the back and heels against the wall and the head positioned such that an imaginary line from the lateral corner of the eye to the superior junction of the auricle is parallel to the floor. B, Rib–pelvis distance test is used to detect occult lumbar vertebral fractures. A positive test result is defined as a distance of less than or equal to 2 fingerbreadths between the inferior margin of the ribs and the superior surface of the pelvis in the midaxillary line.

Hand grip strength is measured using a small hydraulic hand grip or isometric dynamometer and is defined as the maximal force recorded while the patient squeezes the device with arms straight to the side.^{33,34}

METHODS

We searched MEDLINE for articles from 1966 through August 2004, with a search strategy similar to that used by other authors in this series.³⁵ We used several National Library of Medicine Medical Subject Headings to encompass osteopenia, osteoporosis, and spinal fracture disease states: “exp osteoporosis,” “exp spinal fracture,” “exp metabolic bone disease” (for osteopenia), and “exp bone density.” The MEDLINE search was supplemented with a manual review of the bibliographies of all identified articles, additional review articles including recent osteoporosis guidelines, 4 clinical skills textbooks,³⁶⁻³⁹ and contact with experts in the field. Two authors (A.D.G. and M.T.D.) independently executed the MEDLINE search strategy and reviewed titles and abstracts from the search results. Two authors (A.D.G. and C.S.C.-E.) then independently reviewed and extracted data from articles or abstracts identified as relevant. We contacted authors for original data when articles reported data on the precision of signs in diagnosing osteoporosis or spinal fracture but did not include enough information to calculate likelihood ratios (LRs).

We included studies in our review if they included original data on the accuracy or precision of the medical history or physical examination in diagnosing osteoporosis, osteopenia, or spinal fracture. We required that the gold standard comparison for the clinical examination parameters be bone densitometry at any site or documented vertebral fracture using either a semiquantitative technique or vertebral morphometry. When BMD values were reported directly, the corresponding T score was obtained with sex-appropriate tables provided by the manufacturer of the densitometer used in the study. Articles were excluded if they contained insufficient data to allow calculation of LRs. We included in our tables and results only the physical examination parameters that are feasible to perform in a clinical setting.

Quality Assessment of Included Articles

Two authors (A.D.G. and C.S.C.-E.) independently assessed the methodologic quality of included articles using criteria adapted from other authors in this series.⁴⁰ Level 1 evidence classifies articles that were independent (neither the test result nor the gold standard result was used to select patients for the study), studied consecutive patients representative of a population for which the test is likely to be used, were blinded, measured the gold standard (BMD measurement or documented fracture) in all patients, and included at least 100 study participants. Level 2 evidence met criteria for level 1 evidence, but fewer than 100 patients were studied. Level 3 evidence was the same as level 2 evidence, but the population was nonconsecutive or nonrepresentative. Studies of lower

levels of evidence were excluded. Disagreements were resolved by discussion and consensus.

Data Analysis

We used raw data from reported studies that met our inclusion criteria to calculate values and 95% confidence intervals for sensitivity, specificity, and positive likelihood ratio (LR+) and negative likelihood ratio (LR-), using SAS statistical software, version 8.0 (SAS Institute Inc, Cary, North Carolina).

RESULTS

Study Characteristics

We identified 246 articles with our search strategy and an additional 79 from reference lists and expert consultation. Fourteen studies met inclusion criteria and were identified for final review (Tables 36-2 and 36-3).

Precision

Table 36-4 lists reported precision estimates for the physical examination maneuvers. Interrater reliability was not reported for studies of height and weight included in this review. Differences in sensitivity and specificity for the same maneuver across different studies could be related to examiner differences that were not reported.

Diagnostic Accuracy

The most clinically relevant cut points and their associated LRs for the physical examination maneuvers are listed in Table 36-5 for osteoporosis and Table 36-6 for vertebral fracture. In general, the patient populations were women, with most patients from osteoporosis clinics or older than 65 years. Translating these results to younger women might yield error that is difficult to quantify. Because many of the examination findings may be measuring similar or identical physiologic phenomena, we do not recommend using the LRs in series.

For postmenopausal women, prediction rules using osteoporosis risk factors, such as the Simple Calculated Osteoporosis Risk Estimation⁴¹ or the Osteoporosis Risk Assessment Instrument,⁴² have some predictive value in selected populations (Table 36-7).^{11,43-46} Variables included in these prediction rules include age, weight, and race, which overlap with the clinical examination. An exhaustive review of prediction rules for the diagnosis of osteoporosis or fracture was not attempted in this study because reviews already exist in the literature.⁴² Although the LR+ of the prediction rules is not clinically informative (1.2-1.7), the LR- is far superior to the physical examination maneuvers listed here (0.02-0.3), making prediction rules much more useful for ruling out osteoporosis or fracture. Thus, clinical prediction rules are the most useful means of identifying women who are at low risk of fracture, in whom BMD screening can safely be deferred.

Table 36-2 Studies Used to Determine the Accuracy of Clinical Examination for Diagnosing Osteoporosis

Source	Setting and Country	Methodologic Quality ^a	Prevalence of Osteoporosis, %	Inclusion Criteria	No. of Patients	Mean Age, y	Diagnosis Used
Height Loss							
Sanila et al ²³	Outpatients, Finland	Level 3	34	Women aged 55-70 y with rheumatoid arthritis, able to walk	61	62	BMD-diagnosed osteoporosis lumbar BMD < 0.9 on Lunar machine
Dargent-Molina et al ⁴⁷	Volunteers for prospective, multicenter trial, France (EPIDOS)	Level 1	50	White women aged ≥ 75 y, general population, without past fractures	4638	80	BMD-diagnosed osteoporosis T score < -3.5 SD
Weight							
Michaelsson et al ⁴⁴	Outpatients, Sweden	Level 1	4	Random sample of women aged 28-74 y, no exclusions	175	51	BMD-diagnosed osteoporosis T score < -2.5
Dargent-Molina et al ⁴⁷	Volunteers for prospective, multicenter trial, France (EPIDOS)	Level 1	50	White women aged ≥ 75 y, general population, without past fractures	4638	80	BMD-diagnosed osteoporosis T score < -3.5 SD
Bedogni et al ⁵¹	Community, Italy	Level 1	8	Women aged ≥ 18 y without disease	1873	Not reported (range, 49-77)	BMD-diagnosed osteoporosis T score < -2.5
Kyphosis							
Ettinger et al ²⁸	Outpatients, California	Level 1	10	Consecutive sample of women aged 65-91 y	610	73 (range, 72-91)	BMD-diagnosed osteoporosis T score < -2.5
Self-reported Humped Back							
Kantorek et al ⁵³	Outpatients, Ohio	Level 1	10	White women aged ≥ 18 y referred for bone density scan	2577	60	BMD-diagnosed osteoporosis at the hip T score < -2.5
Grip Strength							
Di Monaco et al ³³	Outpatients, Italy	Level 3	34	Consecutive postmenopausal, white female volunteers	102	63	BMD-diagnosed osteoporosis T score < -2.5
Foley et al ³⁴	Outpatients, Ohio	Level 1	18	Older, independent adults in the community	73	71	BMD-diagnosed osteoporosis T score < -2.5
Dargent-Molina et al ⁴⁷	Volunteers for prospective, multicenter trial, France (EPIDOS)	Level 1	50	White women aged ≥ 75 y, general population, without past fractures	4638	80	BMD-diagnosed osteoporosis T score < -3.5 SD
Hand Skinfold							
Orme and Belchetz ²⁰	Outpatients, California	Level 3	63	Consecutive women in osteoporosis clinic	225	59	BMD-diagnosed osteoporosis T score < -2.0
Tooth Count							
Earnshaw et al ⁶⁰	Outpatients in multicenter alendronate trial, United Kingdom, United States, and Denmark	Level 1	33	White postmenopausal women aged 45-59 y	1365	53	BMD-diagnosed osteoporosis T score < -2.5
Inagaki et al ⁶⁴	Outpatients, Japan	Level 1	11.5	Community women	190	Not reported (range, 31-79)	BMD-diagnosed osteoporosis quartiles of BMD reported according to aluminum standard (results calculated in current report using lowest quartile)

Abbreviations: BMD, bone mineral density; EPIDOS, the European Patient Information and Document Service.

^aSee Table 1-7 for a description of Evidence Levels.

Height Loss

Three studies of postmenopausal women using recalled heights found an association between height loss and vertebral fractures, with 2 of the studies including enough data to calculate LRs (Table 36-5).^{23,47,48} In the first study, a height loss of more than 3 cm was useful in classifying patients with and without low BMD (LR+, 3.2; LR-, 0.4).²³ However, the study population was nonconsecutive female patients with rheumatoid arthritis. In a study of women in the general population, Dargent-Molina et al⁴⁷ did not find a strong association between height loss of more than 3 cm and osteoporosis (LR+,

1.1; LR-, 0.6). The third study, based on 13732 women in the Fracture Intervention Trial, reported that a self-reported height loss greater than 4 cm since age 25 years was associated with an odds ratio (OR) of 2.8 for vertebral fractures.⁴⁸ Thus, although height loss is a potentially useful examination tool, the generalizability of this measure is uncertain.

Arm Span–Height Difference

Versluis et al²¹ reported that with age, height declined at twice the rate of arm span. The mean difference in arm span and height was 1.4 cm in women aged 55 to 59 years and

Table 36-3 Studies Used to Determine the Accuracy of Clinical Examination for Diagnosing Spinal Fracture

Source	Setting and Country	Methodologic Quality	Prevalence of Fracture, %	Inclusion Criteria	No. of Patients	Mean Age, y	Diagnosis Used
Arm Span–Height Difference							
Versluis et al ²¹	General practice, The Netherlands	Level 1	3.4 if aged 55–59 y, 21.9 if aged 80–84 y	White women, aged 55–84 y, healthy, in general practices	449	67.6	Vertebral fractures by morphometry
Wang et al ⁵⁰	Osteoporosis clinic, Australia	Level 1	26 in men, not reported in women	White male and female healthy volunteers aged 18–92 y compared with consecutive osteoporosis clinic patients aged 45–90 y	480	63 for women, 66 for men	Vertebral fractures by morphometry
Wall–Occiput Distance							
Siminoski et al ³¹	Outpatients, Canada	Abstract only	29	Women aged > 18 y referred to osteoporosis clinic	216	53	Thoracic vertebral fractures by morphometry
Rib–Pelvis Distance							
Siminoski et al ³²	Outpatients, Canada	Level 1	14	Consecutive women in osteoporosis clinic	781	56.8	Lumbar vertebral fractures by morphometry

Table 36-4 Precision Data Reported in the Studies Used in the Review

Source	Precision Estimate Used	Precision Estimate (95% CI)
Height Loss		
Sanila et al ²³	Coefficient of repeatability	2.3 cm For height, with range of 0.4 mm to 0.5 cm for tape measure positions
Arm Span–Height Difference		
Versluis et al ²¹	Intraobserver mean differences	Height, 1 mm (−2.3 to 2.5 mm); arm span, 0.9 mm (−2.5 to 4.3 mm)
	Interobserver mean differences	Height, −1.6 mm (−3.2 to 0.1 mm); arm span, −4.6 mm (−7.7 to −1.5 mm)
Kyphosis		
Ettinger et al ²⁸	Coefficient of variation (2 independent technicians)	13% For kyphosis index
Wall–Occiput Distance		
Siminoski et al ³¹	Not reported	Single examiner measured 3 times
Rib–Pelvis Distance		
Di Monaco et al ³³	Interobserver κ	$\kappa = 0.87$ For cutoff of 2 finger breadths or less
Grip Strength		
Di Monaco et al ³³	Coefficient of variation	3%
Hand Skinfold		
Orme and Belchetz ²⁰	Mean of 3 measurements	Reproducible to within 0.2 mm

Abbreviation: CI, confidence interval.

increased to 3.2 cm in women aged 80 to 84 years. Finding an arm span–height difference of 5 cm or greater yielded an LR+ of 1.6 and an LR− of 0.8 for spinal fracture based on these data (Table 36-6). Verhaar et al⁴⁹ reported that an arm span–height difference cutoff of 3 cm resulted in a sensitivity of 58% and a specificity of 56% for BMD-diagnosed osteoporosis, for an LR+ of 1.3. Wang et al⁵⁰ found no association between arm span and vertebral fractures in both men and women (LR+ for men, 1.0; LR+ for women, 0.9). We con-

clude that the arm span–height difference does not predict vertebral deformities or BMD-diagnosed osteoporosis.

Weight

For women, the relationship between both low weight and body mass index (BMI) and osteoporosis has been consistently reported.⁴⁴ In cohort studies examining clinical risk factors in women, weight lower than 70 kg (154 lb) is the sin-

Table 36-5 Clinical Signs and Symptoms in the Diagnosis of Osteoporosis

Source	Cutoff Values	Sensitivity, %	Specificity, %	LR+ (95% CI)	LR− (95% CI)
Height Loss					
Dargent-Molina et al ⁴⁷	>3 cm	92	13	1.1 (1.0-1.1)	0.6 (0.4-0.9)
Sanila et al ²³	>3 cm	68	72	3.2 (1.7-6.1)	0.4 (0.2-0.7)
Weight					
Dargent-Molina et al ⁴⁷	<60 kg	82	56	1.9 (1.8-2.0)	0.3 (0.3-0.4)
Bedogni et al ⁵¹	<51 kg	22	97	7.3 (5.0-10.8)	0.8 (0.7-0.9)
Michaelsson et al ⁴⁴	<60 kg			3.6 (2.2-56)	
	60-70 kg			0.3 (0-19)	
	>70 kg			0.2 (0-2.5)	
Kyphosis					
Ettinger et al ²⁸	25	92	3.1 (1.8-5.3)	0.8 (0.7-1.0)	
Self-reported Humped Back					
Kantor et al ⁵³	20.6	97	3.0 (2.2-4.1)	0.85 (0.8-0.9)	
Grip Strength					
Foley et al ³⁴	<40 lb	31	88	2.6 (0.9-7.5)	0.8 (0.5-1.1)
	<60 lb	91	27	1.3 (1.0-1.6)	0.3 (0.1-2.2)
Dargent-Molina et al ⁴⁷	<59 kPa	84	27	1.2 (1.1-1.2)	0.6 (0.5-0.7)
	<44 kPa	41	76	1.7 (1.5-1.9)	0.8 (0.7-0.9)
Di Monaco et al ³³	<20 kg	88	41	1.5 (1.0-2.1)	0.3 (0.1-0.6)
Hand Skinfold					
Orme and Belchetz ²⁰	<2.1 mm	93	20	1.2 (1.0-1.3)	0.4 (0.2-0.8)
Tooth Count					
Earnshaw et al ⁶⁰	<22 teeth	30	70	1.0 (0.8-1.2)	1.0 (0.9-1.1)
Inagaki et al ⁶⁴	<20 teeth	27	92	3.4 (1.4-8.0)	0.8 (0.6-1.0)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR−, negative likelihood ratio.

Table 36-6 Clinical Signs and Symptoms in the Diagnosis of Spinal Fracture

Source	Cutoff Values	Sensitivity, %	Specificity, %	LR+ (95% CI)	LR− (95% CI)
Arm Span–Height Difference					
Versluis et al ²¹	>5 cm	39	76	1.6 (1.1-2)	0 (0.6-1.0)
Wang et al ⁵⁰	>6.6 cm for men	62	37	1.0 (0.7-1.4)	1.0 (0.6-1.7)
	>2.5 cm for women	48	48	0.9 (0.7-1.2)	1.1 (0.8-1.4)
Wall-Occiput Distance					
Siminoski et al ³²	>0 cm	88	46	3.8 (2.9-5.1)	0.6 (0.5-0.7)
Rib-Pelvis Distance					
Siminoski et al ³²	≤2 Finger breadths	88	46	3.8 (2.9-5.1)	0.6 (0.5-0.7)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR−, negative likelihood ratio.

gle best predictor of low BMD^{11,45,46} and is an important variable in 4 of the 5 prediction rules reviewed here. Bedogni et al⁵¹ reported that body weight allowed a better classification of BMD than did BMI, with women weighing fewer than 51 kg having a much greater risk for osteoporosis than do women weighing more (LR+, 7.3; Table 36-5).

The cross-sectional survey by Michaelsson et al⁴⁴ demonstrated that body weight was the best predictor of BMD among measures of body size in women. In this study,

women weighing fewer than 60 kg had a greater risk for osteoporosis than women who weighed more (LR+, 3.6). Women weighing 60 to 70 kg or more than 70 kg had a lower risk for osteoporosis (LR+, 0.3, and LR+, 0.2, respectively). Study limitations included a 20% participation rate and a low prevalence of osteoporosis.

Dargent-Molina et al⁴⁷ found current body weight to be the strongest predictor of very low bone mass (defined as a T score < -3.5 SD). When BMD was measured in the 50% of

Table 36-7 Selection Criteria and Decision Rules Reported for Bone Mineral Density Testing Among Postmenopausal Women Considering Treatment^{11,41-45a}

Guideline/Rule	Selection Cut Point	Scoring System ^b
Simple Calculated Osteoporosis Risk Estimation ⁴¹	Score ≥ 6 LR+, 1.2 LR-, 0.02	Not black = 5 points Rheumatoid arthritis = 4 points History of minimal trauma fracture after age 45 y = 4 points for each fracture of the wrist, hip, or rib (maximum, 12 points) Never used estrogen therapy = 1 point 3 × first digit of age in years = ____ points – 1 × weight in pounds divided by 10 (truncated to integer) = ____ points
Osteoporosis Risk Assessment Instrument ⁴²	Score ≥ 9 LR+, 1.4 LR-, 0.1	Age: ≥75 y = 15 points 65-74 y = 9 points 55-64 y = 5 points Weight: <60 kg = 9 points 60-69.9 kg = 3 points No current estrogen use = 2 points
National Osteoporosis Foundation ¹¹	Score ≥ 1 LR+, 1.2 LR-, 0.2	Age ≥ 65 y = 1 point Weight < 57.6 kg = 1 point History of minimal trauma fracture after age 40 y = 1 point Family history of fracture = 1 point Current cigarette smoking = 1 point
Age, body size, no estrogen ⁴³	Score ≥ 2 LR+, 1.6 LR-, 0.3	Age > 65 y = 1 point Weight < 63.5 kg = 1 point Never used oral contraceptives or estrogen therapy for ≥ 6 mo = 1 point
Dubbo Osteoporosis Epidemiology Study ⁴⁴	Score > 10 LR+, 1.7 LR-, 0.3	Age: <70 = 1 point 70-79 = 2 points 80-84 = 3 points 85-90 = 4 points 90+ = 16 points Weight: <55 kg = 1 point 55-64 = 2 points 65-69 = 3 points 70-74 = 4 points 75-79 = 6 points Previous fracture: Yes = 2 points No = 1 point

Abbreviations: LR+, positive likelihood ratio; LR-, negative likelihood ratio.

^aThe LR+ and LR- are for patients with findings at or above the threshold score (LR+) or below the threshold score (LR-). Diagnosis of osteoporosis was based on T scores less than -2.5 for all rules.

^bFor each new guideline/rule, sum up the total points to get the score.

women who weighed the least (< 59 kg), the LR+ was 1.9 and the LR- was 0.3.

Although all of the studies that met our inclusion criteria were samples of women, other studies that were excluded because they reported only regression analysis data found similar associations between BMD at all sites and weight in both men and women, with weight having a similar influence in each sex.^{11,51,52}

Thus, body weight lower than 59 kg appears to be a simple and reasonably sensitive but nonspecific measure for selecting women for further diagnostic testing. Heavier patients have a lower likelihood of osteoporosis. However, osteoporosis cannot be ruled out according to weight greater than 59 kg alone because of the broad range of LRs across the 3 studies (Table 36-5).

Kyphosis

Flexicurve measurements in women were converted into kyphosis index values by Ettinger et al,²⁸ with the highest decile of kyphosis index used to classify patients as kyphotic. Ettinger et al²⁸ reported that kyphosis was associated with reduced BMD and significant height loss. The presence of kyphosis was specific though not sensitive for osteoporosis (LR+, 3.1; LR-, 0.8). It is not clear whether the clinician's simple observation of kyphosis without sophisticated measurements would yield the same result (Table 36-5).

Self-reported humped back was reported by Kantor et al⁵³ to be highly specific for hip osteoporosis in more than 2000 women referred for densitometry, with an LR+ of 3.0. The absence of self-reported humped back is not useful (LR-, 0.85; Table 36-5).

Wall-Occiput Distance

Siminoski et al³¹ reported in abstract form that a kyphosis angle greater than 43 degrees or wall-occiput distance greater than 7 cm in women rules in a thoracic fracture with a high degree of accuracy, and a kyphosis angle less than 20 degrees or wall-occiput distance of 0 cm reduces the chance of thoracic fracture but does not reliably rule it out. The 0-cm cut-off seems most pragmatic, with an LR+ of 4.6 for thoracic fracture when a patient cannot place the back of her head to the wall (Figure 36-1 and Table 36-6). In a sample size of 60 elderly women, however, Balzini et al⁵⁴ did not find a relationship between wall-occiput distance and vertebral fractures (data were not presented for calculating LRs).

Rib-Pelvis Distance

Rib-pelvis distance of less than or equal to 2 fingerbreadths was calculated to have an LR+ of 3.8 and an LR- of 0.6 for detecting occult lumbar fractures (Table 36-6).³² Adjusting for patient height does not affect the operating characteristics of this test and is unnecessary. The LRs for vertebral fracture in a woman with 0 and 4 fingerbreadths of rib-pelvis distance are 12 and 0.1, respectively. Thus, a low rib-pelvis distance may increase the posttest probability of lumbar fracture to a level at which further testing is warranted.

Grip Strength

Of the common measures of muscle strength, grip strength is most feasible to evaluate in the typical primary care clinic. Di Monaco et al³³ reported a positive association between grip strength and distal radius BMD in postmenopausal women in multiple regression analysis adjusted for age, years since menopause, years of ovarian activity, body height, body weight, BMI, and calcium and alcohol dietary intake, with an LR+ of 1.5 (Table 36-5).

Foley et al³⁴ examined the relationship between hand grip strength and femur BMD, with the goal of canceling out the effects of other anthropometric data, and did not find a relationship between grip strength and proximal femur BMD for men. In women, it was thought that weight was related both to grip strength and femur BMD, with an LR+ of 1.3 for osteoporosis when a cutoff of less than 27.2 kg (60 lb) on the dynamometer was used.

Several other studies reported a positive association between grip strength and BMD, although reported data were not sufficient to calculate LRs.⁵⁵⁻⁵⁹ Overall, grip strength has insufficient sensitivity and inconsistent results for specificity.

Hand Skinfold

Orme and Belchetz²⁰ studied the skinfold thickness in consecutive women in an osteoporosis clinic compared with normal, younger control women and reported ORs for a range of skinfold thickness of 1.5 to 2.1. These ORs corresponded to an LR+ of 1.2 and an LR- of 0.4 (Table 36-5). Although simple to perform, skinfold thickness does not appear to be useful in the diagnosis of osteoporosis.

Tooth Count

Several studies have not shown a relationship between tooth loss and osteoporosis,⁶⁰⁻⁶³ but inclusion of younger patients may have limited their ability to detect an association.⁶⁴ It is not clear whether population studies reveal women with poor dental hygiene and tooth loss or tooth loss from osteoporosis.

Inagaki et al⁶⁴ reported that among postmenopausal women, the proportion of women with fewer than 20 teeth increased from 7% in the normal BMD group to 32% in the very low BMD group. The age-adjusted odds of having fewer than 20 teeth were significantly greater among women in the very low BMD group compared with the normal BMD group. The LR+ for having very low BMD if fewer than 20 teeth are counted is 3.4, but choosing a threshold of fewer than 22 teeth provides no additional clinical information (Table 36-5).⁶⁰

In a retrospective study, Astrom et al⁶⁵ found that elderly women with the least remaining teeth had twice the risk of hip fracture compared with women with the most teeth. For men, the risk was more than 3-fold. Unfortunately, the cut point number of teeth dividing the patients was not provided. May et al⁶⁶ found an association between self-reported tooth loss and BMD of the hip and spine using bone densitometry in older men that was independent of age, BMI, and

Box 36-1 Physical Examination Maneuvers Suggesting Presence of Osteoporosis or Spinal Fracture**WALL-OCCIPUT DISTANCE**

Inability to touch occiput to the wall when standing with back and heels to the wall

WEIGHT

Fewer than 51 kg

RIB-PELVIS DISTANCE

Fewer than 2 fingerbreadths between the inferior margin of the ribs and the superior surface of the pelvis in the midaxillary line

TOOTH COUNT

Fewer than 20 teeth

SELF-REPORTED HUMPED BACK

Patient report that back has become humped

cigarette use. Other population-based studies reviewed demonstrated variable positive correlations between tooth counts and BMD.⁶⁷⁻⁷² Overall, tooth counts are easy to do, and fewer than 20 teeth can reasonably lead the clinician to screen further for osteoporosis.

CLINICAL SCENARIOS—RESOLUTIONS

CASE 1 The reluctant 64-year-old woman has a pretest probability of approximately 22% for osteoporosis at any site (Table 36-1). Her low weight (< 51 kg) has an LR of 7.3, thus increasing her posttest probability of osteoporosis to 67%. She decides that this level of risk makes the drive to the testing center worthwhile.

CASE 2 The prevalence of grade 2 or grade 3 vertebral deformities in women aged 75 to 79 years is approximately 29%.⁷³ The LR+ of a positive wall-occiput maneuver is 4.6, resulting in a posttest probability of 65%. This 78-year-old patient is likely to have vertebral fractures. If spine radiographs confirm the presence of vertebral fractures, then she should be considered for osteoporosis therapy. BMD testing to confirm osteoporosis is not required but may help guide therapeutic decisions.

CASE 3 Although the 58-year-old woman does not meet current screening guidelines for dual-energy x-ray absorptiometry, the low rib-pelvis distance detected on her physical examination increases the probability that she already has occult vertebral fracture from 3.4% to 12%.⁷⁵ Her self-reported humped back increases the probability that she has osteoporosis from 15% to 37%, prompting early assessment of her bone density.

THE BOTTOM LINE

No single physical examination finding or combination of findings is sufficient to rule in osteoporosis or spinal fracture

without further testing. The risk factor prediction rules for osteoporosis quoted in this article have more informative negative LRs than any of the physical findings and may reduce the need for testing in low-risk women. Several convenient examination maneuvers, including low body weight (< 51 kg), inability to place the back of the head against a wall when standing upright, low tooth count, self-reported humped back, and rib-pelvis distance, can significantly increase the likelihood of osteoporosis or spinal fracture and identify additional women who would benefit from earlier screening. **Box 36-1**.

Although the major osteoporosis clinical focus has been on women, the hip fracture incidence in 80-year-old men is similar to that in 75-year-old women.⁷³ A review of male osteoporosis suggests that the risk factors for men are the same (eg, BMD and body weight), although the level of risk is different from that for women.⁷⁴ Because osteoporosis develops at a later age in men, meaningful research is needed to determine whether the examination findings have similar properties in men or whether there is an age at which men should be screened for BMD similar to the recommendations for women.

Author Affiliations at the Time of the Original Publication

Ambulatory Care Service (Drs Green and Bastian) and Geriatrics Research Education and Clinical Center (Drs Colón-Emeric and Lyles), Durham Veterans Affairs Medical Center, and Center for the Study of Aging and Human Development (Drs Colón-Emeric, Bastian, and Lyles) and Department of Internal Medicine (Drs Green, Colón-Emeric, Bastian, Drake, and Lyles), Duke University Medical Center, Durham, North Carolina.

Acknowledgments

Dr Lyles was supported by grants AG11268 from the National Institute on Aging, 2031AH94004 from the Bureau of Health Professions, RR-30 from the Division of Research Resources, General Clinical Research Centers Program, National Institutes of Health, and by the Veterans Affairs Medical Research Service. Dr Colón-Emeric is supported by the Paul A. Beeson Award, National Institute on Aging grant AG024787, and the Claude D. Pepper Older Americans Independence Center. Dr Bastian is supported by the Veterans Affairs Career Development Award from Health Services Research and Development.

We acknowledge the following internal reviewers: Adi Cohen, MD, Darnel DeWalt, MD, and Margaret L. Gourlay, MD. We thank Lesa Hall-Young (Medical Media, Durham Veterans Affairs Medical Center, Durham, North Carolina) for technical assistance with the figures.

REFERENCES

1. Riggs BL, Melton LJ. *Osteoporosis: Etiology, Diagnosis, and Management*. Philadelphia, PA: Lippincott-Raven; 1995.
2. Pachucki-Hyde L. Assessment of risk factors for osteoporosis and fracture. *Nurs Clin North Am*. 2001;36(3):401-408.
3. Kramer AM, Steiner JF, Schlenker RE, et al. Outcome and costs after hip fracture and stroke. *JAMA*. 1997;277(5):396-404.
4. Lyles K, Gold D, Shipp K, Pieper C, Martinez S, Mulhausen PL. Osteoporotic vertebral compression fractures: their association with impaired functional status. *Am J Med*. 1993;94(6):595-601.

5. Magaziner J, Simonsick EM, Kashner TM, Hebel JR, Kenzora JE. Predictors of functional recovery one year following hospital discharge for hip fracture: a prospective study. *J Gerontol.* 1990;45(3):M101-M107.
6. Randell A, Nguyen TV, Bhalerao N, Silverman SL. Deterioration in quality of life following hip fracture: a prospective study. *Osteoporos Int.* 2000;11(5):460-466.
7. Ray N, Chan J, Thalmer M, Melton L. Medical expenditures for the treatment of osteoporosis in the United States in 1995: report from the National Osteoporosis Foundation. *J Bone Miner Res.* 1997;12(1):24-35.
8. Jette A, Harris B, Cleary P, Campion E. Functional recovery after hip fracture. *Arch Phys Med Rehabil.* 1987;68(10):735-740.
9. Melton LJ III, Eddy DM, Johnston CC Jr. Screening for osteoporosis. *Ann Intern Med.* 1990;112(7):516-528.
10. Nelson HD, Hefland M, Woolf SH, Allan JD. Screening for postmenopausal osteoporosis: a review of the evidence for the US Preventative Services Task Force. *Ann Intern Med.* 2002;137(6):529-541.
11. National Osteoporosis Foundation. *Physician's Guide to Prevention and Treatment of Osteoporosis.* Washington, DC: National Osteoporosis Foundation; 2000.
12. Goddard D, Kleerekoper M. The epidemiology of osteoporosis: practical implications for patient care. *Postgrad Med.* 1998;104(4):54-72.
13. Eastell R, Cedel SL, Wahner HW, Riggs BL, Melton LJ III. Classification of vertebral fractures. *J Bone Miner Res.* 1991;6(3):207-215.
14. Nelson HD, Morris CD, Kraemer DF, et al. *Osteoporosis in Postmenopausal Women: Diagnosis and Monitoring.* Rockville, MD: Agency for Healthcare Research and Quality; 2002. AHRQ Publication No. 01-E032.
15. Snelling AM, Crespo CF, Schaeffer M, Smith S, Walbourn L. Modifiable and nonmodifiable factors associated with osteoporosis in postmenopausal women: results from the Third National Health and Nutrition Examination Survey, 1988-1994. *J Womens Health Gend Based Med.* 2001;10(1):57-65.
16. Zimmerman SI, Girman CJ, Buie VC, et al. The prevalence of osteoporosis in nursing home residents. *Osteoporos Int.* 1999;9(2):151-157.
17. Melton LJ III, Lane AW, Cooper C, Eastell R, O'Fallon WM, Riggs BL. Prevalence and incidence of vertebral deformities. *Osteoporos Int.* 1993;3(3):113-119.
18. Chappard D, Alexandre C, Robert JM, Riffat G. Relationships between bone and skin atrophies during aging. *Acta Anat (Basel).* 1991;141(3):239-244.
19. Robinson RJ, al-Azzawi F, Iqbal SJ, Abrams K, Mayberry JF. The relation of hand skin-fold thickness to bone mineral density in patients with Crohn's disease. *Eur J Gastroenterol Hepatol.* 1997;9(10):945-949.
20. Orme SM, Belchetz PE. Is a low skinfold thickness an indicator of osteoporosis? *Clin Endocrinol (Oxf).* 1994;41(3):283-287.
21. Versluis RG, Petri H, van de Ven CM, et al. Usefulness of arm span and height comparison in detecting vertebral deformities in women. *Osteoporos Int.* 1999;9(2):129-133.
22. Payette H, Kertgoat MJ, Shatenstein B, Boutier V, Nadon S. Validity of self-reported height and weight estimates in cognitively intact and impaired elderly individuals. *J Nutr Health Aging.* 2000;4(4):223-228.
23. Sanila M, Kotaniemi A, Viikari J, Isomaki H. Height loss rate as a marker of osteoporosis in postmenopausal women with rheumatoid arthritis. *Clin Rheumatol.* 1994;13(2):256-260.
24. Cummings SR, Nevitt MC, Browner WS, et al. Risk factors for hip fracture in white women: Study of Osteoporotic Fractures Research Group. *N Engl J Med.* 1995;332(12):767-773.
25. Must A, Phillips SM, Naumova EN, et al. Recall of early menstrual history and menarcheal body size: after 30 years, how well do women remember? *Am J Epidemiol.* 2002;155(7):672-679.
26. Must A, Willett WC, Dietz WH. Remote recall of childhood height, weight and body build by elderly subjects. *Am J Epidemiol.* 1993;138(1):56-64.
27. Norgan NG, Cameron N. The accuracy of body weight and height recall in middle-aged men. *Int J Obes Relat Metab Disord.* 2000;24(12):1695-1698.
28. Ettinger B, Black DM, Palermo L, Nevitt MC, Melnikoff S, Cummings SR. Kyphosis in older women and its relation to back pain, disability and osteopenia: the study of osteoporotic fractures. *Osteoporos Int.* 1994;4(1):55-60.
29. Chow RK, Harrison JE. Relationship of kyphosis to physical fitness and bone mass in post-menopausal women. *Am J Phys Med.* 1987;66(5):219-227.
30. Cortet B, Houvenagel E, Puisieux F, Roches E, Garneir P, Delcambre B. Spinal curvatures and quality of life in women with vertebral fractures secondary to osteoporosis. *Spine.* 1999;24(18):1921-1925.
31. Siminoski K, Lee K, Warshawski R. Accuracy of physical examination for detection of thoracic vertebral fractures. *J Bone Miner Res.* 2003;18(suppl 2):S82.
32. Siminoski K, Warshawski RS, Jen H, Lee K. Accuracy of physical examination using the rib-pelvis distance for detection of lumbar vertebral fractures. *Am J Med.* 2003;115(3):233-236.
33. Di Monaco M, Di Monaco R, Manca M, Cavanna A. Handgrip strength is an independent predictor of distal radius bone mineral density in postmenopausal women. *Clin Rheumatol.* 2000;19(6):473-476.
34. Foley KT, Owings TM, Pavol MJ, Grabiner MD. Maximum grip strength is not related to bone mineral density of the proximal femur in older adults. *Calcif Tissue Int.* 1999;64(4):291-294.
35. Metlay JP, Kapoor WN, Fine MJ. Does this patient have community-acquired pneumonia? diagnosing pneumonia by history and physical examination. *JAMA.* 1997;278(17):1440-1445.
36. Bickley L. *Bate's Guide to Physical Examination and History Taking.* Philadelphia, PA: Lippincott Williams & Wilkins; 1999.
37. DeGowin R. *DeGowin and DeGowin's Diagnostic Examination.* 6th ed. New York, NY: McGraw-Hill; 1994.
38. Cabot RC. *Cabot and Adams Physical Diagnosis.* 13th ed. Baltimore, MD: Williams & Wilkins; 1942.
39. Delp MH, Manning RT. *Major's Physical Diagnosis: An Introduction to the Clinical Process.* 9th ed. Philadelphia, PA: WB Saunders; 1981.
40. Holleman DR Jr, Simel DL. Does the clinical examination predict airflow limitation? *JAMA.* 1995;273(4):313-319.
41. Lydick E, Cook K, Turpin J, Melton M, Stine R, Byrnes C. Development and validation of a simple questionnaire to facilitate identification of women likely to have low bone density. *Am J Manag Care.* 1998;4(1):37-48.
42. Cadarette SM, Jaglal SB, Kreiger N, McIsaac WJ, Darlington GA, Tu JV. Development and validation of the Osteoporosis Risk Assessment Instrument to facilitate selection of women for bone densitometry. *CMAJ.* 2000;162(9):1289-1294.
43. Weinstein L, Ullery B. Identification of at-risk women for osteoporosis screening. *Am J Obstet Gynecol.* 2000;183(3):547-549.
44. Michaelsson K, Bergstrom R, Mallmin H, Holmberg L, Wolk A, Ljunghall S. Screening for osteopenia and osteoporosis: selection by body composition. *Osteoporos Int.* 1996;6(2):120-126.
45. Cadarette SM, Jaglal SB, Murray TM, McIsaac WJ, Joseph L, Brown JP. Evaluation of decision rules for referring women for bone densitometry by dual-energy x-ray absorptiometry. *JAMA.* 2001;286(1):57-63.
46. Nguyen TV, Center JR, Pocock NA, Eisman JA. Limited utility of clinical indices for the prediction of symptomatic fracture risk in postmenopausal women. *Osteoporos Int.* 2004;15(1):49-55.
47. Dargent-Molina P, Poitiers F, Breart G; EPIDOS Group. In elderly women weight is the best predictor of a very low bone mineral density: evidence from the EPIDOS Study. *Osteoporos Int.* 2000;11(10):881-888.
48. Vogt TM, Ross PD, Palermo L, et al; Fracture Intervention Trial Research Group. Vertebral fracture prevalence among women screening for the Fracture Intervention Trial and a simple clinical tool to screen for undiagnosed vertebral fractures. *Mayo Clin Proc.* 2000;75(9):888-896.
49. Verhaar HJ, Koele JJ, Neijzen T, Dessens JA, Duursma SA. Are arm span measurements useful in the prediction of osteoporosis in postmenopausal women? *Osteoporos Int.* 1998;8(2):174-176.
50. Wang XF, Duan Y, Henry M, Kin BT, Seeman M. Body segment lengths and arm span in healthy men and women and patients with vertebral fractures. *Osteoporos Int.* 2004;15(1):43-48.
51. Bedogni G, Simonini G, Viaggi S, et al. Anthropometry fails in classifying bone mineral status in postmenopausal women. *Ann Hum Biol.* 1999;26(6):561-568.
52. May H, Murphy S, Khaw KT. Age-associated bone loss in men and women and its relationship to weight. *Age Ageing.* 1994;23(3):235-240.
53. Kantor S, Ossa KS, Hoshaw-Woodard SL, Lemeshow S. Height loss and osteoporosis of the hip. *J Clin Densitom.* 2004;7(1):65-70.
54. Balzini L, Vannucchi L, Benvenuti F, et al. Clinical characteristics of flexed posture in elderly women. *J Am Geriatr Soc.* 2003;51(10):1419-1426.
55. Kritz-Silverstein D, Barrett-Connor E. Grip strength and bone mineral density in older women. *J Bone Miner Res.* 1994;9(1):45-51.
56. Taaffe DR, Pruitt L, Lewis B, Marcus R. Dynamic muscle strength as a predictor of bone mineral density in elderly women. *J Sports Med Phys Fitness.* 1995;35(2):136-142.
57. Bauer DC, Browner WS, Cauley JA, et al. Factors associated with appendicular bone mass in older women. *Ann Intern Med.* 1993;118(9):657-665.
58. Sinaki M, Fitzpatrick LA, Ritchie CK, Montesano A, Wahner HW. Site-specificity of bone mineral density and muscle strength in women: job related physical activity. *Am J Phys Med Rehabil.* 1998;77(6):470-476.

59. Sinaki M, Wahner HW, Offord KP. Relationship between grip strength and related regional bone mineral content. *Arch Phys Med Rehabil.* 1989;70(12):823-826.
60. Earnshaw SA, Keating N, Hosking DJ, et al. Tooth counts do not predict bone mineral density in early postmenopausal Caucasian women. *Int J Epidemiol.* 1998;27(3):479-483.
61. Klemetti E, Vainio P. Effect of bone mineral density in skeleton and mandible on extraction of teeth and clinical alveolar height. *J Prosthet Dent.* 1993;70(1):21-25.
62. Mercier P, Inoue S. Bone density and serum minerals in cases of residual alveolar ridge atrophy. *J Prosthet Dent.* 1981;46(3):250-255.
63. Elders PJ, Habets LL, Netelenbos JC, Van der Linden LW, Van der Stelt PF. The relation between periodontitis and systemic bone mass in women between 46 and 55 years of age. *J Clin Periodontol.* 1992;19(7):492-496.
64. Inagaki K, Kurosu Y, Kamiya T, et al. Low metacarpal bone density, tooth loss, and periodontal disease in Japanese women. *J Dent Res.* 2001;80(9):1818-1822.
65. Astrom J, Backstrom C, Thidevall G. Tooth loss and hip fractures in the elderly. *J Bone Joint Surg Br.* 1990;72(2):324-325.
66. May H, Reader R, Murphy S, Khaw KT. Self reported tooth loss and bone mineral density in older men and women. *Age Ageing.* 1995;24(3):217-221.
67. Kribbs PJ, Chesnut CH, Ott SM, Kilcoyne RF. Relationships between mandibular and skeletal bone in an osteoporotic population. *J Prosthet Dent.* 1989;62(6):703-707.
68. Krall EA, Dawson-Hughes B, Papas A, Garcia RI. Tooth loss and skeletal bone density in healthy postmenopausal women. *Osteoporos Int.* 1994;4(2):104-109.
69. Krall EA, Garcia RI, Dawson-Hughes B. Increased risk of tooth loss is related to bone loss at the whole body, hip, and spine. *Calcif Tissue Int.* 1996;59(6):433-437.
70. Phillips HB, Ashley FP. The relationship between periodontal disease and a metacarpal bone index. *Br Dent J.* 1973;134(6):237-239.
71. Habets LL, Bras J, Borgmeyer-Hoelen AM. Mandibular atrophy and metabolic bone loss. *Int J Oral Maxillofac Surg.* 1988;17(3):208-211.
72. Daniell HW. Postmenopausal tooth loss. *Arch Intern Med.* 1983;143(9):1678-1682.
73. De Laet C, van Hout BA, Burger H, Hofman A, Pols HA. Bone density and risk of hip fracture in men and women: cross sectional analysis. *BMJ.* 1997;315(7102):221-225.
74. Kaufman JM, Johnell O, Abadie E, et al. Background for studies on the treatment of male osteoporosis: state of the art. *Ann Rheum Dis.* 2000;59(10):765-772.
75. Melton LJ 3rd. Epidemiology of spinal fractures. *Spine.* 1997;22(24 suppl):2S-11S.

Prepared by Cathleen S. Colón-Emeric, MD,
and David L. Simel, MD, MHS

Reviewed by Kenneth W. Lyles, MD

CLINICAL SCENARIO

You are treating a 68-year-old man with a history of chronic obstructive pulmonary disease (COPD). He has not used long-term oral steroids and has not had a previous fracture. He weighs 72 kg. Should he be referred for bone densitometry?

- A body mass index (BMI) less than 25 in older women is the single best finding for detecting women with osteoporosis, performing better than decision rules. However, a BMI greater than 25 is not as informative as the decision rules for identifying women at the lowest risk of osteoporosis.
- In women at higher risk for osteoporosis, historical height loss identifies those most likely to have vertebral fractures. However, historical height loss should not be used as a screening test for osteoporosis for most postmenopausal women.

UPDATED SUMMARY ON OSTEOPOROSIS

Original Review

Green AD, Colón-Emeric CS, Bastian L, Drake MT, Lyles KW. Does this woman have osteoporosis? *JAMA*. 2004; 292(23):2890-2900.

UPDATED LITERATURE SEARCH

We repeated the literature search used in the original article, confined to 2004 to April 2006 and restricted to adult patients. We limited this by cross-linking to “exp physical examination/ or physical exam” and with the text words “sensitivity” or “specificity.” The strategy yielded 9 abstracts that we reviewed, of which 4 met our inclusion criteria. One of the included articles evaluated a prediction model for osteoporosis in Asian men.¹ From a review of the references, we found a second article that evaluated a similar prediction model in US male veterans.² We were unable to obtain a copy of a third article, although it evaluated a relatively small number of patients and would have therefore been of lower quality.

NEW FINDINGS

- Useful data are now available for osteoporosis in men. The variables in a simple prediction model show that the risk factors (body size and age) for men are similar to those for women.
- The clinical diagnosis of kyphosis from gross observations is about 85% accurate, but it does not diagnose osteoporosis, as well as more formal measures of kyphosis.

Details of the Update

A new clinical model was developed prospectively in healthy Argentinean postmenopausal women attending a menopause clinic.³ These women were attending the clinic for a variety of reasons; therefore, the study sample was without referral bias. The only clinical sign evaluated in the study was kyphosis, measured with 85% accuracy from simple clinical observations. The study systematically collected multiple risk factors to develop a 5-variable model for predicting osteoporosis measured on hip bone mineral density. The independent predictors (>10 years of menopause, calcium intake < 1200 mg/d, kyphosis, BMI < 25, and history of fractures) were validated prospectively in a new set of patients.

Few data exist for osteoporosis in men. A large, prospective study of Asian men yielded a simple prediction model with variables similar to the variables predicting osteoporosis in women, with weight and age as the variables in the final model (the Osteoporosis Self-assessment Test [OST]).⁴ A study in US male veterans, using the same scoring system, yielded similar results.² Cut points vary substantially within the population, in part because of differences in the average weight of the population.

Women may report height loss, but does it predict vertebral fractures? A study conducted in women who were referred to an endocrinologist for an osteoporosis evaluation showed that a difference of more than 6 cm between the measured height and tallest recalled height makes a vertebral fracture highly likely.⁵ However, the study population had a 57% prevalence of vertebral fractures. In population studies in which the prevalence will be lower (10%-25%) among women older than 50 years, the absence of such a large degree of height loss decreases the likelihood ratio (LR) to 0.76 but will not definitively rule out a fracture. Using a threshold of “any” height loss resulted in

a positive likelihood ratio (LR+) of 1.0 and a negative LR of 0.75, results that are not informative.

IMPROVEMENTS IN THE DATA PRESENTED IN THE ORIGINAL PUBLICATION

We qualitatively compared the Simple Calculated Osteoporosis Risk Estimate (SCORE)⁶ and the Osteoporosis Risk Assessment Instrument (ORAI)⁷ because both were recommended by the Canadian Preventive Health Services Task Force. The SCORE questionnaire is more efficient at detecting women unlikely to have osteoporosis (LR, 0.02 for a

score < 6) but is more complex to calculate than the ORAI. The ORAI also has good measurement properties and has only 3 variables (age, weight, and estrogen use). We recalculated the confidence intervals (CIs) for the ORAI (not reported in the original article) and found that for an ORAI greater than or equal to 9, the LR is 1.6 (95% CI, 1.4-1.8); for women with a more normal ORAI score of less than 9, the LR is 0.13 (95% CI, 0.04-0.40). However, we observed that with fewer women receiving hormone replacement therapy, almost all postmenopausal women would score higher than the cut point, which limits its utility in clinical practice.

Table 36-8 Univariate Findings for Osteoporosis in Women

Finding	LR+ (95% CI)	LR- (95% CI)
Kyphosis detected with clinical observation	1.5 (1.0-2.2)	0.73 (0.51-1.0)
BMI < 25	4.5 (2.5-8.3)	0.48 (0.34-0.69)

Abbreviations: BMI, body mass index; CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

Table 36-9 Univariate Findings for Vertebral Fracture in Women

Finding	LR+ (95% CI)	LR- (95% CI)
Historical height loss	4.6 (2.5-8.4)	0.76 (0.67-8.4)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

Table 36-10 Multivariate Findings for Osteoporosis in Women

Risk Factors in the Model	No. of Risk Factors Present	Probability, %
> 10 y Menopause	5	98
Calcium intake < 1200 mg/d	4	73-95
Kyphosis	3	21-85
BMI < 25 kg	2	7-42
Personal fractures	1	3-33
	0	4

Abbreviation: BMI, body mass index.

Table 36-11 Osteoporosis Self-assessment Tool in Men

Test	OST Score ^a	LR+ (95% CI)	LR- (95% CI)
Asian men	≤ -1	2.2 (1.8-2.8)	0.40 (0.27-0.59)
US veterans	≤ 3	2.7 (2.1-3.5)	0.11 (0.03-0.41)
Summary		2.4 (2.0-2.9)	0.35 (0.23-0.53)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio; OST, Osteoporosis Self-assessment Tool.

^aOST Score = 0.2 × (body weight [kg] – age [y]). Multiply the value in parentheses by the coefficient and round the score down to the nearest integer.

CHANGES IN THE REFERENCE STANDARD

None.

RESULTS OF LITERATURE REVIEW

While univariate findings can be used for identifying osteoporosis (Table 36-8) or vertebral fracture (Table 36-9) in women, a multivariate approach is preferred (Table 36-10).

A multivariate score has now been developed for men (Table 36-11).

EVIDENCE FROM GUIDELINES

Since 2004, a search that restricts “exp osteoporosis” to guidelines yields 26 articles that are mostly from specialty groups. The Canadian Task Force on Preventive Health Care recommends bone densitometry screening risk assessment every 1 to 2 years in women.⁸ Those who are 65 years of age or older have a previous fragility fracture, weigh fewer than 60 kg, have a SCORE questionnaire result greater than or equal to 6, or have an ORAI score greater than or equal to 9 should be screened with bone densitometry.⁸ These latter 2 (ie, scoring questionnaire results) were the 2 decision rules in which normal scores make osteoporosis unlikely.

CLINICAL SCENARIO—RESOLUTION

In the absence of previous fracture or corticosteroid use, there is no current recommendation to guide osteoporosis screening for men. The prevalence of osteoporosis in men with COPD is approximately 10%.

Using his age and weight, this patient’s score on the OST is 0 ($0.2 \times [72 \text{ kg} - 68 \text{ y}] = 0.8$, rounded down to nearest integer = 0). The LR+ for a score less than or equal to 1 is 3.8. Thus, the posttest probability of osteoporosis in this patient is 30%, high enough to warrant further screening with dual-energy x-ray absorptiometry.

OSTEOPOROSIS—MAKE THE DIAGNOSIS

PRIOR PROBABILITY

The prior probability of osteoporosis in women depends on age and ethnicity (Tables 36-12 and 36-13).

Table 36-12 Age-Dependent Prevalence of Osteoporosis in White Women

Age, y	Prevalence, %
50-59	15
60-69	22
70-79	39
≥80	70

Table 36-13 Age-Dependent Prevalence in Nonwhite Women

Women > 50 y	Prevalence, %
Non-Hispanic black women	12
Mexican American women	19
Other ethnicity	28

Comparable data for men have not been adequately validated.

POPULATION FOR WHOM OSTEOPOROSIS SHOULD BE CONSIDERED

Age beyond menopause and low BMI (<25) or weight (<60 kg) are the most important predictors of osteoporosis in women. Older age and low BMI might also be the most important factors in men. Any older patient with a minimal trauma fracture or kyphosis should be screened for osteoporosis.

DETECTING THE LIKELIHOOD OF OSTEOPOROSIS

The SCORE and ORAI questionnaires have the best measurement properties for screening (see Tables 36-14 and 36-15), but the ORAI is a bit easier to use. The OST has not been as extensively validated in women but is one of the few tests with evidence in men.

Table 36-14 Osteoporosis Risk Assessment Instrument

Item	Scoring	LR (95% CI)
Age	≥ 75 y = 15 Points	Total score ≥ 9 points, LR = 1.6 (1.4-1.8)
	65-74 y = 9 Points	
	55-64 y = 5 Points	
Weight	< 60 kg = 9 Points	Total score < 9 points, LR = 0.13 (0.04-0.40)
	60-69.9 kg = 3 Points	
No current estrogen use	2 Points	

Abbreviations: CI, confidence interval; LR, likelihood ratio.

Table 36-15 Osteoporosis Self-assessment Tool

Test	Abnormal	LR+ (95% CI)	LR- (95% CI)
Score = 0.2 (body weight [kg] – age [y]), rounded down to nearest integer	Depends on population; ≤ –1 in Asian men, ≤ 3 in US male veterans	2.4 (2.0-2.9)	0.35 (0.23-0.53)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

REFERENCE STANDARD TESTS

Bone mineral densitometry with T score values less than or equal to 2.5 SDs below the mean of young, healthy population.

REFERENCES FOR THE UPDATE

- Kung AWS, Ho AYY, Ross PD, Reginster JY. Development of a clinical assessment tool in identifying Asian men with low bone mineral density and comparison of its usefulness to quantitative bone ultrasound. *Osteoporos Int*. 2005;16(7):849-855.
- Adler RA, Tran MT, Petkov VI. Performance of the osteoporosis self-assessment screening tool for osteoporosis in American men. *Mayo Clin Proc*. 2003;78(6):723-727.^a
- Masoni A, Morosano M, Pezzotto SM, et al. Construction of two instruments for the presumptive detection of post-menopausal women with low spinal bone mass by means of clinical risk factors. *Maturitas*. 2005;51(3):314-324.
- Kung AWS, Ho AYY, Ross PD, Reginster JY. Development of a clinical assessment tool in identifying Asian men with low bone mineral density and comparison of its usefulness to quantitative bone ultrasound. *Osteoporos Int*. 2005;16(4):849-855.^a
- Siminoski K, Warshawski RS, Jen H, Lee K. The accuracy of historical height loss for the detection of vertebral fractures in postmenopausal women. *Osteoporos Int*. 2006;17(2):290-296.
- Lydick E, Cook K, Turpin J, Melton M, Stine R, Byrnes C. Development and validation of a simple questionnaire to facilitate identification of women likely to have low bone density. *Am J Manage Care*. 1998;4(1):37-48.
- Cadarette SM, Jaglal SB, Kreiger N, McIsaac WJ, Darlington GA, Tu JV. Development and validation of the Osteoporosis Risk Assessment Instrument to facilitate selection of women for bone densitometry. *CMAJ*. 2000;162(9):1289-1294.
- Cheung AM, Feig DS, Kapral M, Diaz-Granados N, Dodin S; Canadian Task Force on Preventive Health Care. Prevention of osteoporosis and osteoporotic fractures in postmenopausal women: recommendation statement from the Canadian Task Force on Preventive Health Care. *CMAJ*. 2004;170(11):1665-1667.

^aFor the Evidence to Support the Update for this topic, see <http://www.JAMAevidence.com>.

This page intentionally left blank

EVIDENCE TO SUPPORT THE UPDATE:

36

Osteoporosis

TITLE Performance of the Osteoporosis Self-assessment Screening Tool for Osteoporosis in Men.

AUTHORS Adler RA, Tran MT, Petkov VI.

CITATION Mayo Clin Proc. 2003;78(6):723-727.

QUESTION Can the Osteoporosis Self-assessment Tool (OST) be used to detect older men at high risk for osteoporosis?

DESIGN Cross-sectional.

SETTING Rheumatology and pulmonary clinics at a single Veterans Affairs health center.

PATIENTS One hundred eighty-one men (69% white, 30% black) without previous dual-energy x-ray absorptiometry (DXA) measurement recruited from clinic population. Mean age was 64.3 years but with a wide range (32–87 years). The mean weight of the participants was also high, at 91 kg (mean body mass index, 29).

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

Participants completed questionnaires including self-reported age, weight, and other risk factors for osteoporosis. The OST was calculated with the following formula:

$$\text{OST score} = 0.2 \times (\text{body weight [kg]} - \text{age [y]})$$

with the result rounded down to the nearest integer

MAIN OUTCOME MEASURES

All participants had bone density measured by DXA as the diagnostic standard. A T score of less than or equal to -2.5 at any site was used to define osteoporosis.

MAIN RESULTS

Overall, 16% of participants had osteoporosis. The OST had good discriminative properties (area under the curve, 0.836). The authors reported the sensitivity and specificity at various

Table 36-16 Likelihood Ratios for Osteoporosis Self-assessment Tool Depending on the Cut Point Value

OST Score	Sensitivity, %	Specificity, %	LR+ (95% CI)	LR- (95% CI)
≤ 3	95	66	2.7 (2.2-3.5)	0.1 (0.03-0.40)
≤ 2	82	74	3.2 (2.3-4.3)	0.2 (0.1-0.5)
≤ 1	75	80	3.8 (2.6-5.5)	0.3 (0.2-0.6)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio; OST, Osteoporosis Self-assessment Tool.

cut points (Table 36-16). Adding other clinical risk factors from the questionnaire to the OST score did not improve its discriminative ability.

The authors recommend a cut point of 3 as most appropriate for their population. This differs from the recommended cut points from other studies of the OST in Asian men (cut point ≤ -1), community-dwelling elderly men in Rotterdam and Baltimore (≤ 2), and Asian and white women (≤ 1).

CONCLUSIONS

LEVEL OF EVIDENCE Level 1.

STRENGTHS The OST is simple to use and can be calculated with either self-reported or routinely collected information. It appears to be useful across a wide range of populations of both men and women.

LIMITATIONS The sample was a subset of male veterans likely at higher risk for osteoporosis because of rheumatologic and pulmonary diseases and therapies. There may be an additional selection bias because it is unclear how men were recruited into the study.

The cut point used for the OST seems to depend substantially on the population, in part because of large variations in average weight. Different cut points may be indicated in different clinical situations. If the goal is to exclude men who would otherwise have screening ordered, the higher cut point should be used. Conversely, if the goal is to identify men for screening who would otherwise not be tested, the lower cut point is more appropriate.

Reviewed by Cathleen S. Colón-Emeric, MD, MHSc, and David L. Simel, MD, MHS

TITLE Development of a Clinical Assessment Tool in Identifying Asian Men With Low Bone Mineral Density and Comparison of Its Usefulness to Quantitative Bone Ultrasonography.

AUTHORS Kung AWS, Ho AYY, Ross PD, Reginster JY.

CITATION *Osteoporosis Int.* 2005;16(7):849-855.

QUESTION Can a simple prediction rule be developed for detecting osteoporosis in older men?

DESIGN Prospective volunteers.

SETTING Patients were recruited from the community at “public road shows, health fairs, or health talks on osteoporosis” in Hong Kong. The recruitment period lasted more than 5 years.

PATIENTS Southern Chinese men older than 50 years who volunteered. Young men (20-39 years) were also invited to participate as controls.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

Risk factors were recorded during a structured interview. The only physical examination assessments were height (measured with a stadiometer) and weight (measured with the patient shoeless, in light indoor clothing, on a balance beam scale). The reference standard was bone mineral density (BMD) of the lumbar spine and left femur.

MAIN OUTCOME MEASURES

The BMD T scores were obtained by comparison with the healthy young men recruited to the study ($n = 124$). Osteoporosis was defined by a femoral neck BMD T score less than or equal to 2.5.

MAIN RESULTS

Out of 906 men who were invited to participate, 74 men declined and 56 were excluded for other known illnesses

Table 36-17 Likelihood Ratios for Osteoporosis Self-assessment Tool

OST Score	Sensitivity	Specificity	LR+ (95% CI)	LR- (95% CI)
≤ -1	0.73	0.68	2.2 (1.8-2.8)	0.40 (0.27-0.59)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio; OST, Osteoporosis Self-assessment Tool.

associated with bone disease. The data for the remaining men (mean age, 65 years) were divided into a model development sample of 420 and a validation sample of 356. The prevalence of osteoporosis was 16% ($n = 126$).

$$\text{Osteoporosis score} = 0.2 \times (\text{body weight [kg]} - \text{age [y]})$$

(Multiply the value in parentheses by the coefficient and round the score *down* to the nearest integer.)

Osteoporosis score ≤ -1 , increased risk

Osteoporosis score > -1 , low risk

The results shown in Table 36-17 were also confirmed in a separate validation set where the sensitivity (0.71) and specificity (0.68) were essentially the same.

CONCLUSIONS

LEVEL OF EVIDENCE Level 1.

STRENGTHS This was a prospective study of community men who volunteered for screening. The data were studied in a validation set.

LIMITATIONS The men were invited into the study through community health activities. Though there may have been some volunteer bias, the number of patients is large.

This is one of the few studies of osteoporosis in men. Compared with models for women, the variables are almost identical, except that age replaces the measure for menopause. Although the variables in the prediction model were similar to those for women, the model for men appears to be slightly better at identifying cases of osteoporosis and slightly less efficient at confirming the absence of osteoporosis. The model is extremely easy to use.

Reviewed by David L. Simel, MD, MHS

CHAPTER 37

Does This Child Have Acute Otitis Media?

Russell Rothman, MD, MPP

Thomas Owens, MD

David L. Simel, MD, MHS

CLINICAL SCENARIO

A mother notices that her 15-month-old child has a low-grade fever and is tugging at his ears after several days of cough and runny nose. The child attends day care services and had 1 previous episode of acute otitis media (AOM) about 4 months ago. In the physician's office, he is afebrile but somewhat irritable and has clear rhinorrhea, mild posterior pharyngeal erythema, and normal chest auscultatory findings. Cerumen occludes the view of his right tympanic membrane, whereas the left tympanic membrane shows normal landmarks and good mobility on pneumatic otoscopy. After removal of the cerumen from his right ear, landmarks are visible on a slightly erythematous tympanic membrane. The tympanic membrane shows normal mobility on pneumatic otoscopy.

WHY IS THIS AN IMPORTANT QUESTION TO ANSWER WITH THE CLINICAL EXAMINATION?

Acute otitis media can be a difficult and controversial diagnosis to make, but studies suggest that AOM is responsible for more than 30 million clinic visits a year in the United States, at a total cost exceeding \$5 billion. This makes AOM one of the most commonly diagnosed and expensive childhood illnesses.¹⁻⁴ Studies have shown that by age 1 year, up to 60% of all children have been diagnosed as having at least 1 episode of AOM, and by age 3 years, more than 80% of children have had at least 1 episode.^{1,5} The best estimates of the prevalence of AOM are based on the National Ambulatory Medical Care Survey. In 1990, the percentage of office visits with otitis media as the principal diagnosis was 17.4% for children aged 0 to 2 years, 18.1% for children aged 2 to 5 years, 10.5% for children aged 6 to 10 years, and 5.2% for children aged 11 to 15 years.⁶ The most common potential risk factors for diagnosis of AOM include age younger than 2 years, male sex, day care attendance, fall or winter season, exposure to cigarette smoke, genetic factors, and history of AOM.^{1,7} Breastfeeding appears to be protective.⁷

Making a correct diagnosis of AOM is often difficult, particularly in young children. Distinguishing between AOM and otitis media with effusion (OME) can be particularly challenging. Several studies suggest that physicians are uncertain of their diagnosis of AOM as much as 40% of the time.⁸ This uncertainty probably contributes to overdiagnosis, as shown in a study that when physicians estimate the odds that a patient has AOM are 50% or less, 3 of 4 will still prescribe antibiotics and 1 of 4 will still prescribe antibiotics when the odds of AOM are $\leq 25\%$.⁹ Various definitions and diagnostic criteria for AOM may also contribute to overdiagnosis. In a study by Hayden,¹⁰ 18 criteria sets for diagnosing AOM were used in 26 articles, and 165 surveyed clinicians identified 147 unique criteria. Recently, an expert panel convened

by the Agency for Healthcare Research and Quality (AHRQ) released a complicated definition requiring the presence of a middle ear effusion and rapid onset of associated symptoms (Box 37-1).^{6,11}

Overdiagnosis of AOM is thought to be common^{7,12,13} and contributes to increased antibiotic use and bacterial resistance. Overdiagnosis may also result in unnecessary specialty referrals and increased use of tympanostomy tubes. In addition, improper diagnosis of AOM in younger children may hinder the proper diagnosis of other underlying causes of fever or illness.

Box 37-1 Agency for Healthcare Research and Quality Definition of Acute Otitis Media⁶

Presence of middle ear effusion, demonstrated by actual presence of fluid in the middle ear, as diagnosed by tympanocentesis, or physical presence of fluid in the external ear canal as a result of tympanic membrane perforation or indicated by limited or absent mobility of the tympanic membrane, as diagnosed by pneumatic otoscopy, tympanogram, or acoustic reflectometry with or without the following:

Opacification, not including erythema

Full or bulging tympanic membrane

Hearing loss

AND

Rapid onset (during a course of 48 hours) of 1 or more of the following signs or symptoms with or without anorexia, nausea, or vomiting:

Otalgia (or pulling of ear in an infant)

Otorrhea

Irritability in infant or toddler

Fever

Anatomic/Physiologic Origins

Genetic, infectious, immunologic, and environmental factors contribute to an underlying predisposition to ear infections.² The eustachian tube, shorter and angled much less steeply in children than in adults, plays a critical role by more easily allowing the reflux of organisms from the nasopharynx into the middle ear.² When the tube becomes congested, as it may with a viral infection in the upper respiratory tract, negative pressure within the middle ear causes secretions to accumulate, and this leads to the proliferation of pathogenic organisms. The bacterial agents most commonly identified in AOM include *Streptococcus pneumoniae*, *Haemophilus influenzae*, and *Moraxella catarrhalis*.⁵ Coinfection with viruses is also observed in 30% to 40% of cases and may play a role in the virulence of symptoms, but less than 10% of AOM is caused by viruses alone.^{5,14,15} Most ear infections resolve without any specific treatment, so the exact role of bacterial or viral pathogens remains unclear.

The buildup of infectious debris behind the tympanic membrane, along with inflammatory mediators, produces the symptoms and signs of AOM. An effusion changes the tympanic membrane's appearance from transparent to opaque and can distort or bulge the membrane, making it difficult to visualize normal landmarks (Figure 37-1). Erythema of the tympanic membrane is related to vascular congestion of the membrane and is thought to represent a nonspecific sign related to irritation of the drum or crying.^{2,12}

How to Elicit Symptoms and Signs

Common but usually nonspecific symptoms associated with the diagnosis of AOM include fever, ear pain, ear pulling, irritability, cough, and rhinitis. In a study of 354 children younger than 15 years (mean, 3.8 years) presenting for an acute illness, 90% of children in whom AOM was diagnosed had fever, ear pain, crying, and irritability alone or in combination, but 72% of children without AOM also presented with these symptoms.^{12,16}

To properly examine the ear for AOM, clinicians should use a pneumatic otoscope to visualize the landmarks and mobility of the tympanic membrane. After the patient is placed in a restrained or other safe position, the otoscope speculum is placed into the external auditory canal. The largest-sized speculum that can comfortably fit into the canal is recommended because a small speculum can limit the visual field and potentially cause pain by irritating the bony canal.^{2,17} A study by Cavanaugh¹⁸ suggested that children older than 18 months should have a soft-tipped speculum to provide an adequate seal and prevent air leakage when performing pneumatic otoscopy. It is also important that the otoscope have a bright light source for visualizing the tympanic membrane. Barriga et al¹⁹ tested otoscopes in clinics and emergency departments and found that 22% were inadequate because of either a worn bulb or a weak battery source.

To properly examine the tympanic membrane, one should evaluate the position, color, landmarks, degree of translucency, and mobility. The position refers to whether the drum

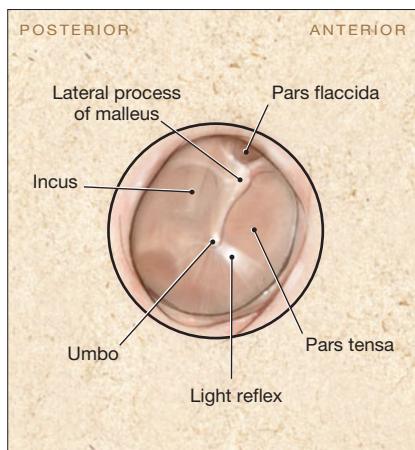


Figure 37-1 Tympanic Membrane Landmarks

appears to be bulging toward the examiner (suggestive of AOM), neutral (normal), or retracted away from the examiner (observed in chronic OME). The tympanic membrane can appear red, pink, yellow (with pus behind the drum), or pearly gray or translucent (normal). Landmarks that should be visible in a normal ear include the pars flaccida, the malleus, and the light reflex below the umbo (Figure 37-1). With a translucent tympanic membrane, the outline of the incus can sometimes be visualized as well. An opaque drum may be a sign of infection or middle ear effusion and can result in a diminished light reflex.

A bulb attachment can test the mobility of the drum with the slightest pressure or release. A study by Cavanaugh²⁰ suggests that only 10 to 15 mm H₂O of positive pressure is needed to assess drum mobility, whereas bulb attachments can easily create pressures of 1000 mm H₂O or more. Forceful pressing of the bulb creates excessive positive pressure that causes pain; in this instance, pain on insufflation does not diagnose infection. The correctly applied positive or negative pressure creates synchronous movement of the normal drum. An immobile drum or one with reduced mobility suggests the presence of a middle ear effusion.

The tympanic membrane can sometimes be difficult to visualize because of patient behavior or the buildup of cerumen in the ear canal. Apprehensive infants and young children can often be sufficiently restrained by having the parent seat the child in his or her lap, using his or her legs to wrap around the child's legs and arms to restrain the child's arms and head. The examiner should hold the otoscope with part of the hand touching the child's head so that the otoscope will move with the child's head and prevent injury.

In a study of 279 children with AOM, 29% required cerumen removal to make a proper diagnosis.²¹ Studies have not adequately compared various modes for physically removing cerumen, though the most common methods cited by generalists are the use of a wire loop, a blunt cerumen curette, or gentle irrigation with room-temperature water. One small randomized trial compared 2 ceruminolytic agents, liquid docusate sodium and triethanolamine polypeptide, applied at an emergency department visit with or without irrigation 15 minutes later. Liquid docusate sodium was highly effective compared with triethanolamine polypeptide, with successful cerumen removal in 82% of patients (number needed to treat for benefit, 3; 95% confidence interval [CI], 2-4).²²

Other techniques used in the diagnosis of AOM include tympanocentesis, tympanometry, and acoustic reflectometry. Tympanocentesis is performed through an otoscope with a special attachment or an otomicroscope. A tuberculin syringe needle is placed into the inferior portion of the tympanic membrane to aspirate fluid.² This technique can be diagnostic and is considered the criterion standard for detecting the presence of fluid in the diagnosis of AOM. However, tympanocentesis is rarely practiced in the primary care setting, where most AOM is managed.¹² Tympanometry and acoustic reflectometry both require the use of additional medical equipment. For tympanometry, a specialized probe is inserted into the canal to form a seal and measure the

amount of reflected sound energy. The amount of reflected energy is used to estimate tympanic membrane motility. In acoustic reflectometry, tympanic membrane motility is also estimated according to sound reflecting from the middle ear, but no seal is required. Both techniques assess tympanic membrane motility and generally have been studied only for detecting an effusion in patients with OME rather than AOM.^{1,7,12,23}

METHODS

Search Strategy and Quality Review

We searched MEDLINE from January 1966 to May 2002 for English-language articles that examined the role of symptoms and signs in the diagnosis of AOM. Multiple MEDLINE search strategies were applied by a single author (T.O.) using techniques that have been used by other authors in this series.^{24,25} We also examined bibliographies of selected articles and used general and specialty textbooks.^{1,2,7,26-29} From 397 identified references, 50 complete articles were retrieved for review by 2 authors (R.R. and T.O.). Among these, we found 17 articles that specifically examined symptoms and signs that were directly relevant to the diagnosis of AOM.^{4,10,16,23,30-42} Articles on the diagnosis of persistent OME were generally excluded because most of these studies were performed by comparing detection of an effusion by pneumatic otoscopy or tympanometry with the presence of an effusion at the time of surgery for myringotomy, rather than in ambulatory settings. In addition, persistent OME is a disease with different pathophysiology and, possibly, different diagnostic characteristics than AOM.

The 17 identified articles underwent independent quality review by 2 authors (R.R. and T.O.). Quality was assessed with an established methodologic filter for assessing internal validity that has been used and explained by other authors in this series.^{24,25} Each article was assigned a level of evidence (1-4) and consensus was reached by both reviewers. Tympanocentesis was considered the pathologic criterion standard, but only 1 study that assessed physical examination findings used this standard.²³ We therefore also included articles that used a standardized clinical definition of AOM when examining articles that dealt with symptoms. Although using a clinical criterion standard was not ideal and might lead to accusations of circular reasoning, the quality of the literature for this common problem left us little choice. However, we believed it was justified to examine these articles because most physicians make a diagnosis according to clinical criteria, and physicians make decisions to treat according to these criteria.

No article met evidence level 1 or 2, which required using an independent blind comparison of signs or symptoms against a criterion standard among consecutive patients. All articles reviewed were graded as evidence levels 3 to 5, but we retained only the level 3 and 4 articles. Level 3 studies used an independent, blind comparison of symptoms to the criterion standard and nonconsecutive patients suspected to have the targeted condition. Level 4 studies had a nonindependent

comparison of symptoms to the criterion standard and “grabbed” a sample of patients with the target condition and, perhaps, some healthy individuals. The excluded level 5 studies used a nonindependent comparison of symptoms to a standard of uncertain validity. When possible, we used published raw data from the identified articles to calculate sensitivity, specificity, likelihood ratios (LRs), and 95% CIs, with conventional definitions.⁴³

For articles in which data were presented stratified by multiple age groups, we present data for all age groups combined unless otherwise noted. Pooled analyses of multiple studies were not performed because of the small number and heterogeneity of studies available. In one study, published data were presented of the utility of physical examination findings compared with tympanocentesis for 2 individual clinicians who were examining 2 separate groups of children.²³ In that study, 64% of children presenting with acute symptoms (such as ear pain, fever, respiratory symptoms, vomiting, or diarrhea) underwent tympanocentesis, whereas 38% of patients without acute symptoms underwent tympanocentesis. Tympanocentesis was performed in any child suspected to have a middle ear effusion on pneumatic otoscopy. In our analysis of these data,²³ we calculated LRs excluding patients with perforation because these patients did not undergo tympanocentesis. To correct for verification bias, we made the conservative assumption that children who did not undergo tympanocentesis had normal-appearing ears (normal color, position, or mobility).⁴⁴ LRs were adjusted by the calculated verification fraction for each clinical sign subset (color, position, and mobility). The correction for verification bias protects against overly optimistic estimates of the examiner’s ability to rule out AOM and overly pessimistic estimates of the ability to rule in AOM. Because the color of the tympanic membrane appeared to have ordinal properties (eg, normal, slightly red, distinctly red, cloudy), we described the overall accuracy of this finding by the area under the receiver operating characteristic curve.

RESULTS

From the 397 references initially identified, we found 6 articles that satisfied inclusion criteria. This included 1 article concerning precision, 4 articles on accuracy of symptoms, and 1 article on accuracy of signs (Table 37-1).^{4,16,23,35,36,41}

Precision of Symptoms and Signs

To our knowledge, no studies concerning precision of symptoms have been published, and there are only a few studies on precision of signs. A comparison of diagnoses among practitioners would be important, especially during training, when medical students and house staff learn to interpret otoscopic findings from their instructors. Recently, Steinbach et al⁴ compared diagnoses of AOM among pediatric residents with diagnoses made by otolaryngologists. Complete examinations were available for 43 children, but the study found only fair agreement between the residents and the otolaryngologists. Overall agreement on diagnosis of AOM between the 2 types of practitioners had a κ statistic of 0.30 (fair). κ Statistics on tympanic membrane features such as erythema, color, effusion, mobility, and position were fair to slight ($\kappa = 0.40, 0.40, 0.31, 0.21$, and 0.16 , respectively). Correlations between pediatric residents and otolaryngologists comparing tympanometry in the detection of an effusion were also fair ($\kappa = 0.25$ and 0.30 , respectively).

Accuracy of Symptoms and Signs

Symptoms

Sensitivity, specificity, and positive and negative LRs derived from articles that examined the role of symptoms in the diagnosis of AOM are included in Table 37-2.^{16,35,36,41} The presence of ear pain appears to be the symptom most useful in making the diagnosis of AOM. Ear pain has a positive LR (LR+) of 3.0 to 7.3 but is present in only 50% to 60% of children with AOM. With a baseline prevalence for

Table 37-1 Studies Meeting Inclusion Criteria for Accuracy of Symptoms and Signs in Diagnosis of Acute Otitis Media

Source, y	Evidence Level	No. of Patients	Age Range, y	Criterion Standard	Limitations
Symptoms					
Niemela et al, ¹⁶ 1994	4	354	1 mo-15 y	Clinical diagnosis	Majority of children examined by specialists Children had a high incidence of recurrent acute otitis media Not blinded
Heikkinen and Ruuskanen, ³⁵ 1995	4	302	0.6-4.2 y	Clinical diagnosis	Not blinded
Ingvarsson, ³⁶ 1982	4	171	0-15 y	Clinical diagnosis	Referred to otolaryngologist for otalgia Not blinded
Kontiokari et al, ⁴¹ 1998	4	138	0.6-6.9 y	Clinical diagnosis	Not blinded
Signs					
Karma et al, ²³ 1989	3	2911	6 mo-2.5 y	Tympanocentesis	All examinations performed by either 1 pediatrician or 1 otolaryngologist

AOM of 20% among children aged 5 years or younger who make an acute pediatric office visit (estimated from the National Ambulatory Medical Care Survey), the presence of ear pain increases the probability of AOM to approximately 43% to 65%.

Fever is often cited as a primary symptom of AOM^{27,28} but shows variability in usefulness. One study shows that the likelihood slightly increases with a fever, but 2 studies found no effect, with the positive LR (LR+) approaching 1.0. The absence of fever seems to confer little change in the likelihood of AOM.

Kontiokari et al⁴¹ examined the ability of parents to predict whether their child had AOM. Parents were fairly accurate and showed similar ability to predict that their child did have AOM (LR+, 3.4) and that their child did not have AOM (LR-, 0.4). These findings are partially tempered by the fact that the physicians were not blinded to parental predictions, and this may have biased their ultimate diagnoses. We suspect that parents learn from their children's symptoms with each febrile or upper respiratory tract illness, so that more experienced parents may have better diagnostic acumen, but the effect of parental experience on accuracy and LR of diagnosing otitis media has not been evaluated. Thus, we do not know whether parents of children with frequent infections of any type are more or less

able to accurately assess ear involvement with each childhood illness episode.

A final symptom that deserves mention is ear pulling. Ear pulling has long been debated as a possible sign of AOM because parents and primary caregivers frequently observe this phenomenon.⁵ Many physicians have been taught that ear pulling is not a useful sign because children pull at their ears because "they are there." In the study by Niemela et al,¹⁶ "ear rubbing" appeared to have some predictive ability for the diagnosis of AOM (LR+, 3.3; 95% CI, 2.1-5.1). The only other study that we know of that has addressed this symptom is a small but often-referenced study by Baker,³⁰ who examined 100 consecutive children with a chief complaint of ear pulling; 20 children had ear pulling as their sole complaint, whereas 80 children had other symptoms. Of the 20 children with ear pulling as the sole complaint, none met Baker's unspecified criteria for AOM compared with 12 of the other 80 children.

Any conclusions about symptoms that can be drawn from the studies in Table 37-2 are limited by the study designs. Two of the 4 studies^{16,36} involve "spectrum bias," in which a spectrum of patients are used who are not representative of the population as a whole. Failure to incorporate an appropriate spectrum of patients can affect the sensitivity and specificity of findings.⁴⁵⁻⁴⁷ In the 2 studies

Table 37-2 Accuracy of Symptoms

Source and Symptoms	Sensitivity, %	Specificity, %	LR+ (95% CI)	LR- (95% CI)
Niemela et al, ¹⁶ 1994				
Ear pain	54	82	3.0 (2.1-4.3)	0.6 (0.5-0.7)
Ear rubbing	42	87	3.3 (2.1-5.1)	0.7 (0.6-0.8)
Fever	40	48	0.8 (0.6-1.0)	1.2 (1.0-1.5)
Cough	47	45	0.9 (0.7-1.1)	1.2 (0.9-1.4)
Rhinitis	75	43	1.3 (1.1-1.5)	0.6 (0.4-0.8)
Excessive crying	55	69	1.8 (1.4-2.3)	0.7 (0.5-0.8)
Poor appetite	36	66	1.1 (0.8-1.4)	1.0 (0.8-1.1)
Vomiting	11	89	1.0 (0.6-1.8)	1.0 (0.9-1.1)
Sore throat	13	74	0.5 (0.3-0.8)	1.2 (1.1-1.3)
Headache	9	76	0.4 (0.2-0.7)	1.2 (1.1-1.3)
Heikkinen and Ruuskanen, ³⁵ 1995				
Ear pain	60	92	7.3 (4.4-12)	0.4 (0.4-0.5)
Fever	69	23	0.9 (0.8-1.0)	1.4 (0.9-2.0)
Cough	84	17	1.0 (0.9-1.1)	1.0 (0.6-1.6)
Rhinitis	96	8	1.0 (1-1.1)	0.5 (0.2-1.4)
Restless sleep	64	51	1.3 (1.1-1.6)	0.7 (0.5-0.9)
Ingvarsson, ³⁶ 1982				
Ear pain	100	NA	NA	NA
Fever	79	70	2.6 (1.9-3.6)	0.3 (0.2-0.5)
Upper respiratory tract infection	96	29	1.4 (1.2-1.6)	0.3 (0.2-0.5)
Kontiokari et al, ⁴¹ 1998				
Parental suspicion of AOM	70	80	3.4 (2.8-4.2)	0.4 (0.3-0.5)

Abbreviations: AOM, acute otitis media; CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio; NA, not applicable.

identified in this analysis, patients were often treated by specialists and had a higher incidence of recurrent AOM or chronic OME. These patients may differ from those in primary care clinics, and this can potentially affect the generalizability of the results.

Another significant design limitation in all 4 included studies is their use of a clinical diagnosis of AOM, rather than tympanocentesis, as the criterion standard. Because the diagnosis of AOM potentially requires the presence of the symptoms that are being examined, an “incorporation bias” can occur when tympanocentesis is not performed as a confirmatory test. Incorporation bias⁴⁶ typically overestimates sensitivity and specificity. This bias may be further exaggerated because examiners who make the diagnosis of AOM also elicit the history in a nonblinded fashion. The bias created by using a clinical diagnosis as the criterion standard should improve the LRs for the symptoms; if that is the case, then it is possible that few symptoms would prove themselves independently useful in methodologically stronger studies.

Signs

Table 37-3 presents the results from the only study that has examined signs in the diagnosis of AOM.²³ The selective performance of tympanocentesis in this study created verification bias, which overestimates sensitivity and underestimates specificity and LR+.^{45,48} Fortunately, the investigators provided clinical examination findings for all patients, allowing us to correct for verification bias. This study suggests that a tympanic membrane that is cloudy (adjusted LR+, 34), bulging (adjusted LR+, 51), or distinctly immobile (adjusted LR+, 31) is highly suggestive of AOM. In contradiction to

what is often taught to physicians in training, a tympanic membrane that is distinctly red, defined as “hemorrhagic, strongly red, or moderately red,” also suggests otitis media (adjusted LR+, 8.4), whereas a drum that is only slightly red (adjusted LR+, 1.4) is not helpful. These data suggest that color of the tympanic membrane can be treated as an ordinal variable, ranging from normal through redness to cloudy (Table 37-3), with the likelihood of AOM increasing with the intensity of redness (the area under the receiver operating characteristic curve as a measure of accuracy of tympanic membrane color is 0.88 [SE, 0.003]).

After correction for verification bias, normal color or normal mobility make otitis media much less likely (LR = 0.2 for both). Given a baseline prevalence of 20% among children at an acute office visit, the probability of AOM decreases to less than 5% when the tympanic membrane is normal in either color or mobility. The independence of the findings of color, position, and mobility has not been assessed. Although it would seem that abnormalities in 2 or all 3 of these components would be more important than the finding of just 1 abnormality, we cannot quantify the effect of increasing numbers of abnormal findings.

Means of Improvement

Because AOM is so prevalent in the pediatric population and more accurate diagnosis of AOM can potentially lead to a decrease in antibiotic use and other costs, the improvement of diagnostic skills for AOM is clearly important. This improvement can be achieved by using more standardized diagnostic criteria and by improving diagnostic skills. A survey by Rosenfeld⁸ suggested that application of the AHRQ recommended criteria for AOM could reduce the rate of diagnosis of AOM by more than 20% by excluding cases that do not have evidence of a middle ear effusion.

Tools to improve diagnostic skills include teaching otoscopes that have 2 viewing areas,⁴⁹ videotapes, mannequin models, computer- and Web-based applications, and the use of more controlled settings, such as children undergoing myringotomy procedures. The American Academy of Pediatrics, for example, supports a multimedia “virtual classroom” Web site designed to help clinicians improve their skills in the diagnosis and treatment of otitis media.⁵⁰

Several studies have documented that clinicians can improve their diagnostic accuracy by practicing pneumatic otoscopy in children who are scheduled to undergo myringotomy.^{37,51} In this setting, clinicians perform ear examinations before anesthetization and in the operating room and compare their findings with the results of myringotomy. In addition, clinicians receive feedback from skilled, previously validated otoscopists. Pichichero and Poole⁵² have demonstrated that videotaped pneumatostoscopic examinations and infant mannequin models may be used to assess and potentially improve accuracy in the diagnosis of AOM and the performance of tympanocentesis.

Despite studies suggesting that diagnostic accuracy in AOM can be improved, current training remains poor. A recent survey by Steinbach and Sectish³ revealed that only

Table 37-3 Accuracy of Signs²³

Signs	Unadjusted LR+	Adjusted LR+ (95% CI) ^a
Color		
Cloudy	11	34 (28-42)
Distinctly red ^b	2.6	8.4 (6.7-11)
Slightly red	0.4	1.4 (1.1-1.8)
Normal	0.1	0.2 (0.19-0.21)
Position		
Bulging	20	51 (36-73)
Retracted	1.3	3.5 (2.9-4.2)
Normal	0.4	0.5 (0.49-0.51)
Mobility		
Distinctly impaired	8.4	31 (26-37)
Slightly impaired	1.1	4.0 (3.4-4.7)
Normal	0.04	0.2 (0.19-0.21)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

^aResults reported by Karma et al²³ were calculated by combining data reported from 2 groups. Results are rounded so that precision is not overstated and results remain clinically meaningful.

^b“Distinctly red” was described qualitatively as “hemorrhagic, strongly red, or moderately red.”

59% of pediatric residency programs currently provide a formal curriculum (defined as a “structured and consistent part of the residency program, not an occasional occurrence”) for training residents in the diagnosis and treatment of AOM. The formal curriculum that is provided usually consists of fewer than 3 didactic lectures per year, with limited assessment of resident performance.

CLINICAL SCENARIO—RESOLUTION

This child is certainly at risk of AOM because he is in the age group in which AOM is common, he has a history of AOM, and he has had a preceding upper respiratory tract infection. None of his presenting symptoms are predictive of AOM. On examination, the left ear appears normal. Cerumen is correctly removed from the right ear to better visualize the drum. This drum was described by the physician as “slightly” red but with normal motility. The precision of diagnosing an ear as slightly red as opposed to “distinctly” red is not known; however, according to the work of Karma et al,²³ a slightly red tympanic membrane does not have a high enough LR to independently confirm AOM, nor does it meet standardized criteria for AOM (such as from the AHRQ). The slightly red drum may be related to irritation from the cerumen removal and is not suggestive of AOM (probability is only 26%, given a baseline prevalence of 20% and an LR of 1.4 for a slightly erythematous membrane). The normal motility more definitively suggests that AOM is not likely in this scenario (probability would be only 5%, with a baseline prevalence of 20% and an LR of 0.2).

THE BOTTOM LINE

The diagnosis of AOM can be difficult, and studies examining this condition are somewhat limited. The studies we reviewed suggest that ear pain may be an important symptom but that other symptoms are not reliable. Although physical examination results are limited by the existence of only 1 well-performed study, a tympanic membrane that is cloudy, bulging, or distinctly immobile is highly suggestive of AOM. The presence of a distinctly red tympanic membrane also appears useful, although not as important as cloudiness of the tympanic membrane. Children with normal color and mobility of their tympanic membranes are much less likely to have otitis media than those with abnormalities. The discovery that erythema may be useful contradicts the instruction many clinicians receive and therefore deserves further study.

Many of the studies on the accurate diagnosis of AOM are limited by spectrum bias that affects generalizability and by lack of an acceptable criterion standard. These limitations are difficult to overcome. For example, it would be difficult to design a study in which tympanocentesis can be performed in children with a low suspicion for AOM. On the other hand, including data on all patients, as in the study by Karma et al²³ (Table 37-1), allows investigators to conduct practical studies with correction for verification bias that improves their validity. Future studies can be improved by using a gen-

eral population of at-risk children, more standardized diagnostic criteria, and independent examinations by blinded examiners. Studies also need to assess the precision and accuracy of characterizing physical findings, as Karma et al²³ have done, in an ordinal rather than dichotomous manner (eg, describing color as normal, slightly red, or distinctly red rather than just normal vs red). Because we do not know the relative importance of multiple abnormal findings vs 1 abnormal finding, an assessment of the independent importance of color, position, and mobility would allow clinicians to properly weigh the relative importance of these findings and, perhaps, lead to the development of a grading scheme that permits more accurate estimates of the likelihood of otitis media.

Despite the limitations of the current studies, we recommend that pneumatic otoscopy be performed when otitis media is considered to assess not just drum color and appearance but also mobility. Clinicians need to appreciate the amount of uncertainty in the diagnosis of AOM and how this may contribute to their decision to treat or not treat with antibiotics. Standard criteria for AOM, such as the AHRQ guidelines, which include the detection of a middle ear effusion, should also be considered because these can result in more uniform diagnoses and, it is hoped, decrease the rate of overdiagnosis. The use of training videos and other techniques may improve physical examination performance, but this will be more helpful after more studies have established the relationship between signs and the diagnosis of AOM.

Author Affiliations at the Time of the Original Publication

Departments of Medicine and Pediatrics, Vanderbilt University Medical Center, Nashville, Tennessee (Dr Rothman); and Departments of Medicine and Pediatrics, Duke University Medical Center (Dr Owens), and Durham Veterans Affairs Medical Center and Duke University Medical Center (Dr Simel), Durham, North Carolina.

Acknowledgments

We thank Peter Blomgren, MD, Amelia Drake, MD, David Witsell, MD, MHS, and Michael Steiner, MD, for their helpful reviews of an early version of the manuscript.

REFERENCES

- Rosenfeld R, Bluestone C. *Evidence Based Otitis Media*. St Louis, MO: BC Decker Inc; 1999.
- Bluestone C, Klein J. *Otitis Media in Infants and Children*. 3rd ed. New York, NY: WB Saunders; 2001.
- Steinbach WJ, Sectish TC. Pediatric resident training in the diagnosis and treatment of acute otitis media. *Pediatrics*. 2002;109(3):404-408.
- Steinbach WJ, Sectish TC, Benjamin DK Jr, Chang KW, Messner AH. Pediatric residents' clinical diagnostic accuracy of otitis media. *Pediatrics*. 2002;109(6):993-998.
- Ruuskanen O, Heikkinen T. Otitis media: etiology and diagnosis. *Pediatr Infect Dis J*. 1994;13(1 suppl 1):S23-S26.
- Marcy M. *Management of Acute Otitis Media*. Rockville, MD: Agency for Healthcare Research and Quality; 2001:1-159.
- Klein JO, Pelton JS. Epidemiology, pathogenesis, clinical manifestations, and complications of acute otitis media. In: Feigin RD, Friedman EM, eds. *Up-to-Date*. Available at: <http://www.uptodate.com>. Accessed June 14, 2002. Subscription required.

8. Rosenfeld RM. Diagnostic certainty for acute otitis media. *Int J Pediatr Otorhinolaryngol.* 2002;64(2):89-95.
9. Gonzalez-Vallejo C, Sorum PC, Stewart TR, Chessare JB, Mumpower JL. Physicians' diagnostic judgments and treatment decisions for acute otitis media in children. *Med Decis Making.* 1998;18(2):149-162.
10. Hayden GF. Acute suppurative otitis media in children: diversity of clinical diagnostic criteria. *Clin Pediatr (Phila).* 1981;20(2):99-104.
11. Chan LS, Takata GS, Shekelle P, Morton SC, Mason W, Marcy SM. Evidence assessment of management of acute otitis media, II: research gaps and priorities for future research. *Pediatrics.* 2001;108(2):248-254.
12. Pichichero ME. Acute otitis media, I: improving diagnostic accuracy. *Am Fam Physician.* 2000;61(7):2051-2056.
13. Weiss JC, Yates GR, Quinn LD. Acute otitis media: making an accurate diagnosis. *Am Fam Physician.* 1996;53(4):1200-1206.
14. Arola M, Ruuskanen O, Ziegler T, et al. Clinical role of respiratory virus infection in acute otitis media. *Pediatrics.* 1990;86(6):848-855.
15. Pichichero ME. Acute otitis media, II: treatment in an era of increasing antibiotic resistance. *Am Fam Physician.* 2000;61(8):2410-2416.
16. Niemela M, Uhari M, Jounio-Ervasti K, Luotonen J, Alho OP, Vierimaa E. Lack of specific symptomatology in children with acute otitis media. *Pediatr Infect Dis J.* 1994;13(9):765-768.
17. Pelton SI. Otoscopy for the diagnosis of otitis media. *Pediatr Infect Dis J.* 1998;17(6):540-543.
18. Cavanaugh RM Jr. Obtaining a seal with otic specula: must we rely on an air of uncertainty? *Pediatrics.* 1991;87(1):114-116.
19. Barriga F, Schwartz RH, Hayden GF. Adequate illumination for otoscopy: variations due to power source, bulb, and head and speculum design. *Am J Dis Child.* 1986;140(12):1237-1240.
20. Cavanaugh RM Jr. Pediatricians and the pneumatic otoscope: are we playing it by ear? *Pediatrics.* 1989;84(2):362-364.
21. Schwartz RH, Rodriguez WJ, McAveney W, Grundfast KM. Cerumen removal: how necessary is it to diagnose acute otitis media? *Am J Dis Child.* 1983;137(11):1064-1065.
22. Singer AJ, Sauris E, Viccellio AW. Ceruminolytic effects of docusate sodium: a randomized, controlled trial. *Ann Emerg Med.* 2000;36(3):228-232.
23. Karma PH, Penttila MA, Sipila MM, Kataja MJ. Otoscopic diagnosis of middle ear effusion in acute and non-acute otitis media, I: the value of different otoscopic findings. *Int J Pediatr Otorhinolaryngol.* 1989;17(1):37-49.
24. Whited JD, Grichnik JM. Does this patient have a mole or a melanoma? *JAMA.* 1998;279(9):696-701.
25. Metlay JP, Kapoor WN, Fine MJ. Does this patient have community-acquired pneumonia? diagnosing pneumonia by history and physical examination. *JAMA.* 1997;278(17):1440-1445.
26. Zitelli BJ, Davis HW. *Atlas of Pediatric Physical Diagnosis.* 3rd ed. St Louis, MO: Mosby-Wolfe; 1997.
27. Behrman RE, Kliegman R, Jenson HB, Nelson WE. *Nelson Textbook of Pediatrics.* 16th ed. Philadelphia, PA: WB Saunders Co; 2000.
28. Goroll AH, Mulley AG. *Primary Care Medicine: Office Evaluation and Management of the Adult Patient.* 4th ed. Philadelphia, PA: Lippincott Williams & Wilkins; 2000.
29. Noble J, Greene HL. *Textbook of Primary Care Medicine.* 3rd ed. St Louis, MO: Mosby; 2001.
30. Baker RB. Is ear pulling associated with ear infection? *Pediatrics.* 1992;90(6):1006-1007.
31. Bluestone CD, Cantekin EI. Design factors in the characterization and identification of otitis media and certain related conditions. *Ann Otol Rhinol Laryngol.* 1979;88(5 pt 2 suppl 60):13-28.
32. Cantekin EI, Bluestone CD, Fria TJ, Stool SE, Beery QC, Sabo DL. Identification of otitis media with effusion in children. *Ann Otol Rhinol Laryngol.* 1980;89(3 pt 2):190-195.
33. Finitzo T, Friel-Patti S, Chinn K, Brown O. Tympanometry and otoscopy prior to myringotomy: issues in diagnosis of otitis media. *Int J Pediatr Otorhinolaryngol.* 1992;24(2):101-110.
34. Hayden GF, Schwartz RH. Characteristics of earache among children with acute otitis media. *Am J Dis Child.* 1985;139(7):721-723.
35. Heikkinen T, Ruuskanen O. Signs and symptoms predicting acute otitis media. *Arch Pediatr Adolesc Med.* 1995;149(1):26-29.
36. Ingvarsson L. Acute otalgia in children—findings and diagnosis. *Acta Paediatr Scand.* 1982;71(5):705-710.
37. Kaleida PH, Stool SE. Assessment of otoscopists' accuracy regarding middle-ear effusion: otoscopic validation. *Am J Dis Child.* 1992;146(4):433-435.
38. Mains BT, Toner JG. Pneumatic otoscopy: study of inter-observer variability. *J Laryngol Otol.* 1989;103(12):1134-1135.
39. Pukander J. Clinical features of acute otitis media among children. *Acta Otolaryngol.* 1983;95(1-2):117-122.
40. Toner JG, Mains B. Pneumatic otoscopy and tympanometry in the detection of middle ear effusion. *Clin Otolaryngol.* 1990;15(2):121-123.
41. Kontiokari T, Koivunen P, Niemela M, Pokka T, Uhari M. Symptoms of acute otitis media. *Pediatr Infect Dis J.* 1998;17(8):676-679.
42. Gimsing S, Bergholtz LM. Otoscopy compared with tympanometry. *J Laryngol Otol.* 1983;97(7):587-591.
43. Simel DL, Samsa GP, Matchar DB. Likelihood ratios with confidence: sample size estimation for diagnostic test studies. *J Clin Epidemiol.* 1991;44(8):763-770.
44. Simel DL, Halvorsen RA Jr, Feussner JR. Quantitating bedside diagnosis: clinical evaluation of ascites. *J Gen Intern Med.* 1988;3(5):423-428.
45. Begg CB. Biases in the assessment of diagnostic tests. *Stat Med.* 1987;6(4): 411-423.
46. Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *N Engl J Med.* 1978;299(17):926-930.
47. Mulherin SA, Miller WC. Spectrum bias or spectrum effect? subgroup variation in diagnostic test evaluation. *Ann Intern Med.* 2002;137(7):598-602.
48. Begg CB, Greenes RA. Assessment of diagnostic tests when disease verification is subject to selection bias. *Biometrics.* 1983;39(1):207-215.
49. Teele DW, Klein JO. Use of a teaching pneumatic otoscope. *JAMA.* 1979;242(24):2664-2665.
50. Berman S, Stool S, Johnson C, et al. *American Academy of Pediatrics Otitis Media With Effusion Toolkit.* Denver, CO: University of Colorado Health Sciences Center; 1997. Updated 2002. <http://www.aap.org/otitismedia/www/>. Accessed June 7, 2008.
51. Silva AB, Hotaling AJ. A protocol for otolaryngology-head and neck resident training in pneumatic otoscopy. *Int J Pediatr Otorhinolaryngol.* 1997;40(2-3):125-131.
52. Pichichero ME, Poole MD. Assessing diagnostic accuracy and tympanocentesis skills in the management of otitis media. *Arch Pediatr Adolesc Med.* 2001;155(10):1137-1142.

Prepared by David L. Simel, MD, MHS, and Russell Rothman, MD

Reviewed by Sheri Keitz, MD

CLINICAL SCENARIO

A 12-month-old boy is brought to the clinic by his parents for evaluation of a possible ear infection. The parents state that the child has had fever, with temperature to 38.5°C, moderate irritability, and decreased appetite. They also note that the child has had 2 previous ear infections and that he is now acting as he did when he had the previous infections, raising their concern about another possible ear infection. On examination, the child is moderately irritable and crying. His left tympanic membrane appears to be grey, with good mobility on pneumatic otoscopy. His right tympanic membrane displays both distinctly red and distinctly impaired mobility on pneumatic otoscopy.

UPDATED SUMMARY ON ACUTE OTITIS MEDIA

Original Review

Rothman R, Owens T, Simel DL. Does this child have acute otitis media? *JAMA*. 2003;290(12):1633-1640.

UPDATED LITERATURE SEARCH

We repeated the literature search through April 2006 on acute otitis media and reviewed 66 new potential titles, of which 14 were promising enough for further evaluation. We attempted to identify prospective studies of children evaluated for possible acute otitis media; of the 14 titles, 8 full articles were retrieved and 1 was retained.¹ Because tympanocentesis is the best reference standard, we cross-checked the original 66 articles with the text word "tympanocentesis" and found 2 articles that described a clinical score using a combination of findings.^{2,3} We retained only the earlier of these 2 articles because it presented data in a fashion that allowed a calculation of likelihood ratios (LRs).

NEW FINDINGS

Healthy children who cry are unlikely to have red tympanic membranes (<3%).

Details of the Update

In a study of healthy infants and toddlers receiving injections for vaccinations, the color of the tympanic membrane was compared before and after the injection (which, together with the original examination, often induced crying).¹ On examination, clinicians rated the degree of crying and the color of the tympanic membrane on an 8-point scale from shades of clear or grey to pink and red. Although the clinicians undoubtedly knew the study hypothesis and could have been biased, the color was compared with a standardized chart developed from a digital computer program. Crying healthy children almost never had "lightly red" ears (2/242 tympanic membranes; 95% confidence interval [CI], 0%-3%) and none had frankly red ears. On the other hand, the tympanic membranes of children who cried more during the second examination had a greater likelihood of a 2-color increment increase toward faint pink/pink compared with those of children who did not have increased crying (19% vs 5%).

A large group of children at an Israeli medical center underwent tympanocentesis as part of an antibiotic clinical trial.² The investigators created a clinical score (Table 37-4) with face validity, though it has not been validated. The score combined the results of the child's temperature, parental assessment of irritability, and tugging of the ears, together with the clinician's assessment of redness and bulging membranes. Each item was graded on a 0 to 3 ordinal scale and summed across domains to create the clinical score (range, 0-15). The reference standard was the result of middle ear fluid culture, which was positive for 75% of children (*Haemophilus influenzae* or *Streptococcus pneumoniae*), but 37% of the patients had received antibiotics before the tympanocentesis. This bias could misclassify affected patients as not having otitis media, but it is impossible to know whether sensitivity or specificity is more affected because antibiotics could also affect the clinical score.

IMPROVEMENTS IN THE DATA PRESENTED IN THE ORIGINAL PUBLICATION

None.

CHANGES IN THE REFERENCE STANDARD

None.

Table 37-4 Unvalidated Clinical Score for Acute Otitis Media^a

Score	Temperature (°C)	Irritability	Tugging	Redness	Bulging
0	<38	Absent	Absent	Absent	Absent
1	38-38.5	Mild	Mild	Mild	Mild
2	38.6-39	Moderate	Moderate	Moderate	Moderate
3	>39	Severe	Severe	Severe	Severe

^aThe score is obtained by summing the individual scores for each finding. A maximum score is 15. When both ears are involved, use the score from the worst ear.

RESULTS OF LITERATURE REVIEW

Univariate Findings for Acute Otitis Media

We found no new valid data on the LR of individual symptoms and signs for acute otitis media.

Multivariate Findings for Acute Otitis Media

Overall, clinical scores (see Table 37-4) greater than or equal to 9 had an LR of 1.3 (95% CI, 1.0-1.7), 6 to 8 had an LR of 0.94 (95% CI, 0.70-1.3), and scores less than or equal to 5 had an LR of 0.32 (95% CI, 0.16-0.62), which suggests that the score is not particularly useful in diagnosing acute otitis. The utility of the clinical score may have been diluted by the inclusion of clinical signs in the score. Only tympanic redness and bulging had individual scores that were statistically greater among those with positive culture results vs negative culture results. The importance of tympanic redness or bulging fits with the criteria recommended by the Agency for Healthcare Research and Quality as reported in the original Rational Clinical Examination article on acute otitis media.⁴ In the study, more than 70% of patients had acute otitis, which is much greater than expected in the normal population. This spectrum of patients, with high rates of otitis, could overestimate the sensitivity and underestimate the specificity of the scale. The data suggest that the clinical score from Table 37-4 may have too many variables because temperature, irritability, and tugging did not differ between patients with culture-positive and culture-negative results.

EVIDENCE FROM GUIDELINES

The American Academy of Pediatrics and the American Academy of Family Practice released new guidelines on the diagnosis and treatment of acute otitis media.⁵ These guidelines for diagnosing acute otitis are similar to the guidelines sponsored by the Agency for Healthcare Research and Quality.⁴

CLINICAL SCENARIO—RESOLUTION

The prevalence of acute otitis media in children aged 0 to 2 years at ambulatory visits is about 20%. The parental suspicion of acute otitis media may be modestly helpful; with an LR of 3.4, this would make the posttest probability of acute otitis 46%. The distinctly red tympanic membrane is likely not just related to crying and suggests an acute infection; with an LR of 8.4, this sign increases the posttest probability to 68%. A distinctly impaired tympanic membrane mobility on pneumatic otoscopy is the most helpful finding (LR, 31) and raises the posttest probability to 89%. The combination of these symptoms and signs may make acute otitis media even more likely.

REFERENCES FOR THE UPDATE

- Yamamoto LG, Sumida RN, Yano SS, Derauf DC, Martin PE, Eakin PJ. Does crying turn tympanic membranes red? *Clin Pediatr*. 2005;44(8):693-697.^a
- Leibovitz E, Satran R, Piglansky L, et al. Can acute otitis media caused by *Haemophilus influenzae* be distinguished from that caused by *Streptococcus pneumoniae*? *Pediatr Infect Dis J*. 2003;22(6):509-515.^a
- Polachek A, Greenberg D, Lavi-Givon N, et al. Relationship among peripheral leukocyte counts, etiologic agents and clinical manifestations of acute otitis media. *Pediatr Infect Dis J*. 2004;23(5):406-413.
- Marcy M. *Management of Acute Otitis Media*. Rockville, MD: Agency for Healthcare Research and Quality; 2001:1-159.
- American Academy of Pediatrics Subcommittee on Management of Acute Otitis Media. Diagnosis and management of acute otitis media. *Pediatrics*. 2004;113(5):1451-1465.

^aFor the Evidence to Support the Update for this topic, see <http://www.JAMAevidence.com>.

ACUTE OTITIS MEDIA—MAKE THE DIAGNOSIS

PRIOR PROBABILITY

The prevalence of clinically diagnosed otitis media is high, with a rate of 17.4% for all visits to US pediatricians for 0- to 24-month-old children, 18% for 2- to 5-year-old children, 10% for 6- to 10-year-old children, and 5.2% for 11- to 15-year-old children. A baseline prevalence of 20% is a reasonable anchor for child visits to the emergency department.

POPULATION FOR WHOM ACUTE OTITIS MEDIA SHOULD BE CONSIDERED

The diagnosis should be considered for a child complaining of ear symptoms. Among infants, the rapid onset of ear pulling, ear drainage, irritability, or fever should prompt an otoscopic evaluation for otitis media.

DETECTING THE LIKELIHOOD OF ACUTE OTITIS MEDIA

Healthy children who cry before and during the examination are unlikely to have distinctly red tympanic membranes. Therefore, discovering red tympanic membranes should not be attributed solely to crying. The most useful findings are tympanic membrane color, mobility, and position (Table 37-5).

REFERENCE STANDARD TESTS

Tympanocentesis is the reference standard, but most studies use a standardized clinical definition.

Table 37-5 Detecting the Likelihood of Acute Otitis Media

Signs ^a	LR+ (95% CI)
Tympanic membrane color	
Cloudy	34 (28-42)
Distinctly red	8.4 (6.7-11)
Slightly red	1.4 (1.1-1.8)
Normal	0.20 (0.19-0.21)
Tympanic membrane mobility ^b	
Distinctly impaired	31 (26-37)
Slightly impaired	4.0 (3.4-4.7)
Normal	0.20 (0.19-0.21)
Tympanic membrane position	
Bulging	51 (36-73)
Retracted	3.5 (2.9-4.2)
Normal	0.50 (0.49-0.51)
Symptoms	LR+ (95% CI) or Range LR- (95% CI) or Range
Parental suspicion of otitis media	3.4 (2.8-4.2) 0.4 (0.3-0.5)
Ear rubbing	3.3 (2.1-5.1) 0.7 (0.6-0.8)
Ear pain	3.0-7.3 0.4-0.6

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

^aLRs adjusted for verification bias.

^bAssessed with pneumatic otoscopy.

This page intentionally left blank

EVIDENCE TO SUPPORT THE UPDATE:

37

Otitis Media, Child

TITLE Can Acute Otitis Media Caused by *Haemophilus influenzae* be Distinguished From That Caused by *Streptococcus pneumoniae*?

AUTHORS Leibovitz E, Satran R, Piglansky L, et al.

CITATION *Pediatr Infect Dis*. 2003;22(6):509-514.

QUESTION Does a previously proposed clinical score based on symptoms and signs accurately identify infants and young children with acute otitis media?

DESIGN Prospective, nonconsecutive enrollment.

SETTING The study took place in an Israeli pediatric emergency department. All reported study subjects were enrolled in various antibiotic efficacy trials.

PATIENTS Infants and children (aged 3 to 36 months) treated in the emergency department during a 5-year period who had (1) symptoms of acute otitis (parental report of fever, irritability, and ear tugging) and signs (redness or bulging of the tympanic membrane); (2) illness duration less than or equal to 7 days; and (3) no tympanostomy tubes or spontaneous perforation of at least 24 hours.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

All children underwent an examination by an otolaryngologist who also performed a tympanocentesis. The score was based on a previously reported score, summed across 5 findings, with a maximum score of 15.¹ When both ears were involved, the score for the worse ear was used (Table 37-6).

Table 37-6 Clinical Score for Acute Otitis Media

Score	Temperature (°C)	Irritability	Tugging	Redness	Bulging
0	<38	Absent	Absent	Absent	Absent
1	38-38.5	Mild	Mild	Mild	Mild
2	38.6-39	Moderate	Moderate	Moderate	Moderate
3	>39	Severe	Severe	Severe	Severe

MAIN OUTCOME MEASURES

The means of the clinical score were calculated for patients according to an infection with *Haemophilus influenzae*, *Streptococcus pneumoniae*, or mixed infections. Viral cultures were not performed.

MAIN RESULTS

Of 372 enrolled study subjects with complete data (Table 37-7), 96% were younger than 2 years. The culture results were negative in 94 (25%) patients, but only 63% of patients had not received antibiotics in the preceding 72 hours.

Culture-positive patients had a higher clinical score than culture-negative patients (9.3 vs 8.4; $P = .01$). There was no difference in score between culture-positive and culture-negative children treated with antibiotics, whereas those who were antibiotic naive showed a statistical difference (culture-positive score 9.11 vs culture negative score 8.1; $P = .02$ [confidence intervals not provided]). The authors evaluated age as a second confounder. The differences in culture-positive patients' score vs culture-negative patients only barely reached statistical significance in infants aged 3 to 6 months (8.9 ± 2.7 vs 7.4 ± 2.0 ; $P = .05$). Among older children, the scores were not statistically different between culture-positive and culture-negative results for patients. The scores did not

Table 37-7 Clinical Data of the Patients

Score	All Patients No.	Received Previous Antibiotics ^a		Received no Previous Antibiotics ^a	
		LR+ (95% CI)	Prevalence, %	LR+ (95% CI)	Prevalence, %
≥9	201	1.3 (1.0-1.7)	65	1.10 (0.83-1.50)	49
6-8	140	0.94 (0.70-1.30)	26	1.40 (0.70-2.80)	43
≤5	31	0.32 (0.16-0.62)	9	0.26 (0.09-0.75)	8
					0.34 (0.14-0.82)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio.

^aEugene Leibovitz, MD, kindly provided the data that allowed calculation of the likelihood ratios for those who received antibiotics and those who were antibiotic naive.

distinguish between bacterial etiologies, which was an outcome unaffected by antibiotic exposure. When the individual components of the score were analyzed, the scores for redness and bulging were the only results statistically higher for culture-positive vs culture-negative patient results. Tympanic membranes were redder in patients with *H influenza* ($P = .001$) or *S pneumonia* ($P = .05$) compared with those with negative cultures. Similarly, tympanic membranes bulged more in children who were culture positive (*H influenza*, $P < .001$; *S pneumonia*, $P = .04$).

Because antibiotics could have affected the clinical findings, we calculated the likelihood ratios from data provided by the author. Among children who had not received antibiotics, 78% had a culture-positive result. There was no clinical difference in the likelihood ratios at any of the 3 specified thresholds, whether or not the child received antibiotics (Table 37-7).

CONCLUSIONS

LEVEL OF EVIDENCE Level 3.

STRENGTHS All these children received tympanocentesis and culture.

LIMITATIONS All the children were referred for enrollment in antibiotic treatment trials of otitis media.

With a 78% probability of otitis among antibiotic-naïve patients, the clinicians were effectively identifying the children most likely to have acute otitis media. The disproportionately high prevalence of children with scores greater than or equal to 9 is appropriate for a randomized clinical trial enrolling children according to their clinical diagnosis. However, it creates verification bias when using the data to determine the accuracy of the clinical diagnosis because children would have been referred for tympanocentesis only when they were highly likely to have otitis media. Typically, sensitivity is overestimated and specificity is underestimated with verification bias. In addition to verification bias, higher clinical scores may have led to earlier initiation of antibiotics. Among children who did not receive previous antibiotics, the clinical score provides little information. With a prevalence of 78% culture positive, a clinical score greater than or equal to 9 increases the probability of acute otitis media to 84%, whereas a clinical score less than or equal to 5 decreases the probability to 55%.

In other research, only redness and bulging have been shown as useful for diagnosing otitis media. The scoring system includes 3 other measures, none of which were statistically significant as independent items. By including those measures in the score, the clinical efficiency of the score system might have been decreased.

Reviewed by David L. Simel, MD, MHS

REFERENCE FOR THE EVIDENCE

- Dagan R, Leibovitz E, Greenberg D, Yagupsky P, Fliss DM, Leiberman A. Early eradication of pathogens from middle ear fluid is associated with improved clinical outcome. *Pediatr Infect Dis J*. 1998;17(9):776-782.

TITLE Does Crying Turn Tympanic Membranes Red?

AUTHORS Yamamoto LG, Sumida RN, Yano SS, Derauf DC, Martin PE, Eakin PJ.

CITATION *Clin Pediatr*. 2005;44(8):693-697.

QUESTION Among healthy infants and toddlers subjected to immunizations, does crying affect the color of the tympanic membrane?

DESIGN Prospective, convenience sample.

SETTING Pediatrics office.

PATIENTS Infants and toddlers (age ≤ 30 months) at routine healthy child checks during which immunizations were given.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

A physician examined the child's ears before immunization and rated the child's degree of crying and color of each tympanic membrane. Crying was assessed on an ordinal 0 to 4+ scale (0 = "no crying at all"; 4 = loudest and most intense crying ever heard). Color was rated on a 1 to 9 scale according to a sheet with color images produced by Adobe Photoshop (Adobe Systems, Inc, San Jose, California) with standardized red-blue-green values. The colors were described as no color, light gray, gray, faint pink, pink, darker pink, light red, red, and "can't see the tympanic membrane." A second independent examiner repeated the examination after the immunization.

MAIN OUTCOME MEASURES

Change in crying was compared to the color of the tympanic membrane perceived by the physician. The color had to be rated as pink or more red (≥ 5 on color scale) to be considered an "increase in redness."

MAIN RESULTS

Of the 121 children, 53 were not crying during the first examination compared with only 17 who were not crying during the second examination. At the second examination, 64 subjects were rated as crying the same or less compared with the rating by the first examiner. Only 2 tympanic membranes of the 242 examined ears exhibited light redness and none were frankly red. When children were crying more during the second exam-

ination, 19% had increased redness by 2 categories or more vs 5% of those crying the same or less ($P < .001$).

CONCLUSIONS

LEVEL OF EVIDENCE Level 4.

STRENGTHS Independent examiners.

LIMITATIONS The validity of the study depends on the colors chosen. However, standardizing the colors from a computer program is likely better than a subjective assessment of redness. The examining physicians were not blinded, so they likely knew the study hypothesis. The analysis was based on the number of ears examined rather than children examined, which increased the apparent sample size and might have inflated the statistical significance. Approximately 18 physi-

cians were used in 5 clinic sites, and a different physician performed the initial and second examinations. For better assessment of the validity of the color scale, it would have been helpful if the authors had assessed the level of physician agreement with a κ statistic.

The authors include a good discussion of some of the limitations to generalizability of their results. However, the near absence of red tympanic membranes strongly suggests that healthy children who cry do not develop frankly red tympanic membranes, though they can display pinkish hues. The authors appropriately noted that sick, febrile children might have higher rates of greater-intensity crying and more flushing even though fever is not a strong predictor of otitis media.

Reviewed by David L. Simel, MD, MHS

This page intentionally left blank

CHAPTER 38

Does This Patient Have Parkinson Disease?

Goutham Rao, MD

Laura Fisch, MD, MPH

Sukanya Srinivasan, MD, MPH

Frank D'Amico, PhD

Tadao Okada, MD

Carolyn Eaton, MD

Craig Robbins, MD, MPH

CLINICAL SCENARIO

A 68-year-old man presents with a 3-month history of right arm tremor at rest. His movements have been slower and he has difficulty getting out of a chair. Physical examination reveals rigidity in the upper limbs. He walks with small steps and has limited ability to swing his arms. His facial expressions are limited.

WHY ANSWER THIS QUESTION WITH A CLINICAL EXAMINATION?

With a prevalence estimated between 150 and 200 per 100 000, Parkinson disease (PD) is one of the most common neurologic disorders.¹ It is more prevalent in older persons, affecting 1% of those older than 65 years and 2% of those older than 85 years.²

Although common, the diagnosis of PD is challenging. Laboratory tests are not available and conventional imaging studies are not helpful. The best reference standard is, unfortunately, neuropathologic (depletion of brain stem pigmented neurons and proliferation of Lewy bodies).³ Serial neurologic evaluation with or without concomitant treatment can also be used.⁴ The response to an acute levodopa challenge has been used as a diagnostic tool. This test is problematic for a number of reasons: its sensitivity and specificity are low, acute levodopa administration is associated with significant adverse effects, there is lack of agreement on what constitutes a threshold response, and the test is expensive and inconvenient.⁵ The clinical examination, therefore, is the basis for initial diagnosis. Classic clinical features of PD include tremor at rest, bradykinesia, rigidity, and postural instability.

There is evidence that the accuracy of diagnosis in some settings is improving. In a 1991 study by Rajput et al⁶ among 41 patients diagnosed clinically with PD by neurologists, the disease was confirmed neuropathologically at autopsy in 31 (positive predictive value [PPV] of 76%). Hughes et al⁷ evaluated the accuracy of clinical diagnosis among 100 patients with PD, 86 of whom were followed up by neurologists, 7 by geriatricians, and 7 by internists. The diagnosis was confirmed at autopsy in 90 persons (PPV = 90%). Another study confirmed PD at autopsy among 72 of 73 patients (PPV = 99%) followed up by neurologists affiliated with a highly specialized movement disorders center.⁸ Despite these improvements and impressive results, it is important to keep in mind that the clinical diagnoses in these studies were often made during a long period and by physicians with a great deal of expertise and experience. The accuracy of clinical diagnosis in other settings is unclear. PD is still mistaken for other neurologic disorders. The most frequent misdiagnoses include progressive supranuclear palsy, multisystem atrophy (MSA) (encompasses the diagnoses Shy-Drager syndrome, olivo-

pontocerebellar atrophy, and striatonigral degeneration), and dementia with Lewy bodies.⁶ The differential diagnosis also includes essential tremor and vascular pseudoparkinsonism. Mistaking PD for other conditions can lead to inappropriate and ineffective treatment. Although a patient with essential tremor, for example, may benefit from a β -blocker, this treatment would have no effect on the tremor of PD. Inappropriate treatment based on misdiagnosis also delays the use of dopaminergic medications, which can decrease the severity of symptoms and disability.⁹

Mistaking other disorders for PD is also harmful. Dyskinesias, for example, appear in 15% to 85% of persons within 5 years of treatment with levodopa, and hallucinations occur in 20% of patients.¹⁰ There is also evidence that levodopa causes damage to dopamine neurons, leading to accelerated dopamine degeneration.⁵ Whether the initial diagnosis is correct or not, the disease has serious social and psychological consequences.^{11,12} In summary, the clinical examination is

Box 38-1 Typologic Classification of Tremors

REST TREMOR

Tremor occurring in a body part that is not voluntarily activated and when it is supported completely against gravity.

ACTION TREMORS

Postural

Tremor that occurs while voluntarily maintaining a position against gravity.

Kinetic

Tremor occurring during any voluntary movement.

1. *Simple*. Tremor occurring during voluntary movements that are not target-directed.
2. *Intention*. Tremor whose amplitude increases during visually guided movements (eg, finger-to-nose test).
3. *Task-Specific*. Tremor that appears or is exacerbated by specific tasks (eg, writing).
4. *Isometric*. Tremor that occurs during voluntary muscle contraction against a rigid stationary object (eg, squeezing examiner's hand).

Box 38-2 Three Common Tremor Syndromes

TREMOR OF PARKINSON DISEASE

Slow frequency (4-6/s) tremor at rest. Tremor inhibited during movement and sleep. Aggravated by emotional distress. "Pill-rolling quality."

CLASSIC ESSENTIAL TREMOR

Bilateral, usually symmetric postural or kinetic tremor. Family history of tremor is common. Attenuated by alcohol.

PHYSIOLOGIC TREMOR

Present to differing degrees in all subjects. Enhanced form is easily visible, mainly postural, and has a high frequency (8-12/s). No evidence of underlying neurologic disease. Cause is usually reversible (eg, caffeine).

important in suspected PD because no laboratory or radiologic tests are helpful diagnostically. Misdiagnosis of PD is associated with adverse effects.

Pathophysiologic Characteristics

It is important to distinguish between PD and parkinsonism. *Parkinsonism* refers to any clinical syndrome in which 2 or more features are present such as tremor, rigidity, and bradykinesia. Parkinson disease is a form of primary or idiopathic parkinsonism. Viral infections, environmental toxins, oxidative stress, and heredity have all been suspected as causes.¹³ Secondary or acquired parkinsonism has a variety of causes, including head trauma, cerebrovascular disease, and hydrocephalus.^{14,15} Secondary parkinsonism may persist for months after the drugs that caused it are discontinued. A thorough inquiry into past and current medication use, therefore, is essential when questioning patients presenting with parkinsonism. Parkinson disease begins as neurons and dopamine are lost from the substantia nigra and intracytoplasmic inclusions (Lewy bodies) appear. Symptoms appear when 70% to 80% of dopamine is lost.¹⁶

Symptoms and Signs

Nonspecific insidious symptoms, including generalized malaise, easy fatigability, and subtle personality changes, mark the onset of PD. These may occur years before the appearance of tremor, limb rigidity, bradykinesia, and postural instability.¹⁶ Numerous secondary manifestations appear unpredictably and are as varied as disordered sleep (42% of patients),¹⁷ constipation (50%), pain (50%), depression (40%), and dementia (20%).^{16,18} Signs typically begin unilaterally and then progress asymmetrically.

James Parkinson described the combination of tremor and bradykinesia as a shaking palsy.¹⁸ Seventy-five percent of patients complain initially of a tremor that usually occurs at rest in an upper extremity and is characterized by visible oscillations with a frequency of 4 to 6 per second. Tremor appears intermittently, disappearing during sleep and increasing in severity during times of emotional distress or anxiety. It is often described as pill-rolling, because a rhythmic movement is observed in the hand as the index finger flexes and extends against the thumb repetitively.¹⁹

Some basic features distinguish the tremor of PD from physiologic and essential tremors (Boxes 38-1 and 38-2).^{20,21}

Rigidity, refers to an involuntary stiffness of the skeletal muscles and is a common sign. Electromyogram assessment of parkinsonian patients reveals an alternating discharge pattern in opposing muscle groups, even at rest (eg, triceps and biceps). Resistance to movement of limbs may be smooth or interrupted. *Cog wheeling* refers to the jerky motion of limbs as constant force is applied across a joint, which is similar to the ratcheting of the cogs of gears as they click.²² Unlike rigidity, *spasticity* refers to a selective increase of tone of flexor muscles in the arms and extensor muscles in the legs²³ and suggests a diagnosis other than PD.

Bradykinesia refers to the overall slowing of active movement or slowness in initiating movement. The initial surge of motor activity is inadequate and movements are fragmented into a series of incremental steps. Postural instability in patients with PD presents as changes in gait and balance. Short and shuffling steps are often accompanied by festination. Loss of arm movements commonly appears. The patient may walk with the arms straight down, rather than swinging them back and forth. Gait disturbance is the major cause of disability in many patients. As postural reflex mechanisms are lost, patients become stooped and have a tendency to fall. Those with severe deficits are sometimes confined to a wheelchair or bed.

How to Elicit Signs

Tremor

Tremor can be defined as any rhythmic, involuntary oscillatory movement of a body part. The tremor classification is complex and has overlapping features in different disease syndromes. Nevertheless, the Movement Disorder Society has developed a classification system to help clinicians distinguish tremor types.²⁴ Tremors can be classified as *rest* or *action*.

The classification system divides tremors into 11 syndromes. Patients with PD typically have a slow (frequency of about 4-6/s) tremor at rest. It is easily observed by having patients position their hands on their lap. Physicians should be able to identify the key features of PD and essential and physiologic tremors (Box 38-2).

Precise measurement of tremor frequency and amplitude is sometimes used in diagnostic evaluation. This requires special devices and is beyond the scope of the clinical examination.

Bradykinesia

Involuntary muscle stiffness or rigidity can be shown if resistance to passive movement of the limbs is detected. With the patient relaxed, the examiner places his or her thumb across the antecubital fossa with one hand while passively flexing and extending the elbow several times with the other hand. Rigidity often increases with repeated flexion and extension movements. With cog wheeling, the examiner feels alternate periods of resistance and relaxation. With lead-pipe rigidity, the examiner feels smooth but increased muscle tone throughout passive flexion and extension.²⁵ Rigidity and cog wheeling may be felt in other large joints, but if detected in the arms, there is no need to confirm their presence elsewhere. Many patients with essential tremor manifest a rhythmic resistance to passive movements of a limb while there is voluntary action of another body part. This is not true cog wheeling but is known as Froment sign, which also appears in PD patients.²⁶

Bradykinesia

Bradykinesia refers to a decrease in the speed and amplitude of complex movements. Jobbagy et al²⁷ described 4 maneuvers designed to detect bradykinesia: tapping the fingers, twiddling, pinching and circling, and tapping with the heel (Figure 38-1). Twiddling refers to repeated rotation of

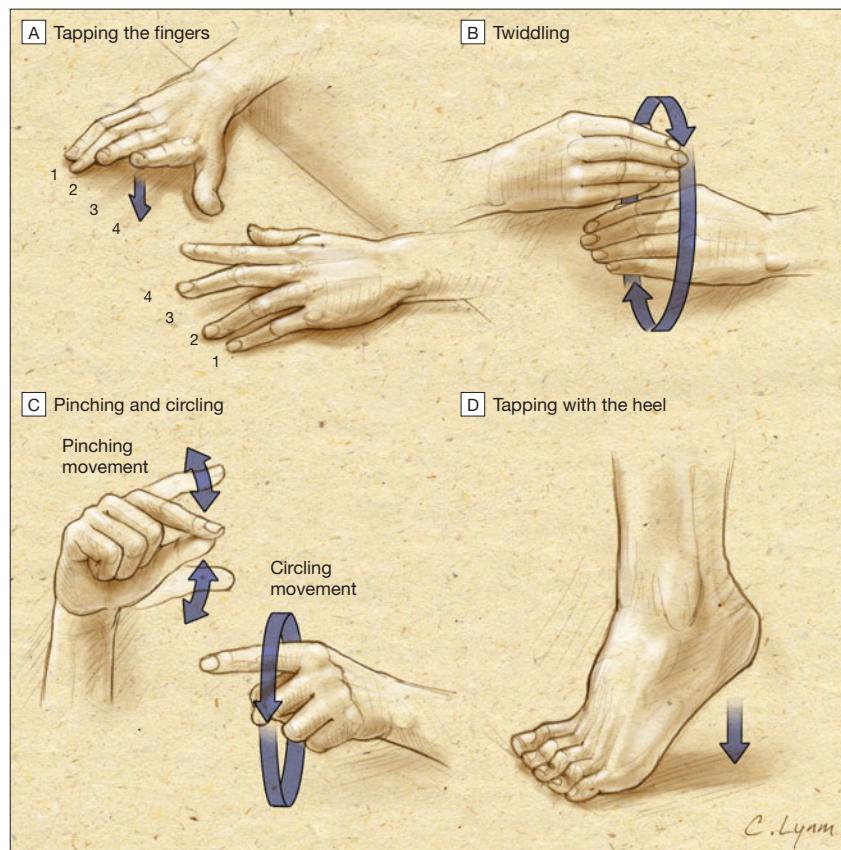


Figure 38-1 Maneuvers to Detect Bradykinesia

the hands in front of the body. The pinching and circling test is a sequence of 6 movements: pinching (opposing thumb and index finger) with the right hand and then with the left hand; circling (rotating the hand in a circle) with the right hand and then with the left hand; pinching with the right hand while simultaneously circling with the left; and pinching with the left hand while simultaneously circling with the right (Figure 38-1). Jobbagy et al²⁷ were able to quantify the performance of patients on these tasks by using a motion analyzer, although a specific threshold

“score” to define bradykinesia was not determined. However, poor performance of these maneuvers is easily detectable and clinicians can use them to confirm the presence of bradykinesia subjectively.

Glabella Tap Reflex

This reflex is tested by percussion of the forehead with the examiner’s index finger or by pulling a fold of skin between the thumb and index finger on the temple lateral to the external canthus and tapping with the thumb. The orbicularis oculi muscle reflexively contracts, causing both eyes to blink. The reflex blinking normally stops after tapping is repeated 5 to 10 times. Persistent blinking is a positive response sometimes referred to as Myerson sign.²⁸ Care should be taken to keep the examiner’s finger above the patient’s eyes to avoid blinking in response to visual threat (Figure 38-2).

Are These Features Found in Other Diseases?

The symptoms and signs of idiopathic PD overlap with those of other neurologic diseases, including MSA and progressive supranuclear palsy.

Like PD, MSA often presents with asymmetric rigidity and akinesia, but only a minority of patients have a resting tremor.²⁹ Half of patients with MSA present with autonomic dysfunction and cerebellar symptoms, and one-quarter demonstrate a transient response to levodopa.^{25,29} Similarly,

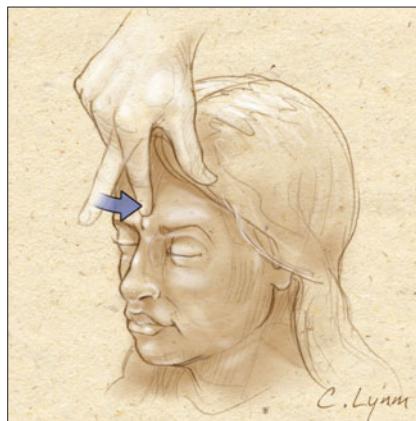


Figure 38-2 Glabella Tap Test

Table 38-1 Grade C Studies Included for Review^a

Source	No. of Subjects	Age, y, Mean (Range)	Patient Population	Reference Standard for Diagnosis of PD	Reason Study Not Grade A
Hughes et al ³⁴	100	64 (31-85)	Diagnosed clinically as having PD	Autopsy findings of depletion of nigral pigmented neurons and proliferation of Lewy bodies	Significant selection bias because patients studied were clinically diagnosed as having PD
Wenning et al ³⁵	138	61 (NA)	Diagnosed clinically as having PD or MSA	Autopsy findings consistent with PD or MSA	Significant selection bias because patients studied were clinically diagnosed as having PD or MSA
Pearce et al ³⁶	100	48 (NA)	Unselected inpatients and outpatients diagnosed as having PD and controls without known neurologic disease	Detailed neurologic examination	Samples of patients who obviously have the condition; comparisons nonindependent; small sample size
Duarte et al ³⁷	128	66 (30-89) ^b	Patients attending a movement disorders polyclinic for the first time	Detailed neurologic evaluation	Convenience sample including many individuals likely to have PD; small sample size
Mutch et al ³⁸	123				Nonindependent comparisons with unclear standard; samples of patients who obviously have the condition; small sample size
Cases	75 (57-89)	35	Diagnosed as having PD	Unclear standard used	
Controls	73 (71-76)	88	From general practices	Neurologic evaluation	
Meneghini et al ³⁹	108	NA	87 Inpatients with neurologic disorders and 21 patients without known neurologic disease	Detailed neurologic evaluation	Samples of patients who obviously have the condition (including many individuals likely to have PD and controls); small sample size; prone to observer bias

Abbreviations: MSA, multisystem atrophy; NA, not available; PD, Parkinson disease.

^aSee Table 1-7 for a summary of Evidence Grades and Levels.

^bFor 37 patients diagnosed as having PD only.

patients with progressive supranuclear palsy seldom present with tremor. Rigidity and postural instability, however, are common.³⁰

Parkinsonism is sometimes also a feature of Alzheimer disease.¹⁴ However, Alzheimer disease is easy to distinguish from PD because other features are much more prominent. Furthermore, unlike in PD, cognitive impairment is present at the onset of Alzheimer disease.

METHODS

Four of the authors (G.R., L.F., T.O., and C.E.) performed independent searches of the MEDLINE database (1966-2001), using a number of Medical Subject Headings ("exp tremor," "exp PD," "essential tremor") combined with the search terms and strategy used for The Rational Clinical Examination series.³¹

All relevant articles were retrieved. The resulting set of articles was divided into 3 parts, each of which was reviewed by a pair of authors. The reference lists of all articles were also carefully searched for additional articles. Articles were included for study if they met the following criteria: dealt primarily with the diagnosis of PD; included patients presenting with 1 or more typical parkinsonian symptoms or signs (eg, tremor, rigidity); final diagnosis confirmed by a suitable criterion standard, such as serial or detailed neurologic evaluation or pathologic confirmation at autopsy; and contained original data from which 2×2 tables could be extracted to calculate the sensitivity, specificity, positive likelihood ratio (LR+), and negative likelihood ratio (LR-) for different signs and symptoms. Because the number of suitable articles was small, additional inclusion criteria such as a minimum sample size or publication after a certain year were not used. However, the quality of articles included was assessed according to criteria previously developed for this series.³¹

The likelihood ratios (LRs) for different diagnostic features were calculated when not available in the original articles. Corresponding 95% confidence intervals (CIs) were determined by the method of Greenland and Robins.³² All values were rounded to 2 significant digits. When identical or similar diagnostic features appeared in more than 1 article and the patients were similar across studies in terms of demographics and illness characteristics, weighted summary LRs (pooled LRs) and the corresponding 95% CIs were calculated with the DerSimonian-Laird random-effects method.³³ We used MetaWin statistical software (version 2; Sinauer Associates, Sunderland, Massachusetts).

RESULTS

Quality of the Evidence

A total of 185 articles were reviewed. All authors agreed about which articles met our selection criteria. We chose 6 articles. Two articles^{34,35} included independent blind comparisons of symptoms and signs of a small number of

patients who had been diagnosed as having PD or MSA according to comparison of clinical records to pathologic results at autopsy. Because the patients studied had already been diagnosed clinically as having PD or MSA, selection bias was a serious problem. These 2 articles provided level 3 evidence, leading to grade C recommendations (see Table 1-7 for summary of Evidence Grades and Levels).

The remaining 4 articles³⁶⁻³⁹ had numerous methodologic biases. Although of lower methodologic quality, they can still be classified as containing level 3 evidence and providing grade C recommendations (Table 38-1).

Table 38-2 Symptoms Evaluated in Patients With Possible Parkinson Disease

Symptom	LR+ (95% CI)	LR- (95% CI)
Tremor		
Arms or legs shake		
Duarte et al ³⁷	1.4 (1.2-1.6)	0.25 (0.08-0.78)
Pearce et al ³⁶	17 (6.3-44)	0.24 (0.13-0.44)
Tremor as initial symptom ³⁴	1.3 (0.90-2.0)	0.60 (0.34-1.1)
Tremor of head or limbs ³⁹	11 (4.8-24)	0.26 (0.12-0.55)
Rigidity		
Muscle stiffness ³⁸	2.3 (1.3-4.3)	0.73 (0.54-0.97)
Paralysis or weakness ³⁹	1.3 (0.60-2.8)	0.93 (0.75-1.2)
Rigidity and bradykinesia ³⁹	4.5 (2.9-7.1)	0.12 (0.03-0.45)
Facies and General Symptoms or Historical Findings		
Face less expressive ³⁷	2.1 (1.4-3.2)	0.54 (0.35-0.84)
Feet freeze ³⁷	3.7 (2.1-6.7)	0.55 (0.39-0.79)
Impaired consciousness ³⁹	0.31 (0.08-11)	1.3 (1.1-1.6)
Bradykinesia		
Difficulty rising from chair		
Duarte et al ³⁷	1.9 (1.3-2.7)	0.58 (0.38-0.90)
Mutch et al ³⁸	5.2 (2.9-9.5)	0.39 (0.25-0.63)
Posture and Motor Tasks		
Loss of balance		
Duarte et al ³⁷	1.6 (1.3-2.2)	0.29 (0.13-0.68)
Mutch et al ³⁸	6.6 (3.4-13)	0.35 (0.21-0.57)
Shuffling gait		
Duarte et al ³⁷	3.3 (2.1-5.0)	0.32 (0.18-0.58)
Mutch et al ³⁸	15 (4.7-47)	0.50 (0.36-0.71)
Trouble turning in bed ³⁸	13 (4.1-43)	0.56 (0.41-0.76)
Trouble buttoning ³⁷	3.0 (2.0-4.4)	0.33 (0.19-0.60)
Trouble opening jars ³⁸	6.1 (3.4-11)	0.26 (0.14-0.48)
Uncontrolled limbs ³⁹	1.3 (0.53-3.1)	0.93 (0.72-1.2)
Fine Motor		
Micrographia		
Duarte et al ³⁷	2.8 (1.8-4.2)	0.44 (0.27-0.71)
Mutch et al ³⁸	5.9 (3.1-9.4)	0.30 (0.17-0.53)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

Table 38-3 Signs Evaluated in Patients With Possible Parkinson Disease

Sign	LR+ (95% CI)	LR- (95% CI)
Tremor		
Tremor ³⁴	1.5 (1.0-2.3)	0.47 (0.27-0.84)
Tremor with rigidity and bradykinesia ³⁴	2.2 (1.2-4.2)	0.50 (0.34-0.75)
Rigidity		
Rigidity ³⁹	2.8 (1.8-4.4)	0.38 (0.19-0.76)
Rigidity with bradykinesia ³⁹	4.5 (2.9-7.1)	0.12 (0.03-0.45)
General Findings		
Glabella tap ³⁶	4.5 (2.8-7.4)	0.13 (0.03-0.47)
Voice softer ³⁷	3.7 (2.4-5.6)	0.25 (0.13-0.49)
Change in speech ³⁹	2.6 (1.2-5.3)	0.73 (0.53-1.0)
Asymmetric disease ³⁴	1.8 (0.98-3.2)	0.61 (0.41-0.91)
Levodopa response ³⁴	1.2 (0.87-1.6)	0.63 (0.31-1.3)
Bradykinesia		
Akinetic/rigid disease ³⁴	0.44 (0.25-0.75)	1.7 (1.1-2.6)
Posture and Motor Tasks		
Difficulty or inability to walk heel to toe ³⁹	2.9 (1.9-4.5)	0.32 (0.15-0.70)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

Selection bias was a major problem in all 4 articles because many of the patients evaluated had either been diagnosed as having PD on initial clinical examination or had obvious parkinsonian features. In one study,³⁹ a screening instrument was administered to patients with an initial diagnosis of PD, peripheral neuropathy, stroke, or epilepsy. The instrument included both a self-administered questionnaire and a set of physical tasks, performance of which was graded subjectively. Neurologists confirming the presence of PD were aware of each patient's initial diagnosis and responses to the screening instrument. This obviously makes the study prone to observer bias. Like most studies with low methodologic quality, these 4 articles³⁶⁻³⁹ reported optimistic LRs.

Precision

Interclinician and intraclinician reliability of symptoms and signs was documented only for the glabella tap sign.³⁶ Precision could not be quantified in the clinicopathologic studies because symptom histories were obtained retrospectively from medical records.

Interclinician reliability in eliciting the glabella tap sign was found to be 88% among patients with intracranial disease and 100% in controls.³⁶ A κ coefficient for interclinician agreement could not be calculated because data about how each clinician scored each patient were not included. No causes for imprecision in assessing symptoms or signs were documented in the selected articles.

Accuracy

Several symptoms, collected by patient self-report in a questionnaire,³⁸ significantly increase the likelihood of PD when present and decrease it when absent. The symptoms are trouble turning in bed, shuffling while walking, micrographia, difficulty rising from a chair, loss of balance, and trouble opening jars. The diagnostic value of tremor as a symptom varied widely among the selected articles, with a range in LRs+ of 1.3 to 11 (Table 38-2).

The lack of tremor as a symptom makes PD less likely (range of LRs-, 0.24-0.60). However, the usefulness of the lack of tremor as a symptom is limited by verification bias in the corresponding studies. Verification bias occurs when confirmatory or criterion standards are selectively applied to patients, depending on the results of their preliminary screening test.⁴⁰ The independent value of tremor detected on neurological examination has an LR+ of only 1.5 (95% CI, 1.0-2.3), while the absence of a tremor detected on examination about halves the likelihood of PD (LR-, 0.47; 95% CI, 0.27-0.84) (Table 38-3).³⁴

Rigidity as a symptom has an LR+ range of 1.3 to 4.5 and makes PD more likely. The absence of rigidity has a broad LR, making it less useful (LR- range, 0.12-0.73) (Table 38-2). As a sign detected on neurological examination (Table 38-3), rigidity as an independent value is more useful (LR+, 2.8; 95% CI, 1.8-4.4; LR-, 0.38; 95% CI, 0.19-0.76).³⁹ When both rigidity and bradykinesia are present, the LR+ for the combination of findings improves to 4.5 (95% CI, 2.9-7.1), while the absence of both findings makes PD much less likely, with an LR- of 0.12 (95% CI, 0.03-0.45).³⁴

The glabella tap sign is useful, with an LR+ of 4.5 (95% CI, 2.8-7.4) and an LR- of 0.13 (95% CI, 0.03-0.47).³⁶ Changes in voice have an LR range of 2.6 to 3.7, while the lack of a voice change makes PD somewhat less likely (LR- range, 0.25-0.73).^{37,39} The results confirm the limited usefulness of the response to levodopa because the LR+ is only 1.2 (95% CI, 0.87-1.6), while the absence of a response has an LR- of 0.63 (95% CI, 0.31-1.3).³⁴

Physicians sometimes must consider whether patients have PD or multiple systems atrophy (MSA), so a series of findings have been compared between the 2 disorders.³⁵ Patients with rigidity as an initial presenting feature of PD are less likely to have PD (LR, 0.53; 95% CI, 0.35-0.80) and more likely to have MSA. The presence of dementia also favors PD over MSA (LR+, 3.2; 95% CI, 1.5-6.8). Not surprisingly, central or autonomic nervous systems findings are much less likely with PD (LR range when central or autonomic findings present, 0.03-0.31), so their presence favors MSA. Bradykinesia and symptoms of depression do not help distinguish between the disease (95% CI for the LR includes 1).

THE BOTTOM LINE

Few studies address the clinical diagnosis of PD rigorously. Nearly 200 years after it was first described, the accurate clin-

ical diagnosis of PD remains a significant challenge. There is a great need for diagnostic studies involving larger numbers of patients in which presenting symptoms and signs are prospectively compared with the final diagnosis, established through a suitable criterion standard, such as autopsy or serial neurologic evaluation.

A number of classic features of PD, when present, do help establish the diagnosis. These include the symptoms of tremor, the combination of rigidity and bradykinesia, loss of balance, micrographia, and shuffling gait. Difficulty with the tasks of turning in bed, opening jars, and rising from a chair should also raise the suspicion of PD. It is difficult to gauge the usefulness of the absence of tremor as a symptom in ruling out PD because of verification bias in the studies in which it was evaluated.

The diagnostic value of the classic combination of tremor, rigidity, and bradykinesia on examination is modest at best. Useful signs include the glabella tap, difficulty walking heel to toe, and the presence of rigidity on examination.

CLINICAL SCENARIO—RESOLUTION

This patient presents with many common features of PD. You can question him about the tasks of turning in bed and opening jars. A sample of his writing may reveal micrographia. A glabella tap test should be performed. Additional positive symptoms or signs would justify empiric treatment with dopaminergic medication, with careful follow-up by a physician experienced in the treatment of this condition.

Author Affiliations at the Time of the Original Publication

Department of Family Medicine, University of Pittsburgh, Pittsburgh, Pennsylvania (Drs Rao, Fisch, and Srinivasan); University of Pittsburgh Medical Center, Pittsburgh, Pennsylvania (Drs D'Amico, Okada, and Eaton); and Kaiser Permanente of Colorado, Aurora (Dr Robbins).

Acknowledgments

We thank Richard Bedlack, MD, Mitchell Heflin, MD, and Amy Rosenthal, MD, for their helpful reviews of the manuscript; Amy Haugh, MLS, for her valuable library assistance; and David L. Simel, MD, MHS, for his valuable guidance.

REFERENCES

- Barbosa ER, Limongi JC, Cummings JL. Parkinson's disease. *Psychiatr Clin North Am*. 1997;20(4):769-790.
- Ben-Shlomo Y. The epidemiology of Parkinson's disease. *Baillieres Clin Neurol*. 1997;6(1):55-68.
- Oppenheimer DR. Diseases of the basal ganglia, cerebellum and motor neurons. In: Adams JH, Corsellis JAN, Duchen IW, eds. *Greenfield's Neuropathology*. 4th ed. New York, NY: John Wiley & Sons Inc; 1984:699-747.
- Jankovic J, Rajput AH, McDermott MP, Perl D. The evolution of diagnosis in early Parkinson disease. *Arch Neurol*. 2000;57(3):369-372.
- Clarke CE, Davies P. Systematic review of acute levodopa and apomorphine challenge tests in the diagnosis of idiopathic Parkinson's disease. *J Neurol Neurosurg Psychiatry*. 2000;69(5):590-594.
- Rajput AH, Rozdilsky B, Rajput A. Accuracy of clinical diagnosis in parkinsonism: a prospective study. *Can J Neurol Sci*. 1991;18(3):275-278.
- Hughes AJ, Daniel SE, Lees AJ. Improved accuracy of clinical diagnosis of Lewy body Parkinson's disease. *Neurology*. 2001;57(8):1497-1499.
- Hughes AJ, Daniel SE, Ben-Shlomo Y, Lees AJ. The accuracy of diagnosis of parkinsonian syndromes in a specialist movement disorder service. *Brain*. 2002;125(pt 4):861-870.
- Olanow CW, Watts RL, Koller WC. An algorithm (decision tree) for the management of Parkinson's disease. *Neurology*. 2001;56(suppl 5): S1-S88.
- Koller W. Levodopa in the treatment of Parkinson's disease. *Neurology*. 2000;55(11 suppl 4):S2-S7.
- Lee KS, Merriam A, Owen A, Chew B, Tan TC. The medical, social and functional profile of Parkinson's disease patients. *Singapore Med J*. 1994;35(3):265-268.
- Haberman B. Spousal perspectives of Parkinson's disease in middle life. *J Adv Nurs*. 2000;31(6):1409-1415.
- Siderowf A. Parkinson's disease. *Neurol Clin*. 2001;19(3):565-578.
- Playfer JR. Parkinson's disease. *Postgrad Med J*. 1997;73(859):257-264.
- Bower JH, Dickson DW, Taylor L, Maraganore DM, Rocca WA. Clinical correlates of the pathology underlying parkinsonism: a population perspective. *Mov Disord*. 2002;17(5):910-916.
- Ng DC. Parkinson's disease diagnosis and treatment. *West J Med*. 1996;165(4):234-240.
- Kumar S, Bhatia M, Behari M. Sleep disorders in Parkinson's disease. *Mov Disord*. 2002;17(4):775-781.
- Calne DB, Snow BJ, Lee C. Criteria for diagnosing Parkinson's disease. *Ann Neurol*. 1992;32(suppl):S125-S127.
- Fahn S. Differential diagnosis of tremors. *Med Clin North Am*. 1972;56(6):1363-1375.
- Martinelli P, Gabellini S, Gulli R, Lugaresi E. Different clinical features of essential tremor: a 200-patient study. *Acta Neurol Scand*. 1987;75(2): 106-111.
- Lou JS, Jankovic J. Essential tremor: clinical correlates in 350 patients. *Neurology*. 1991;41(2 pt 1):234-238.
- Colcher A, Simuni T. Clinical manifestations of Parkinson's disease. *Med Clin North Am*. 1999;83(2):327-347.
- Goetz CG, Pappert EJ. *Textbook of Clinical Neurology*. Philadelphia, PA: WB Saunders Co; 1999:226.
- Deuschl G, Bain P, Brin M; Ad Hoc Scientific Committee. Consensus statement of the Movement Disorder Society on tremor. *Mov Disord*. 1998;13(suppl 3):2-23.
- Quinn N. Parkinsonism: recognition and differential diagnosis. *BMJ*. 1995;310(6997):447-452.
- Findley LJ. Classification of tremors. *J Clin Neurophysiol*. 1996;13(2): 122-132.
- Jobbagy A, Furnée EH, Harcos P, Tarczy M, Krekule I, Komjathi L. *Analysis of Movement Patterns Aids the Early Detection of Parkinson's Disease: Proceedings of the 19th International Conference, Chicago, Ill, October 30 to November 2, 1997*. Washington, DC: Institute of Electrical and Electronics Engineers; 1997:1760-1763.
- Haerer A. *DeJong's The Neurologic Examination*. 5th ed. New York, NY: JB Lippincott Co; 1992:185.
- Alder CH. Differential diagnosis of Parkinson's disease. *Med Clin North Am*. 1999;83(2):349-367.
- Stoessl AJ, Rivest J. Differential diagnosis of parkinsonism. *Can J Neurol Sci*. 1999;26(suppl 2):S1-S4.
- Holleman DR Jr, Simel DL. Does the clinical examination predict airflow limitation? *JAMA*. 1995;273(4):313-319.
- Greenland S, Robins J. Estimation of a common effect parameter from sparse follow-up data. *Biometrics*. 1985;41(1):55-68.
- DerSimonian R, Laird N. Meta-analysis in clinical trials. *Control Clin Trials*. 1986;7(3):177-188.
- Hughes AJ, Ben-Shlomo Y, Daniel SE, Lees AJ. What features improve the accuracy of clinical diagnosis in Parkinson's disease? a clinicopathologic study. *Neurology*. 1992;42(6):1142-1146.
- Wenning GK, Ben-Shlomo Y, Hughes A, Daniel SE, Lees A, Quinn NP. What clinical features are most useful to distinguish definite multiple

- system atrophy from Parkinson's disease? *J Neurol Neurosurg Psychiatry*. 2000;68(4):434-440.
36. Pearce J, Aziz H, Gallagher JC. Primitive reflex activity in primary and symptomatic parkinsonism. *J Neurol Neurosurg Psychiatry*. 1968;31(5): 501-508.
37. Duarte J, Claveria LE, De Pedro-Cuesta J, Sempere AP, Coria F, Calne DB. Screening Parkinson's disease. *Mov Disord*. 1995;10(5):643-649.
38. Mutch WJ, Smith WC, Scott RF. A screening and alerting questionnaire for parkinsonism. *Neuroepidemiology*. 1991;10(3):150-156.
39. Meneghini F, Rocca WA, Anderson DW, et al. Validating screening instruments for neuroepidemiologic surveys. *J Clin Epidemiol*. 1992;45(4): 319-331.
40. Mower WR. Evaluating bias and variability in diagnostic test reports. *Ann Emerg Med*. 1999;33(1):85-91.

Prepared by Goutham Rao, MD
Reviewed by Richard Bedlack, MD

CLINICAL SCENARIO

A 70-year-old man presents with a 3-month history of progressive tremor and difficulty in writing letters. His wife states that he has trouble eating with a spoon. Physical examination reveals tremor in both hands. Otherwise, he feels well and continues to play golf.

UPDATED SUMMARY ON PARKINSON DISEASE

Original Review

Rao G, Fisch L, Srinivasan S, et al. Does this patient have Parkinson disease? *JAMA*. 2003;289(3):347-353.

UPDATED LITERATURE SEARCH

Our literature search used the parent search strategy for The Rational Clinical Examination series, combined with the subject headings “exp tremor,” “exp Parkinson disease,” and “essential tremor,” published in English from 2002 to July 2004. No new original articles address the sensitivity and specificity of clinical findings for Parkinson disease (PD). We found 1 new systematic review that addresses key questions related to the diagnosis and management of PD. Much of the recent research in the diagnosis of PD has focused on functional neuroimaging with positron emission tomography and single-photon-emission computed tomography.

CHANGES IN THE REFERENCE STANDARD

There are no laboratory or radiographic reference standards for PD. The best reference standard is serial clinical assessments performed by a specialist. Newer imaging methods may eventually be routinely helpful in identifying PD at an early stage or in patients with atypical signs and symptoms and in following the progression of disease. However, they have not yet been evaluated with enough methodologic rigor to allow clinicians to be confident in the results.

NEW FINDINGS

One systematic review confirms that PD remains a clinical diagnosis and that neuroimaging, the levodopa challenge, and other diagnostic tests are not useful.¹

IMPROVEMENTS IN THE DATA PRESENTED IN THE ORIGINAL PUBLICATION

Tables 38-2 and 38-3 were revised because they originally included data for helping to distinguish PD from multiple systems atrophy. These tables now show only the likelihood ratios for making the diagnosis of PD.

RESULTS OF LITERATURE REVIEW

No data suggest new useful findings for PD. Symptoms common in PD include tremor, rigidity, bradykinesia, and micrographia (unusually small handwriting). Difficulty with tasks such as turning in bed, opening jars, or rising from a chair are also common. The most useful clinical findings are the glabellar tap (see Figure 38-1) and heel-to-toe tests.

EVIDENCE FROM GUIDELINES

No guidelines explicitly address the diagnosis of PD.

CLINICAL SCENARIO—RESOLUTION

The patient should be asked about key symptoms of PD, including loss of balance (positive likelihood ratio [LR+], 1.6-6.6), shuffling gait (LR+, 3.3-15), and rigidity or stiffness (LR+, 1.3-4.5). It is not clear why he has trouble eating with a spoon. This could be a result of his tremor or because of slowness of movements (bradykinesia). The patient should be asked about difficulty rising from a chair (LR+, 1.9-5.2), a specific manifestation of bradykinesia. His difficulty in writing letters could also be attributed to tremor, bradykinesia, or micrographia.

Further evaluation reveals that the patient's tremor is of slow frequency (4-6/s) and occurs at rest, which is typical of the tremor of PD. By contrast, classic essential tremor is usually postural (occurs while maintaining a position against gravity) or kinetic (occurs during voluntary movements.) Physiologic tremor is usually postural and of higher frequency (8-12/s) than the tremor of PD.

The glabella test should be performed (LR+, 4.5), and whether the patient is able to walk heel to toe (LR+, 2.9)

should also be assessed. Asking the patient to rise from a chair and checking to see whether his upper limbs are rigid during passive movement is also useful in establishing the diagnosis (combination of bradykinesia and rigidity; LR+, 4.5).

Parkinson disease is a progressive disease, and patients present with different combinations of clinical features of various severity at different stages. It is therefore, important to evaluate the patient periodically before establishing the diagnosis and initiating treatment.

PARKINSON DISEASE—MAKE THE DIAGNOSIS

Parkinson disease remains a clinical diagnosis. No accurate laboratory or radiologic test is currently available. An acute challenge with levodopa, followed by monitoring for improvement of symptoms, is not useful diagnostically.

PRIOR PROBABILITY

Parkinson disease affects 1% of people older than 65 years and 2% of those older than 85 years.² The prevalence among older patients presenting with general neurologic complaints is almost certainly much higher, but precise figures are unavailable.

POPULATION FOR WHOM PARKINSON DISEASE SHOULD BE CONSIDERED

Adults with a tremor or other symptoms noted in Tables 38-4 and 38-5 (see Figure 38-2 for assessment of bradykinesia).

DETECTING THE LIKELIHOOD OF PARKINSON DISEASE

Table 38-4 Useful Symptoms for Detecting Parkinson Disease

Symptom (No. of Studies)	LR+ ^a	LR- ^a
Shuffling gait (2)	3.3-15	0.32-0.50
Bradykinesia (difficulty rising from a chair) (2)	1.9-5.2	0.39-0.58
Loss of balance (2)	1.6-6.6	0.29-0.35
Tremor (4)	1.4-11	0.24-0.60
Rigidity (3)	1.3-4.5	0.12-0.93

Abbreviations: LR+, positive likelihood ratio; LR-, negative likelihood ratio.

^aValues represent the range across the studies. Different definitions of symptoms precluded meta-analysis.

Table 38-5 Useful Signs for Detecting Parkinson Disease

Sign ^a	LR+ (95% CI)	LR- (95% CI)
Rigidity and bradykinesia	4.5 (2.9-7.1)	0.12 (0.03-0.45)
Glabella tap	4.5 (2.8-7.4)	0.13 (0.03-0.47)
Difficulty walking heel to toe	2.9 (1.9-4.5)	0.32 (0.15-0.70)
Rigidity	2.8 (1.8-4.4)	0.38 (0.19-0.76)
Asymmetry of disease	1.8 (0.98-3.2)	0.61 (0.41-0.91)
Tremor	1.5 (1.0-2.3)	0.47 (0.27-0.84)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

^aAll likelihood ratios are from single studies.

REFERENCE STANDARD TESTS

Serial clinical examinations performed by a specialist.

REFERENCES FOR THE UPDATE

- Levine CB, Fahrback KR, Siderowf AD, Estok RP, Ludensky VM, Ross SD. *Evidence Report/Technology Assessment Number 57*. Rockville, MD: Agency for Healthcare Research and Quality; 2003. AHRQ Publication No. 03-E040. Prepared by Metaworks, Inc, under Contract No. 290-97-0016.^a

- Ben-Schlomo Y. The epidemiology of Parkinson's disease. *Ballieres Clin Neurol*. 1997;6(1):55-68.

^aFor Evidence to Support the Update for this topic, see <http://www.JAMAevidence.com>.

Parkinsonism

TITLE Diagnosis and Treatment of Parkinson's Disease: A Systematic Review of the Literature.

AUTHORS Levine CB, Fahrbach KR, Siderowf AD, Estok RP, Ludensky VM, Ross SD.

CITATION Agency for Healthcare Research and Quality. *Evidence Report/Technology Assessment No. 57*. Rockville, MD: Agency for Healthcare Research and Quality; 2003. Prepared by Metaworks, Inc., under Contract No. 290-97-0016. AHRQ Publication No. 03-E040. <http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=hstat1.chapter.123680>. Accessed June 7, 2008.

QUESTIONS Though they did assess a number of diagnostic techniques, the authors began by addressing 2 specific questions related to the diagnosis of Parkinson disease (PD):

1. What are the results of neuroimaging studies (computed tomography [CT], magnetic resonance imaging [MRI], positron emission tomography [PET], single photon emission computed tomography [SPECT]) or other diagnostic tests in determining the diagnosis of PD?
2. What are the accuracy, sensitivity, and specificity of an L-dopa challenge diagnosing PD?

DESIGN Qualitative systematic review (ie, formal systematic review without meta-analysis) of studies about diagnosis. Meta-analytic techniques were used for pharmacologic studies about treatment.

DATA SOURCES MEDLINE, Current Contents, and Cochrane Library databases and bibliographies of all publications included for review and recent review articles.

STUDY SELECTION AND ASSESSMENT To identify studies about diagnosis, the sources above were searched for articles published between January 1, 1990, and December 31, 2000. Inclusion criteria for diagnostic studies included (1) adult patients with potential diagnosis of PD; (2) addressing of any diagnostic test to establish or support a diagnosis of PD; and (3) observational (prospective, retrospective, and cross-sectional) or intervention designs (randomized controlled trials, nonrandomized controlled trials, and uncontrolled case series).

The review excluded letters, case reports, editorials, commentaries, unpublished study reports, animal or in vitro studies, studies written in languages other than English, studies published before 1990, studies with fewer than 10 patients, crossover studies, studies in which results for PD population could not be separated from results from other populations, or studies not pertaining to diagnosis or treatment of PD. All eligible studies were rated for both quality and level of evidence.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

Studies included for review included the following categories of diagnostic testing: apomorphine or L-dopa challenge tests, autopsy studies, clinical or laboratory tests, color vision testing, MRI, olfactory testing, PD Test Battery (tests of motor function, olfaction and depression quantified with a score between 0 and 1.0), PET scans, SPECT scans, and other scans. The authors found that clinical diagnosis is the reference diagnostic standard in many studies, but they point out that this is problematic because the clinical diagnosis may be wrong in up to 25% of cases. Long-term response to L-dopa was also used but is not a valid reference standard.

MAIN OUTCOME MEASURES

Sensitivity, specificity, positive and negative predictive value when available.

MAIN RESULTS

There is insufficient evidence to determine the diagnostic accuracy and therefore to recommend the use of the following to diagnose PD: apomorphine and L-dopa challenge tests; SPECT, PET, or other scans; olfactory tests; color vision tests; or blood and cerebrospinal fluid tests.

Sensitivities for the PD Test Battery ranged from 69% to 95%. Specificities ranged from 64% to 95%. Methodologic problems with studies of the PD Test Battery, however, limit its usefulness. The authors do not recommend the battery for diagnosing PD.

The sensitivity and specificity of clinical diagnosis of PD at the first patient visit ranged from 53% to 90% and 74% to 94%, respectively (compared with subsequent neuropathology studies at autopsy). At the last visit, these values increased to a sensitivity of 60% to 87% and specificity of 82% to 97%. The positive predictive value increased from 34% to 61% at the first visit to 43% to 75% at the last visit. The median negative predictive value was more than 95% at both visits. The authors conclude that the clinical diagnosis of PD is modestly accurate and improves over time and that autopsy is the only acceptable reference standard.

CONCLUSIONS

LEVEL OF EVIDENCE Systematic review.

STRENGTHS This methodologically sound systematic review addresses specific diagnostic questions, establishes explicit search and inclusion criteria for diagnostic studies, and uses widely accepted methods to assess the value of those studies.

LIMITATIONS The last articles included for review were published in 2000.

The review substantiates the role of the clinical examination for establishing the diagnosis, rather than additional tests. The diagnostic value of specific clinical features (eg, bradykinesia) was not explicitly evaluated. The role of repeated examinations is important for physicians because the diagnostic accuracy improves with following the patient's serial symptoms and signs.

Reviewed by Goutham Rao, MD

CHAPTER 39

Is This Patient Allergic to Penicillin?

Alan R. Salkind, MD

Paul G. Cuddy, PharmD

John W. Foxworth, PharmD

CLINICAL SCENARIOS

CASE 1 An 18-year-old male college student presents with group A streptococcal pharyngitis, and you prescribe penicillin.¹ The patient informs you that he developed a rash after taking about half a penicillin prescription for a respiratory tract infection 3 years ago. The rash was bright red, was restricted to the extremities and trunk, and resolved several days after penicillin was discontinued.

CASE 2 A 26-year-old pregnant woman has syphilis. She recalls an “itchy rash” and trouble breathing after taking penicillin 4 years ago; she thinks the rash appeared about 3 days into the course of penicillin. Penicillin is the recommended antibiotic for syphilis in pregnancy, even for patients with a true penicillin allergy.²

WHY IS IT IMPORTANT TO DETERMINE WHETHER PATIENTS HAVE TRUE PENICILLIN ALLERGY?

Penicillin, a β -lactam antibiotic, and its semisynthetic chemical derivatives (such as ampicillin and amoxicillin) and other β -lactam antibiotics (including cephalosporins, carbapenems, and monobactams) remain first-line or acceptable alternative treatments for many infections.³ However, the use of drugs containing penicillin is often limited by an unconfirmed or questionable history of penicillin hypersensitivity provided by the patient. Because fear of penicillin anaphylaxis is common among clinicians encountering a patient with a self-reported history of penicillin allergy, many clinicians overdiagnose penicillin allergy in patients who have not had a true allergic reaction to penicillin. Some clinicians may simply accept a diagnosis of penicillin allergy from a patient without obtaining a detailed history of the reaction.^{4,5} Some patients, when asked, have no firsthand recall of an allergic response to penicillin, the patient perhaps having been informed of their allergy by a parent.^{4,5} For example, patients reporting a penicillin allergy have described an “allergic reaction” consisting of fever and yellow spots on the tonsils, which actually related to the illness they were being treated for, rather than penicillin itself.⁴ Unless a detailed medical history and a critical evaluation of the reaction are sought, such patients may incorrectly be labeled as penicillin allergic. In fact, 80% to 90% of patients who report a penicillin allergy are not truly allergic to the drug when assessed by skin testing.⁶⁻⁹ Consequently, penicillin is withheld from many patients who could safely receive the drug or its derivatives, perhaps affecting outcomes.¹⁰ Two studies have shown that incorrectly labeling patients as being allergic to penicillin was associated with increased health care costs.^{11,12}

METHODS

We searched MEDLINE for English-language literature dated from 1966 to October 2000 by using the following Medical Subject Headings and search strategy: (1) “medical history taking” or “physical examination” and “penicillin” or “ β -lactam hypersensitivity” and (2) “reproducibility of results” or “observer variation” and “penicillin” or “ β -lactam hypersensitivity.” A textword search was also performed with “interobserver,” “intraobserver,” “accuracy,” “precision,” “reliability,” “sensitivity,” “specificity,” “skin testing,” and “penicillin” or “ β -lactam hypersensitivity” or “allergy.” The bibliographies of pertinent articles were searched to identify additional references. Included articles were original studies conducted on ambulatory or hospitalized children or adults describing the accuracy or precision of skin testing in the diagnosis of an immunoglobulin E (IgE) –mediated penicillin allergy. Excluded studies investigated allergy to aminopenicillins (amoxicillin and ampicillin) or cephalosporins, did not use both major and minor determinants in the skin testing procedure, or did not provide an explicit definition of penicillin allergy or of a positive skin test result. Data from patients who were reported to have had an uninterpretable or equivocal skin test result were not included in our analysis. Quality measures were applied, as used in The Rational Clinical Examination series (see Table 1-7 for a summary of Evidence Grades and Levels).¹³ Using study quality as a measure of the relative weight that a single study should receive was not used in our analysis, because other authors have highlighted the pitfalls of this practice.^{14,15} Of the 14 studies¹⁶⁻²⁹ meeting our inclusion criteria, 4 studies¹⁶⁻¹⁹ compared the clinical history with the skin test result for penicillin allergy among a group of patients with and without a positive history of penicillin allergy (Table 39-1). Confidence intervals (CIs) for the likelihood ratios (LRs) from individual studies were computed with a previously described method.³⁰

Classification of Penicillin Hypersensitivity Reactions

The frequency of all adverse reactions to penicillin in the general population ranges from 0.7% to 10%.³¹ This wide variation in the frequency of adverse reactions to penicillin exists because of

a number of variables, including exposure history, route of administration, duration of treatment, elapsed time between the reaction and diagnostic skin testing or reexposure, and nature of the initial reaction. Understanding the different classifications of penicillin hypersensitivity reactions aids evaluation of each patient’s risk for an allergic reaction that would preclude administration of a drug that contains penicillin.

Gell and Coombs³² categorized allergic reactions to penicillins by the type of reaction, immune mechanism, and clinical syndrome, whereas Levine³³ classified untoward reactions to penicillin by their time of onset (Table 39-2). Classification of penicillin allergy has been reviewed by several authors^{6,34,35} and is summarized briefly below. We refer the reader to the original works for a more detailed discussion.^{32,33}

Immediate Reactions

Type I, or immediate, reactions are often associated with the systemic manifestations of anaphylaxis, such as diffuse erythema, pruritus, urticaria, angioedema, bronchospasm, laryngeal edema, hyperperistalsis, hypotension, or cardiac arrhythmias, either alone or in combination (Table 39-2). Anaphylactic reactions occur in about 0.004% to 0.015% of penicillin courses and are most commonly observed in adults between the ages of 20 and 49 years.³¹ A history of atopy does not generally place an individual at increased risk for a type I penicillin reaction.³⁶ However, atopic patients may have a higher frequency of severe anaphylactic reactions.³⁶

Type I reactions result when penicillin or its reactive metabolites covalently bind to serum proteins and then crosslink with preformed penicillin-specific IgE antibodies bound to tissue mast cells, circulating basophils, or both. When the bound IgE antibodies are crosslinked by allergen, mast cells are activated to release their mediators. A patient using β -adrenergic antagonists may be at increased risk of death if anaphylaxis occurs.³⁷

Some reactions to penicillin occurring from 1 to 72 hours after administration may also be IgE mediated. These reactions, termed *accelerated reactions*, can be manifested by urticaria, angioedema, laryngeal edema, and wheezing. However, urticaria and angioedema can occur at any time after adminis-

Table 39-1 Studies Assessing the Skin Test for Penicillin Allergy Among Patients With and Without a History of Penicillin Allergy^a

Source, y	Quality of Methods ^b	Setting (Sample Size, % Penicillin Allergic)	Sensitivity	Specificity	LR+ (95% CI)	LR- (95% CI)
Adkinson et al, ¹⁶ 1971	C	Inpatient, nonconsecutive (n = 218, 11.9)	0.61	0.74	2.4 (1.6-3.5)	0.5 (0.3-0.85)
Green et al, ¹⁷ 1977	C	Multicenter study (n = 2947, 8.1)	0.79	0.45	1.4 (1.4-1.5)	0.5 (0.39-0.57)
Sogn et al, ¹⁸ 1992	C	Multicenter study, chronically ill (n = 1298, 12.6)	0.85	0.50	1.7 (1.6-1.9)	0.3 (0.21-0.44)
Gadde et al, ¹⁹ 1993	C	Sexually transmitted disease clinic (n = 5063, 2.5)	0.43	0.85	2.9 (2.4-3.7)	0.7 (0.57-0.77)
Summary					1.9 (1.5-2.5)	0.5 (0.4-0.6)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

^aAn LR+ indicates the likelihood that a patient with a history of penicillin allergy will have a positive penicillin skin test result; an LR- indicates the likelihood that a patient without a history of penicillin allergy will have a positive penicillin skin test result.

^bQuality of methods was based on published criteria. Grade C: independent, blind comparison of sign or symptom, with a gold standard of diagnosis among nonconsecutive patients suspected of having the target condition plus, perhaps, individuals without the target condition; or nonindependent comparison of sign or symptom with a standard of uncertain validity.¹³ Of the included studies, not all patients received penicillin challenge. (See Table 1-7 for a summary of Evidence Grades and Levels.)

Table 39-2 Classification of Penicillin Reactions

Classification	Time of Onset, h	Mediator(s)	Clinical Signs	Skin Testing Useful	Comments
Immediate (type I reaction)	<1	Penicillin-specific IgE antibodies	Anaphylaxis or hypotension, laryngeal edema, wheezing, angioedema, urticaria	Yes	Much more likely with parenteral administration than oral administration; fatal outcome in 1 per 50000 to 1 per 100000 treatment courses; some reactions occurring between 1 and 72 h of exposure may be IgE mediated (see text for details)
Late reactions	>72 h after exposure				
Type II		IgG, complement	Increased clearance of red blood cells, platelets by lymphoreticular system	No	IgE not involved
Type III		IgG, IgM immune complexes	Serum sickness, tissue injury	No	Tissue lodging of immune complexes; drug fever
Type IV			Contact dermatitis	No	
Other (idiopathic)	Usually > 72 h after exposure		Maculopapular or morbilliform rashes	No	1% To 4% of all patients receiving penicillin

Abbreviations: IgE, immunoglobulin E; IgG, immunoglobulin G; IgM, immunoglobulin M.

tration of penicillin. Life-threatening reactions occurring beyond 1 hour of penicillin administration are rare. The patient described in case 1 had none of the features of a serious IgE-mediated penicillin allergy. In contrast, the patient described in case 2 had features that suggest an IgE-mediated accelerated reaction.

Late Reactions

Late penicillin hypersensitivity reactions are those that occur after 72 hours of drug administration. These responses have been classified as types II, III, or IV, depending on the immune mechanism underlying the response (Table 39-2). Because none of these reactions are IgE dependent, skin testing has no role in the evaluation of a patient with type II, III, IV, or idiopathic responses to penicillin.

Some reactions to penicillin are not included in the Gell and Coombs³² classification and have been termed *idiopathic*. Although various immune-mediated responses have been postulated, the exact immunologic mechanisms underlying these responses are not known. The most common idiopathic reaction to drugs containing penicillin is a maculopapular or morbilliform rash. The combined frequency of all rashes occurring in patients taking penicillin is estimated at 1% to 4%.^{38,39} These eruptions are usually symmetric, often confluent erythematous macules and papules that generally spare the palm and soles. They may originate on the extremities of ambulatory patients or overlie pressure areas of bedridden patients.⁹ Rashes associated with ampicillin administration occur in 5.2% to 9.5% of treatment courses.³⁸⁻⁴⁰ Patients with Epstein-Barr virus or cytomegalovirus infections, or with acute or chronic lymphocytic leukemia, are reported to have a higher incidence of ampicillin-associated rash.⁶ The reason for the increased incidence of rash caused by ampicillin remains unknown.

In experimental settings, individuals with histories of type I hypersensitivity reactions to aminopenicillins (ampicillin, amoxicillin, bacampicillin) demonstrate cross-reactivity to

penicillin when assessed by skin testing.⁴¹ Although some of these individuals fail to react to penicillin skin testing and react only to skin testing with aminopenicillins, these occurrences appear less commonly, yet are well documented.^{42,43} In contrast, individuals reporting a history of a nonimmediate reaction are less likely to react to penicillin skin test determinants.⁴² In light of the above, it is prudent to perform a skin test for penicillin in those individuals with a history of an urticarial reaction to aminopenicillin derivatives and administer a drug containing penicillin only in patients with negative skin test results.⁴⁴ Patients without urticarial rashes to aminopenicillins are unlikely to manifest a serious reaction and can generally receive a drug containing penicillin, without further testing.⁴⁴

Drug-independent rashes are common in patients with viral infections, especially those caused by the human immunodeficiency virus, hepatitis B, mumps, echovirus,¹¹ and Coxsackie virus.⁴⁵ Infections with numerous bacteria can also be associated with a rash.⁴⁵ Therefore, patients with some infections who develop a rash while taking penicillin derivatives or penicillin itself should not be automatically labeled as penicillin allergic. Moreover, many patients taking penicillin may also be taking other medications, including other antibiotics, that can cause rashes that are independent of β-lactam compounds.⁹ Maculopapular eruptions caused by drugs containing penicillin may subside spontaneously despite continued use of the drug and may not recur on reexposure.^{9,40} The frequency of a penicillin-associated maculopapular eruption on reexposure to the drug is not known because many clinicians withhold drugs that contain penicillin in this patient population. Green et al¹⁷ reported that 3 (3.5%) of 85 patients with a maculopapular rash associated with penicillin administration had adverse reactions to oral challenge with penicillin. The nature of the oral challenge reaction was not specified, but none were classified as type I reactions. Six (4.5%) of 134 patients with negative penicillin skin test results and a history of a penicillin-associated cutaneous reaction had an adverse response to penicillin readministration.

The nature of the response was not described.¹⁹ Another 3 patients with negative penicillin skin test results and a history of rash caused by penicillin developed a type I reaction to penicillin administration,¹⁹ likely indicating the inaccuracy of the historical information. If a detailed history of a patient's reaction to penicillin indicates that the rash was strictly maculopapular, with no signs of a type I reaction, then it appears to be safe to readminister an antibiotic that contains penicillin.^{20,35}

Penicillin (or any medication) that is clearly associated with the development of exfoliative dermatitis or the Stevens-Johnson syndrome should be discontinued immediately and not readministered to the patient.⁹ Patients with a history of Stevens-Johnson syndrome or exfoliative dermatitis attributable to β -lactam drugs should not undergo a skin test⁹ and should wear a Medic Alert bracelet indicating a severe reaction to the drug.

Cross-Reactivity With Other β -Lactam Antibiotics

Cephalosporins (like penicillins) contain a β -lactam ring.³ The frequency of allergic reactions within 24 hours of cephalosporin administration to patients with a history of penicillin allergy and positive skin test results was 5.6% vs 1.7% for patients with a history of penicillin allergy and negative skin test results.³⁵ Earlier reports suggested that the cross-reaction rate may be higher for first-generation cephalosporins than for subsequent cephalosporins.⁴⁶ Complicating interpretation of these data was the finding that some early first-generation cephalosporins contained trace amounts of penicillin.⁴⁶

One group of investigators challenged 19 patients with well-documented histories of a type I allergy to penicillin with cephalosporins containing side-chain structures expected to lead to cross-reaction.⁴⁷ Seventeen patients tolerated the challenge doses and subsequent courses of the cephalosporin. Both of the patients who had allergic reactions had positive penicillin skin test results to benzylpenicillin only; however, another patient with the same skin test pattern tolerated cephalosporin challenge without incident. Because this study did not contain a control group without penicillin allergy, the relative significance of the penicillin allergy cannot be determined.⁴⁷ In another study, 1 (1.6%) of 62 patients with positive skin test results to penicillin who were challenged with a cephalosporin on the same day as the skin testing developed mild urticaria plus bronchospasm within 24 hours.⁷ Solley et al²² described 27 patients with positive penicillin skin test results, all of whom were treated with cephalosporins without a reaction, whereas 2 (1.5%) of 151 patients with a positive history of penicillin allergy and negative penicillin skin test results had an allergic reaction to cephalosporins. Forty-three treatment courses with cephalosporins were administered to children who had positive skin test results or positive oral challenge to penicillin. Forty-one (95%) of the cephalosporin courses were well tolerated. Two children experienced a mild IgE type-mediated reaction.²⁶

In summary, neither the history nor the penicillin skin test result reliably predicts the probability of allergic reactions to cephalosporins in patients with positive histories of penicillin allergy. Available data suggest that the majority of patients who

Box 39-1 Taking a History of Penicillin Allergy: What to Ask

- What was the patient's age at the time of the reaction?
- Does the patient recall the reaction? If not, who informed the patient of it?
- How long after beginning penicillin did the reaction begin?
- What were the characteristics of the reaction?
- What was the route of administration?
- Why was the patient taking penicillin?
- What other medications was the patient taking? Why and when were they prescribed?
- What happened when the penicillin was discontinued?
- Has the patient taken antibiotics similar to penicillin (eg, amoxicillin, ampicillin, cephalosporins) before or after the reaction? If yes, what was the result?

are allergic to penicillin tolerate cephalosporins without significant reaction. Our approach to a patient with a history of penicillin allergy requiring a cephalosporin is to first determine the likelihood that the patient requiring a cephalosporin had a type I allergic reaction to penicillin (Box 39-1). If a detailed medical history does not suggest a true penicillin allergy, we administer the cephalosporin. When the history is concerning for penicillin allergy, we recommend penicillin skin testing. For patients with negative skin test results, the cephalosporin can be administered. When the penicillin skin test result is positive and an alternate drug cannot be used, cephalosporin desensitization by an experienced practitioner should be considered.⁴⁴

Some investigators have called for broader use of cephalosporin skin testing in patients who are allergic to penicillin and require a cephalosporin.^{26,47} However, protocols for skin testing with cephalosporin compounds are not well standardized, and the negative predictive value of cephalosporin skin testing is not known.^{7,44,46}

Carbapenems and monobactams are β -lactam antibiotics, of which imipenem and aztreonam are respective prototypes. Patients who have positive skin test results to penicillin have also shown a high degree of reactivity to imipenem determinants.⁷ Therefore, carbapenems should not be administered to patients with positive penicillin skin test results or a concerning history of a type I allergic response to penicillin.⁷ Available information indicates that aztreonam may be safely administered to most, if not all, patients with a type I allergic response to penicillin.⁷

RESULTS

WHY IS TAKING A DETAILED CLINICAL HISTORY FOR PENICILLIN ALLERGY IMPORTANT?

The majority of patients with a history of penicillin allergy have no concurrent physical examination findings related to

the adverse response to penicillin. Thus, initial determination of the probability of a true penicillin allergy relies almost solely on a detailed medical history (Box 39-1). For example, a patient receiving penicillin who developed a rash on day 5 of treatment for an upper respiratory tract infection who has since taken multiple courses of drugs containing penicillin without an untoward reaction does not have a true penicillin allergy. In contrast, if a patient described new-onset wheezing 1 hour after a penicillin injection, it is highly probable that this patient had an immediate hypersensitivity reaction to the drug.

When assessing a patient for penicillin allergy, all medications that the patient is (or was) taking should be evaluated for their propensity to cause a reaction similar to the one being attributed to penicillin. For example, a patient receiving penicillin for 4 days without untoward effects who then begins taking an angiotensin-converting enzyme inhibitor and develops angioedema on the third day of administration (day 7 of penicillin therapy) should not be automatically labeled as penicillin allergic.⁹

Serious allergic and fatal reactions to antibiotics that contain penicillin can occur in individuals who have never had a previous allergic reaction to penicillin or who deny any medical exposure to drugs that contain penicillin.⁶ The clinical history, no matter how carefully considered, cannot prevent these rare reactions.

ACCURACY OF THE CLINICAL HISTORY FOR PENICILLIN ALLERGY

Four studies¹⁶⁻¹⁹ compared the clinical history of penicillin allergy to the skin test result and included patients who had positive histories of penicillin allergy and those who did not. We pooled the results of these studies (Table 39-1). The presence of a clinical history suggesting penicillin allergy increases the likelihood that the patient will be allergic to penicillin as assessed by skin testing (summary positive LR, 1.9; 95% CI, 1.5-2.5). The absence of a clinical history suggesting penicillin allergy decreases the likelihood of a positive skin test result by slightly more than half (summary negative LR, 0.5; 95% CI, 0.4-0.6).

The percentages of positive skin test results for patients with a history of anaphylaxis, urticaria, or a maculopapular rash ranged from 17% to 46%, 12% to 16%, and 4% to 7%, respectively, in 2 studies.^{17,19} One study¹⁷ also reported that 18% of patients with a history of angioedema had a positive penicillin skin test result. Limited data are available about the rate of skin test reactivity when the patient's allergic status to penicillin is unknown. Sogn et al¹⁸ found that the proportion of positive skin test results among patients with an unknown history of penicillin allergy was 3% (3/96). In another study of 57 patients with an uncertain allergy to penicillin, 1.7% had a positive skin test reaction.¹⁹ Although the clinical history does help separate those more likely from those less likely to have a penicillin allergy, as demonstrated by skin testing, the history is not precise. The studies¹⁶⁻¹⁹ evaluating the skin test in patients with and without a history of penicillin

allergy had higher positive predictive values for the clinical history than all but 1 of the studies that included only patients with positive histories of penicillin allergy (summary positive predictive value, 19%; 95% CI, 18%-21%). After exclusion of the outlier study,²¹ the positive predictive value for the clinical history of penicillin allergy is 14% (95% CI, 12%-18%). Thus, a clinician would need to perform skin tests on 7 patients with a history suggesting penicillin allergy to find 1 positive reaction.

PENICILLIN SKIN TESTING

Blackley introduced the skin test in 1865 when he scarified a portion of his forearm, sprinkled it with pollen, and noted the development of itching and swelling surrounded by erythema. It is now known that IgE antibodies mediate such reactions.⁴⁸

The penicillin skin test has no place in the treatment of patients without a clinical history of a type I penicillin allergy. It would also be unnecessary in the face of a bona fide history of a life-threatening type I reaction, when equally efficacious antibiotics are available, or if the clinician would still withhold penicillin therapy regardless of skin test results. Some,^{11,20,26} but not all,^{6,7} investigators have suggested elective skin testing for penicillin allergy. Elective skin testing for penicillin allergy may be useful in children because of the frequent outpatient need for antibiotics that contain penicillin. In addition, elective skin testing of adults with positive histories of penicillin allergy might be considered in certain situations. An example of this would be a cancer patient who has a positive history of penicillin allergy who is likely to develop chemotherapy-induced neutropenia and requires a drug containing penicillin promptly for an infection.⁴⁴ Recommendations regarding the general use of elective penicillin skin testing await further study.

However, when the history of type I hypersensitivity is concerning and penicillin therapy is warranted, skin testing is helpful and should be considered. For example, a patient who has a positive history of penicillin allergy and has *Staphylococcus aureus* endocarditis susceptible to an antistaphylococcal penicillin (such as nafcillin or oxacillin) would be an appropriate candidate for skin testing⁴⁹ because vancomycin, an antibiotic often used in patients who are allergic to penicillin and have serious *S aureus* infections, is less effective and more expensive than nafcillin.⁵⁰

Another factor influencing the decision to perform a skin test relates to the ability to do the test in an efficient manner with appropriate reagents and interpretation. A recent study of hospitalized patients showed that the time for skin testing averaged 40 minutes, and the cost for the skin test reagents and equipment was \$17 per patient.¹²

The positive predictive value of skin testing to assess risk for an allergic reaction to penicillin is unclear because patients providing a convincing history of a type I reaction to penicillin who subsequently react to skin testing are unlikely to undergo oral penicillin challenge. However, a limited number of patients with positive skin test results

have been treated with penicillin. The risk of a type I allergic reaction ranges from about 9% in subjects with negative histories of penicillin allergy to 50% to 70% in subjects with positive histories.⁶ Despite the observation that some patients with positive skin test results are able to tolerate penicillin, it is inadvisable to administer penicillin to these patients because of an unfavorable risk-benefit ratio. Patients with positive skin test results who need penicillin should undergo desensitization.⁶

Many studies have used penicillin challenge in subjects with positive histories of penicillin allergy and negative skin test results, and the experiences have been consistent: the majority of subjects tolerated the challenge, and those who did not experience only urticaria or other mild cutaneous reaction. When 6739 patients with positive histories of penicillin allergy and negative skin test results were given penicillin, only 101 (1.5%) developed an IgE-mediated reaction, whereas 43 (0.63%) developed a delayed reaction.¹⁶⁻²⁹ Penicillin anaphylaxis was not reported in subjects with negative skin test results who received a penicillin challenge. Patients with positive histories of penicillin allergy who have negative skin test results may receive a medically supervised oral penicillin challenge. If there is no reaction to the oral challenge, patients can then generally be treated with an oral or parenteral penicillin. When the skin test is properly performed, almost all patients with negative penicillin skin test results can safely receive the drug. Thus, even when the history of a type I reaction is concerning and penicillin is the clear drug of choice, skin testing should be considered because the majority of those patients will have a negative skin test result, and 98% of patients with a negative result will tolerate penicillin without any serious sequelae.^{6,7}

If skin testing seems appropriate after obtaining a detailed history of the patient's reaction to penicillin, both the major determinant (benzyl penicilloyl; commercially available as PrePen; Kremers-Urban, Milwaukee, Wisconsin) and the minor determinant composed of freshly diluted aqueous penicillin G should be used.⁴⁴ A minor determinant mixture (MDM) is not commercially available in the United States. The use of the major determinant reagent alone would detect between 75% and 90% of all potential positive reactions. Including fresh penicillin G as the sole MDM reagent improves identification of patients who may have reactions to the skin test by 5% to 10%.⁶ However, the addition of other minor determinants to the testing protocol may increase identification of patients allergic to penicillin by skin testing to about 99%.^{16,23} The absence of a commercially available MDM solution has hampered the general use of the penicillin skin test. The steps for performing a penicillin skin test are described in detail elsewhere.^{44,51}

Limitations of Skin Testing Compared With Other Diagnostic Techniques

A review identified the essential criteria that any diagnostic test must satisfy, but studies evaluating penicillin skin testing fail to meet several of these criteria.⁵² An independent, blind comparison of a reference standard—oral penicillin

challenge—has never been uniformly applied to all patients who have undergone skin testing. Moreover, few studies have actually subjected all subjects with positive histories of penicillin allergy and negative skin test results to oral challenge. It is clear that in most studies the skin test results influenced the decision to perform the penicillin challenge, thus introducing a built-in bias. These limitations undermine attempts to generate reliable estimates of sensitivity and specificity for penicillin skin testing compared with oral penicillin challenge used as the gold standard. This problem, labeled *reverse workup bias*, can result in biased test estimates because it is likely that patients who do not undergo skin testing differ in important ways from patients for whom testing is undertaken.⁵³

Redelmeier and Sox⁵³ used expert opinion to estimate the probability of severe allergic reactions in 100 patients with a convincing penicillin allergy history who were to receive the drug without previous skin testing. Respondents estimated that 5 to 90 (median, 50) patients would experience a severe reaction to penicillin.⁵³ Accordingly, these authors concluded that skin testing for patients with a “very strong” history of penicillin allergy is not recommended (ie, estimated 50% pretest probability for a severe allergic reaction to penicillin based on a history of penicillin allergy). They reasoned that clinicians would be unwilling to risk a potential serious reaction in these patients even if they had negative skin test results.⁵³ However, at least 50% of patients with a history of an IgE-mediated reaction will have a negative skin test result.^{17,19} Because patients with negative skin test results tolerate penicillin well, patients with histories of a type I reaction should undergo skin testing when they have a strong indication for penicillin therapy. At least 50% of these patients will be identified as candidates for penicillin therapy. Still, if the clinician's treatment threshold is so high that he or she is unwilling to administer penicillin regardless of the clinical situation (given a history of a type I reaction), skin testing clearly has no value.

CLINICAL SCENARIOS—RESOLUTIONS

In case 1, the patient reported a maculopapular rash halfway through a course of penicillin. The pretest probability that this represents a true reaction to penicillin would be 10%, using a conservative estimate for the frequency of any adverse reaction to penicillin.³¹ After a careful medical history is taken from the patient, one might conclude that his experience is inconsistent with a type I reaction. Using a negative LR of 0.5 for a negative history of penicillin allergy, the probability that this patient will experience any adverse reaction to penicillin can be revised to 5.2%, a percentage that is similar to the frequency of any adverse reaction to penicillin in the general population.³¹ In this patient, skin testing should not be performed, and the patient should receive penicillin. Careful history-taking should have increased confidence about the safety of administering penicillin to this patient.

The patient described in case 2 reported, and a detailed history confirmed, an urticarial rash within 72 hours of taking penicillin. Again, using 10% as the pretest probability of any adverse reaction to penicillin,³¹ a 17% posttest probability that this patient has a true penicillin allergy is arrived at by using the positive LR of 1.9. We would perform skin testing on this patient because a negative skin test result virtually excludes a significant reaction to penicillin, whereas a positive skin test result in this patient with a strong indication for penicillin would mandate desensitization.⁶

COMMENT

We identified only 4 studies meeting our inclusion criteria that used penicillin skin testing in patients with and without positive histories of penicillin allergy (Table 39-1). Two of these studies provided no data on the frequency of positive skin test results in patients according to their previous reaction to penicillin.^{16,17} Moreover, none of the studies included in our analysis were independent, blind comparisons of signs or symptoms of penicillin allergy compared with the gold standard, oral penicillin challenge. These methodologic flaws have tempered the quality of the published database for this common clinical problem, leaving us with a pervasive lack of guidelines for determining penicillin allergy.

Nonetheless, encountering patients with a stated penicillin allergy remains an everyday problem for many clinicians, and some clinicians simply prescribe an alternate antibiotic for these patients. However, some alternative antibiotics are more expensive, less effective, or associated with more adverse effects than penicillin, and there is the risk of increasing antimicrobial resistance. Other clinicians turn to the literature, hoping to find a rich evidence-based database to help guide their decision-making process. Regrettably, the methods of diagnosing true penicillin allergy have been inadequately studied, leaving the busy clinician to make the most informed decision possible while recognizing the limitations in the available data.

We provide an approach to the patient with a stated penicillin allergy based on a critical analysis of an admittedly limited database: by systematically documenting signs and symptoms associated with the patient's adverse reaction to penicillin (Box 39-1), the clinician should be able to determine with a higher degree of certainty whether the patient has a true penicillin allergy. Using a more structured approach should allow the clinician to assess the likelihood that the patient had a true penicillin allergy, thereby allowing a more rational decision-making process in consideration of penicillin usage, as illustrated by the resolution of the clinical scenarios.

THE BOTTOM LINE

- Many patients recalling a reaction to penicillin are unsure of specific details and, even when evidence supporting true penicillin allergy is absent, are nevertheless labeled as penicillin allergic by many clinicians.

- A detailed history of the patient's drug reaction can help the clinician determine whether or not the patient's self-reported history is compatible with a true penicillin allergy, permitting penicillin administration to those patients who are unlikely to have true penicillin allergy.
- Eighty percent to 90% of all patients reporting a penicillin allergy have negative penicillin allergy reaction when assessed by skin testing, meaning that penicillin is withheld from many patients who could safely receive the drug.
- Patients who develop a rash while taking penicillins should not be automatically labeled as penicillin allergic without considering other possibilities, such as a rash caused by the infection being treated or by other drugs the patient is taking.
- For patients with a concerning history of penicillin allergy who have a compelling need for penicillin, skin testing should be performed.
- At least 98% of patients with positive histories of penicillin allergy and negative skin test results can tolerate penicillin without any sequelae.

Author Affiliations at the Time of the Original Publication

Department of Medicine and Sections of Infectious Diseases (Dr Salkind), Clinical Pharmacology (Dr Cuddy), and the Department of Medicine and Clinical Pharmacology (Dr Foxworth), The University of Missouri, Kansas City School of Medicine, Kansas City, Missouri.

Acknowledgments

This work was supported by a Sonnenwirth Grant from the University of Missouri, Kansas City School of Medicine.

We appreciate the expert advice offered by Peter Bressler, MD, Gerald Smetana, MD, Vance Fowler, MD, and Russell Hall, MD, during the preparation of this article.

REFERENCES

- Bisno AL, Gerber MA, Gwaltney JM Jr, Kaplan EL, Schwartz RH. Diagnosis and management of group A streptococcal pharyngitis: a practice guideline. *Clin Infect Dis*. 1997;25:574-583.
- 1998 Guidelines for treatment of sexually transmitted diseases. *MMWR Morb Mortal Wkly Rep*. 1997;47(No. RR-1):40-41.
- Wright AJ. The penicillins. *Mayo Clin Proc*. 1999;74(3):290-307.
- Oswald NTA. Penicillin allergy: a suspect label. *BMJ (Clin Res Ed)*. 1983;287(6387):265-266.
- Surtees SJ, Stockton MG, Gietzen TW. Allergy to penicillin: fable or fact? *BMJ*. 1991;302(6784):1051-1052.
- Weiss ME, Adkinson NF. Immediate hypersensitivity reactions to penicillin and related antibiotics. *Clin Allergy*. 1988;18(6):515-540.
- Saxon A, Beall GN, Rohr AS, Adelman AC. Immediate hypersensitivity reactions to beta-lactam antibiotics. *Ann Intern Med*. 1987;107(2):204-214.
- Lin RY. A perspective on penicillin allergy. *Arch Intern Med*. 1992;152(5):930-937.
- deShazo RD, Kemp SF. Allergic reactions to drugs and biologic agents. *JAMA*. 1997;278(22):1895-1906.
- Kirk CR, Dorgan JC, Hart CA. Gas gangrene: a cautionary tale. *BMJ (Clin Res Ed)*. 1988;296(6631):1236-1237.
- Macy E. Elective penicillin skin testing and amoxicillin challenge: effect on outpatient antibiotic use, cost, and clinical outcomes. *J Allergy Clin Immunol*. 1998;102(2):281-285.
- Harris AD, Sauberman L, Kabbash L, Greineder DK, Samore MH. Penicillin skin testing: a way to optimize antibiotic utilization. *Am J Med*. 1999;107(2):166-168.

13. Holleman DR Jr, Simel DL. Does the clinical examination predict airflow limitation? *JAMA*. 1995;273(4):313-319.
14. Dickersin K, Berlin JA. Meta-analysis: state-of-the-science. *Epidemiol Rev*. 1992;14:154-176.
15. Detsky AS, Naylor CD, O'Rourke K, McGeer AJ, L'Abbe KA. Incorporating variations in the quality of individual randomized trials into meta-analysis. *J Clin Epidemiol*. 1992;45(3):255-265.
16. Adkinson NF Jr, Thompson WL, Maddrey WC, Lichtenstein LM. Routine use of penicillin skin testing on an inpatient service. *N Engl J Med*. 1971;285(1):22-24.
17. Green GR, Rosenblum AH, Sweet LC. Evaluation of penicillin hypersensitivity: value of clinical history and skin testing with penicilloyl-polysine and penicillin G. *J Allergy Clin Immunol*. 1977;60(6):339-345.
18. Sogn DD, Evans RI, Shepherd GM, et al. Results of the National Institute of Allergy and Infectious Diseases Collaborative Clinical Trial to test the predictive value of skin testing with major and minor penicillin derivatives in hospitalized adults. *Arch Intern Med*. 1992;152(5):1025-1032.
19. Gadde J, Spence M, Wheeler B, Adkinson NF Jr. Clinical experience with penicillin skin testing in a large inner-city STD clinic. *JAMA*. 1993;270(20):2456-2463.
20. Mendelson LM, Ressler C, Rosen JP, Selcow JE. Routine elective penicillin allergy skin testing in children and adolescents: study of sensitization. *J Allergy Clin Immunol*. 1984;73(1 pt 1):76-81.
21. Sullivan TJ, Wedner HJ, Shatz GS, Yecies LD, Parker CW. Skin testing to detect penicillin allergy. *J Allergy Clin Immunol*. 1981;68(3):171-180.
22. Solley GO, Gleich GJ, Van Dellen RG. Penicillin allergy: clinical experience with a battery of skin-test reagents. *J Allergy Clin Immunol*. 1982;69(2):238-244.
23. Warrington RJ, Simons FER, Ho HW, Gorski BA, Tse KS. Diagnosis of penicillin allergy by skin testing: the Manitoba experience. *CMAJ*. 1978;118(7):787-791.
24. Levine BB, Zolov DM. Prediction of penicillin allergy by immunological tests. *J Allergy*. 1969;43(4):231-244.
25. Bierman CW, Vanarsdale PP Jr. Penicillin allergy in children: the role of immunological tests in its diagnosis. *J Allergy*. 1969;43(5):267-272.
26. Pichichero ME, Pichichero DM. Diagnosis of penicillin, amoxicillin, and cephalosporin allergy: reliability of examination assessed by skin testing and oral challenge. *J Pediatr*. 1998;132(1):137-143.
27. Macy E, Richter PK, Falkoff R, Zeiger R. Skin testing with penicilloate and penilloate prepared by an improved method: amoxicillin oral challenge in patients with negative skin test responses to penicillin reagents. *J Allergy Clin Immunol*. 1997;100(5):586-591.
28. Chandra RK, Joglekar SA, Tomas E. Penicillin allergy: anti-penicillin IgE antibodies and immediate hypersensitivity skin reactions employing major and minor determinants of penicillin. *Arch Dis Child*. 1980;55(11):857-860.
29. Noritake DT, Sue MA, Klaustermeyer WB. Immediate hypersensitivity skin testing for penicillin allergy. *Mil Med*. 1986;151(7):395-399.
30. Simel DL, Samsa GP, Matchar DB. Likelihood ratios with confidence: sample size estimation for diagnostic test studies. *J Clin Epidemiol*. 1991;44(8):763-770.
31. Idsoe O, Guthe T, Willcox RR, De Weck A. Nature and extent of penicillin side-reactions, with particular reference to fatalities from anaphylactic shock. *Bull World Health Organ*. 1968;38(2):159-188.
32. Gell PGH, Coombs RRA. Classification of allergic reactions responsible for clinical hypersensitivity and disease. In: Gell PGH, Coombs RRA, Hachmann PJ, eds. *Clinical Aspects of Immunology*. Oxford, England: Blackwell Scientific Publications; 1975:761-781.
33. Levine BB. Immunologic mechanisms of penicillin allergy: a haptenic model system for the study of allergic diseases in man. *N Engl J Med*. 1966;275(20):1115-1125.
34. Saxon A. Immediate hypersensitivity reactions to β -lactam antibiotics. *Rev Infect Dis*. 1983;5(suppl 2):S368-S379.
35. Shepherd GM. Allergy to β -lactam antibiotics. *Immunol Allergy Clin North Am*. 1991;11:611-633.
36. Adkinson NF Jr. Risk factors for drug allergy. *J Allergy Clin Immunol*. 1984;74:567-572.
37. Lang DM, Alpern MB, Visintainer PF, Smith ST. Increased risk for anaphylactoid reaction from contrast media in patients on beta-adrenergic blockers or with asthma. *Ann Intern Med*. 1991;115(4):270-276.
38. Shapiro S, Slone D, Siskind V, Lewis G, Hershel J. Drug rash with ampicillin and other penicillins. *Lancet*. 1969;2(7628):969-972.
39. Arndt KA, Jick H. Rates of cutaneous reactions to drugs: a report from the Boston Collaborative Drug Surveillance Program. *JAMA*. 1976;235(9):918-923.
40. Kerns DL, Shira JE, Go S, Summers RJ, Schwab JA, Plunket DC. Ampicillin rash in children: relationship to penicillin allergy and infectious mononucleosis. *Am J Dis Child*. 1973;125(2):187-190.
41. Romano A, Torres J, Fernandez J, et al. Allergic reactions to ampicillin: studies on the specificity and selectivity in subjects with immediate reactions. *Clin Exp Allergy*. 1997;27(12):1425-1431.
42. Romano A, Quaratino D, Di Fonzo M, Papa G, Venuti A, Gasbarrini G. A diagnostic protocol for evaluating nonimmediate reactions to ampicillins. *J Allergy Clin Immunol*. 1999;103(6):1186-1190.
43. Vega JM, Blanca M, García JJ, et al. Immediate allergic reactions to amoxicillin. *Allergy*. 1994;49(5):317-322.
44. Executive summary of disease management of drug hypersensitivity: a practice parameter: Joint Task Force on Practice Parameters, the American Academy of Allergy, Asthma and Immunology, the American Academy of Allergy, Asthma and Immunology, and the Joint Council of Allergy, Asthma and Immunology. *Ann Allergy Asthma Immunol*. 1999;83(6 pt 3):665-700.
45. Levin S, Goodman LJ. An approach to acute fever and rash (AFR) in the adult. In: Remington JS, Swartz MN, eds. *Current Clinical Topics in Infectious Diseases*. Cambridge, MA: Blackwell Scientific; 1995:19-75.
46. Anne S, Reisman RE. Risk of administering cephalosporin antibiotics to patients with histories of penicillin allergy. *Ann Allergy Asthma Immunol*. 1995;74(2):167-170.
47. Blanca M, Fernandez J, Miranda A, et al. Cross-reactivity between penicillins and cephalosporins: clinical and immunological studies. *J Allergy Clin Immunol*. 1989;83:381-385.
48. Bousquet J, Michel F-B. In vivo methods for study of allergy: skin tests, techniques, and interpretation. In: Middleton E Jr, Reed CE, Ellis EF, Adkinson NF Jr, Yunginger JW, Busse WW, eds. *Allergy: Principles and Practice*. St Louis, MO: Mosby; 1993:573-594.
49. Dodek P, Phillips P. Questionable history of immediate-type hypersensitivity to penicillin in staphylococcal endocarditis: treatment based on skin test results versus empirical alternative treatment—a decision analysis. *Clin Infect Dis*. 1999;29(5):1251-1256.
50. Karchmer AW. *Staphylococcus aureus* and vancomycin: the sequel. *Ann Intern Med*. 1991;115(9):739-741.
51. Norman PS, Peebles RS. In vivo diagnostic allergy testing methods. In: Rose NR, de Macario EC, Folds JD, Lane HC, Nakamura RM, eds. *Manual of Clinical Laboratory Immunology*. Washington, DC: American Society of Microbiology; 1997:875-880.
52. Jaeschke R, Guyatt G, Sackett DL. Users' guide to the medical literature, III: how to use an article about a diagnostic test, A: are the results of the study valid? *JAMA*. 1994;271:389-391.
53. Redelmeier DA, Sox HC Jr. The role of skin testing for penicillin allergy. *Arch Intern Med*. 1990;150(9):1939-1945.

Prepared by David L. Simel, MD, MHS
Reviewed by John Foxworth, PharmD, and Alan Salkind, MD

CLINICAL SCENARIO

A 12-year-old boy with pharyngitis has a positive rapid streptococcal test result for strep throat. You would like to treat with penicillin V. He has no personal or family history of a penicillin reaction. In the past, he has received amoxicillin without an adverse reaction. Does the absence of a previous reaction guarantee that he is not allergic?

UPDATED SUMMARY ON PENICILLIN ALLERGY

Original Review

Salkind AR, Cuddy PG, Foxworth JW, Simel DL, Rennie D, eds. Is this patient allergic to penicillin? *JAMA*. 2001;285(19):2498-2505.

UPDATED LITERATURE SEARCH

Our literature search combined the search terms “penicillins,” “lactams,” “drug hypersensitivity,” and “skin tests” with “sensitivity and specificity,” “reproducibility of results,” “medical history taking,” and “physical examination.” The search was limited to English-language publications that were in the MEDLINE database from 2000 to July 2004. The search strategy yielded 84 articles that were further limited to 52 articles by confirming that they were cross-referenced with “exp epidemiologic methods” or “prospective studies” as search terms. We found 4 studies that provided the predictive value of a history of penicillin allergy for a subsequent positive penicillin skin test result or response to penicillin on rechallenge. One of these studies was done on a selected population of patients who had experienced an immediate penicillin response.

NEW FINDINGS

- A history of a reaction to penicillin increases the likelihood of a reaction to future doses of penicillin (likelihood ratio [LR], 11; 95% confidence interval [CI], 8.5-14), but most patients with a previous reaction will not be allergic and will not have future reactions.

- Patients with a history of penicillin reaction who have a negative skin test result may still react to courses of penicillin (about 10%), but the chance of a life-threatening reaction is greatly diminished.
- Patients with well-documented *immediate* reaction to penicillin are about 90% likely to react with subsequent courses.

Details of the Update

Two types of studies address the role of the patient history in predicting penicillin allergy. The question is important because only about 15% of patients claiming a penicillin allergy prove to have positive skin test results. Prospective studies either assess responses to skin testing for penicillin allergy or assess responses to subsequent penicillin administrations after an initial penicillin course. About 10% of adult patients with an allergic-like event attributed to penicillin prove to have positive skin test results for penicillin allergy.¹ A negative skin test reaction suggests that 90% of such patients who receive another course of penicillin, according to a medical record review, will use penicillin safely, with no adverse outcome. Of those who do experience an allergic response, the likelihood of an anaphylactic reaction is low (<0.7%).

A study of a large practice research database extracted patients who had 2 separate administrations of penicillin.² The authors looked for coded diagnoses for specific allergic reactions within 30 days after a penicillin course. The LR for predicting a second allergic reaction after the first allergic reaction was 11 (95% CI, 8.5-14). The study has at least 2 important weaknesses. First, although a large percentage of patients with an initial reaction subsequently received a second course of penicillin (48%), the clinicians may have used judgment to exclude the patients most likely to experience a second reaction. With the rates of reactions in the patients receiving a second course of penicillin, the LR for a positive history of allergy is likely greater than 16. Second, neither the individual patients nor the individual patient records were reviewed. The study relies on the accuracy of the physician’s code for each visit.

A study of children referred for skin testing with possible penicillin allergy addresses the problem of lack of confirmation from retrospective studies.³ Individuals who had a negative skin test result received an oral dose of penicillin observed by the physician (none had an adverse reaction). The overall true rate of allergy was 6.7%. Using the commonly held notion that 10% of all patients experience a penicillin reaction, we can

deduce that the LR for a positive history of reaction in these children was 10.6, which is almost identical to the point estimate in the large database study.

A large study reported the 10-year results of patients ($n = 330$) referred for allergy testing in a highly selected population of patients with a much higher prevalence of true allergy (88%). This population was unique in that all the patients had experienced immediate reactions to penicillin. For these patients, the predictive value of the history was much greater in that 61% had a positive skin test result, 11% had a negative skin test result but a positive radioimmunoassay test result, and 27% ($n = 89$) had a negative skin test result. When the patients with a negative skin test result were rechallenged with penicillin in a controlled setting, 49 (55%) had reactions within 1 hour of administration.

IMPROVEMENTS IN THE DATA PRESENTED IN THE ORIGINAL PUBLICATION

No changes in the data presented are required.

CHANGES IN THE REFERENCE STANDARD

The true reference standard for penicillin allergy is a reaction to an oral penicillin challenge that is observed and well documented by a physician. Because a physician may not observe most reactions, there may be uncertainty about attributing the reaction to penicillin. The response to penicillin skin testing can be used as better clinical evidence for penicillin allergy.

Table 39-3 Predictive Value of a Patient's History of Penicillin Allergy for Either a Positive Skin Test Result or Actual Allergic Reaction on Rechallenge

Finding (No. of Studies)	Predictive Value (95% CI) to Identify Patients With a Positive Skin Test Result, %	Predictive Value (95% CI) for No Allergic Response to Penicillin Administration After a Negative Skin Test Result, %
History of penicillin allergy (2) ^{a,3}	14 (12-16)	
History of penicillin allergy with negative skin test result in children ^{3a}		>95
History of penicillin allergy with negative skin test result in adults ^{1b}		90 (86-91)
	LR+	LR- (95% CI)
History of penicillin allergy for predicting an allergic response (anaphylaxis, positive skin test result, or response to a second course of penicillin) ²	>11	0.98 (0.98-0.99)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

^aAll children were given an oral challenge with penicillin after negative skin test results. Zero of 69 children had an allergic response. The value is the lower limit of the 95% CI.

^bAccording to chart review of second penicillin courses after a negative skin test result.

RESULTS OF LITERATURE REVIEW

Univariate Findings for Penicillin Allergy

Skin test results help identify patients who, despite a history of penicillin allergy, have a low probability of a reaction to penicillin when rechallenged (Table 39-3).

EVIDENCE FROM GUIDELINES

The role of routinely taking a history of penicillin allergy for the general population is not addressed by any federal recommendations. However, the Centers for Disease Control and Prevention addresses penicillin allergies in its treatment guidelines for sexually transmitted diseases, primarily because of the central role that penicillin plays in the treatment of neurosyphilis, congenital syphilis, or syphilis in pregnancy.⁴ The guidelines note that 3% to 10% of US adults have experienced reactions after penicillin and that 10% of those remain penicillin allergic. They recommend skin testing with the major and minor penicillin determinants for patients who are at higher risk of a subsequent reaction. When a patient has a negative skin test result, the Centers for Disease Control and Prevention recommends penicillin for the treatment of the above syphilitic conditions, although some experts recommend desensitization for these patients. The recommendations include a protocol for penicillin allergy skin testing.

CLINICAL SCENARIO—RESOLUTION

The lack of a previous reaction does not guarantee that the child will have no future reaction to penicillin or a penicillin derivative, but the risk is low (about 1%). You can confidently prescribe penicillin as the preferred antibiotic. Had the child experienced a previous reaction, the clinician would have to make a decision. The risk that such a child is truly allergic to penicillin is greater, given a previous reaction, even though the absolute risk is low. The decision for how to proceed may depend on your assessment of the previous reaction, the need for penicillin vs alternative antibiotics, and the availability of penicillin testing. If the previous reaction was well documented or an immediate reaction, erythromycin is an inexpensive treatment for streptococcal pharyngitis so that urgent skin testing is not necessary.

REFERENCES FOR THE UPDATE

- Macy E, Mangat R, Burchette R. Penicillin skin testing in advance of need: multiyear follow-up in 568 test result-negative subjects exposed to oral penicillins. *J Allergy Clin Immunol*. 2003;111(5):1111-1115.^a
- Apter AA, Kinman JL, Bilker WB, et al. Represcription of penicillin after allergic-like events. *J Allergy Clin Immunol*. 2004;113(4):764-770.
- Langley JM, Halperin SA, Bortolussi R. History of penicillin allergy and referral for skin testing: evaluation of a pediatric penicillin allergy testing program. *Clin Infect Dis*. 2002;25(5):181-184.^a
- Centers for Disease Control and Prevention. Sexually transmitted disease treatment guidelines 2002. *Morb Mortal Wkly Rep*. 2002;51(No. RR-6):28-29.
- Macy E. How predictive is a history of penicillin allergy [letter]? *JAMA*. 2001;286(10):1174.^a

^aFor the Evidence to Support the Update for this topic, see <http://www.JAMAevidence.com>.

PENICILLIN ALLERGY—MAKE THE DIAGNOSIS

PRIOR PROBABILITY

About 10% of patients will have an adverse reaction to penicillin, but most will not be penicillin allergic. Less than 1% of all patients will have a true allergy to penicillin, as defined by an anaphylactic reaction, positive skin test result, or response to a second dose of penicillin. Physicians should ascertain the nature of the reaction to help decide whether it might have represented a true penicillin allergy. An allergy to penicillin may be difficult to ascertain from the patient's medical history, primarily because many penicillin reactions do not represent allergic reactions. The most important finding from a penicillin history is also the least frequent—patients with severe reactions (eg, toxic epidermal necrolysis, life-threatening anaphylaxis, hemolytic anemia, liver damage) should not be skin tested for penicillin allergy and should not receive penicillin.⁵

POPULATION FOR WHOM PENICILLIN ALLERGY SHOULD BE CONSIDERED

- Patients with a previous allergic response to penicillin-type antibiotics
- Patients with a response to medications that cross-react with penicillin (eg, cephalosporins, carbapenems)

DETECTING THE LIKELIHOOD OF PENICILLIN ALLERGY

The history of a penicillin allergy or a negative skin test result affects the probability of a true penicillin allergy (Table 39-4).

Table 39-4 Likelihood Ratios and Predictive Value for a History of Penicillin Allergy in Predicting True Allergy

Finding (No. of Studies)	LR+ (95% CI)	LR- (95% CI)
History of penicillin allergy for predicting an allergic response (anaphylaxis, positive skin test, or response to a second course of penicillin)	>11	≈1
History of penicillin allergy for predicting response to penicillin skin test result (4)	1.9 (1.5-2.5)	0.5 (0.4-0.6)
Predictive Value		
Predictive value of a negative skin test result in a patient with a history of penicillin allergy (1)	>95% ^a	

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

^aProbability of no allergic-like event with second administration of penicillin.

REFERENCE STANDARD TESTS

Penicillin allergy is confirmed by a reliable history of an immediate anaphylactic reaction, positive skin test reactivity, or well-documented response to a second observed penicillin challenge.

This page intentionally left blank

EVIDENCE TO SUPPORT THE UPDATE:

39

Penicillin Allergy

TITLE Represcription of Penicillin After Allergic-like Events.

AUTHORS Apter AA, Kinman JL, Bilker WB, et al.

CITATION *J Allergy Clin Immunol.* 2004;113(4):764-770.

QUESTION How well does the history of an allergic-like event from penicillin predict subsequent responses after readministration?

DESIGN Analysis of a large database.

SETTING AND PATIENTS United Kingdom General Practice Research Database of 687 general practitioner practices, representative of England and Wales, and comprising 6% of the population. The database contained records for 3375162 patients who received at least 1 prescription of penicillin.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

The General Practice Research Database was assessed for patients who received at least 2 penicillin doses at least 60 days apart. The patients were sorted by those who had an allergic-like response to the first administration and then to subsequent administrations. Allergic reactions were identified by computerized codes within 30 days of the penicillin prescription for anaphylaxis, urticaria, angioedema, erythema multiforme, laryngeal spasm, dermatitis attributed to a drug, toxic epidermal necrolysis, or adverse drug reactions attributed to a medication.

MAIN OUTCOME MEASURES

Tables (2×2) were created for the documented history of penicillin reactions as a predictor for a subsequent reaction.

MAIN RESULTS

With the initial penicillin course, 0.18% of patients had an allergic-like event. Almost 60% of the patients who received

Table 39-5 The Presence of a Previous Penicillin Allergy Predicts a Future Reaction

Test	LR+ (95% CI)	LR- (95% CI)
History of penicillin reaction	11 (8.5-14)	0.98 (0.98-0.99)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

at least 1 prescription for penicillin also received a second prescription ($n = 2017957$). According to the history of the initial response to penicillin, the likelihood ratio (LR) for predicting a second reaction can be derived as shown in Table 39-5.

For patients who had an initial reaction to penicillin, 1.9% had a reaction to the second course of penicillin. For patients who did not have an allergic-like event to the first prescription, 0.17% had a reaction to the second prescription.

The serious reactions to the first penicillin course ($n = 3014$) included anaphylaxis ($n = 16$), angioedema ($n = 106$), laryngospasm ($n = 19$), and toxic epidermal necrolysis ($n = 6$). Most patients had urticaria ($n = 2275$) or erythema multiforme ($n = 237$). The pattern of reactions to the second penicillin course was similar.

About 75% of the prescriptions were for amoxicillin.

CONCLUSIONS

LEVEL OF EVIDENCE Review of a large database with outcomes of uncertain reliability.

STRENGTHS Large database that allows the detection of low-frequency events. The physicians had to concur with the diagnosis, as evidenced by their reporting the diagnostic code.

LIMITATIONS Lack of standardized case definitions. A “case” required a second visit by the patient and appropriate coding, both of which would bias the outcomes to underestimate all allergic-like reactions. Although the specificity of the diagnosis for an allergic-like event seems reasonable, there is an assumption that the event was attributable to the penicillin and not to another drug or to illness. The limitations of this study were addressed in an editorial.¹

It is surprising that 48% of patients with an initial allergic-like event received a second course of penicillin. This could have happened because patients forgot their previous reaction and the physician was therefore unaware or because the previous reaction was attributed to another cause. Few patients (1.89%) had a second reaction. When the authors expanded their case definition of reactions to include bronchospasm, asthma, and eczema, the allergic-like events increased to 9% for patients with a previous reaction. The event rate of 9% matches the event rate for patients with a history of penicillin allergy who have a negative skin test result and then are treated again with penicillin.²

If the physicians were efficient in identifying the patients most likely to have a second reaction, then the positive LR of 11 is underestimated (Table 39-5). To highlight this, we can project the low-event rate of second reactions (1.9%) onto the 3198 initial reactors who did not receive a second course of a penicillin. Had those patients received a second course with an allergic-like event, the positive LR for a previous penicillin reaction would have been 16. The inclusion of those patients creates minimal change in the negative LR (0.95). Given the caveats about the data set, it is probably safe to say that the history of a penicillin reaction documented by a physician confers a positive LR greater than 11 for a second reaction. This LR for penicillin allergy is much higher than the LR for the clinical history in predicting allergy as defined by the response to skin testing.

Reviewed by David L. Simel, MD, MHS

REFERENCES FOR THE EVIDENCE

- Josephson AS. Penicillin allergy: a public health perspective. *J Allergy Clin Immunol*. 2004;113(4):605-606.
- Macy E, Mangat R, Burchette R. Penicillin skin testing in advance of need: multiyear follow-up in 568 test result-negative subjects exposed to oral penicillins. *J Allergy Clin Immunol*. 2003;111(5):1111-1115.

TITLE History of Penicillin Allergy and Referral for Skin Testing: Evaluation of a Pediatric Penicillin Allergy Testing Program.

AUTHORS Langley JM, Halperin SA, Bortolussi R.

CITATION *Clin Invest Med*. 2002;25(5):181-184.

QUESTION Does a history of penicillin allergy predict response to skin testing and oral challenge with penicillin?

DESIGN Prospective, protocol assessment.

SETTING Canadian ambulatory infectious disease clinic.

PATIENTS Seventy-four children referred for possible penicillin allergy. Ninety-six percent had generalized cutaneous eruptions.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

Penicillin allergy was defined by a history of life-threatening anaphylaxis, a positive skin-test result, or no response to an observed oral challenge of penicillin.

MAIN OUTCOME MEASURE

Positive predictive value of the history for penicillin allergy.

MAIN RESULTS

Two patients had “convincing” life-threatening anaphylaxis and 3 had a positive intradermal skin test result. The remaining 69 patients had an oral challenge with penicillin; none had an adverse reaction, so the negative predictive value is 100% (lower 95% confidence interval [CI], 96%). The positive predictive value of the history of penicillin allergy was 6.7% (95% CI, 2.9%-15%).

CONCLUSIONS

LEVEL OF EVIDENCE Positive predictive value study.

STRENGTHS Patients with a negative skin test result for penicillin allergy were given an oral challenge with penicillin.

LIMITATIONS Small population of patients with uncertainty about whether these were consecutive patients. As with all referral studies of penicillin allergy, this likely does not capture the universe of patients with possible penicillin reactions. Nine percent of the patients were referred for cephalosporin reactions.

Although the study was small, the information presented is enhanced by the oral penicillin challenge in patients who had a negative skin test result. Using the generally held notion that about 10% of the population will have a reaction to penicillin, the true allergy rate would be $0.067 \times 0.10 = 0.67\%$. If we take 0.67% as the prior probability and use 6.7% as the posterior probability for a patient with a positive penicillin allergy history, we can solve for the likelihood ratio:

Penicillin allergy likelihood ratio = posterior odds/prior odds, or $(0.067/0.933)/(0.0067/0.9933) = 10.6$.

Reviewed by David L. Simel, MD, MHS

TITLE Penicillin Skin Testing in Advance of Need: Multi-year Follow-up of 568 Test Result–Negative Subjects Exposed to Oral Penicillins.

AUTHORS Macy E, Mangat R, Burchette RJ.

CITATION *Allerg Clin Immunol*. 2003;111(5):1111-1115.

QUESTION Among patients with a history of penicillin allergy, does a negative skin test reaction confirm the lack of penicillin allergy?

DESIGN Retrospective medical record review.

SETTING Allergy clinic as part of a health care management organization.

PATIENTS Patients were adults referred to an allergist for skin testing for suspected penicillin allergy. The symptoms of their allergic reaction were not described.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

The computerized pharmacy records of patients who had a negative skin test result for penicillin were reviewed. Using the narrative of the patient's clinical records, the investigators searched for documentation of an allergic response to penicillin.

MAIN OUTCOME MEASURES

Allergic responses were recorded as anaphylaxis, gastrointestinal reactions, hives, other rashes, or other reactions.

MAIN RESULTS

During 7 years, 1383 patients were skin tested for penicillin allergy. Among this population of patients with a clinical suspicion for penicillin allergy, 137 had positive skin test results (9.9%) (Table 39-6). The charts of the remaining 1246 patients were studied for penicillin exposures; 568 patients received subsequent penicillin challenges. Among the patients with a history of penicillin allergy, with a negative skin test result, and who were challenged with penicillin, 65 had a reaction. None of the reactions were documented as truly anaphylactic (by chart review, upper 95% confidence interval < 0.7%), with most being "hives" (72%) or other rashes (12%).

Table 39-6 The Predictive Value of a History of a Penicillin Allergy Is Modified by Knowing the Response to Skin Testing

Test	Predictive Value (95% CI) for Positive Skin Test Result, %	Predictive Value (95% CI) for No Allergic Response to Penicillin Administration, %	Predictive value (95% CI) for Positive Skin Test Result or Subsequent Allergic Reaction, %
History of penicillin allergy	9.9 (8.4-12)		15 (13-17)
History of penicillin allergy with negative skin test result		90 (86-91)	

Abbreviation: CI, confidence interval.

CONCLUSIONS

LEVEL OF EVIDENCE Level 4.

STRENGTHS Large sample size of patients referred for a possible penicillin reaction.

LIMITATIONS Retrospective chart review, relying on non-standardized clinical documentation of reactions. Patients may have received care outside of the health care management organization, so their results would not have been captured.

By defining true penicillin allergy as a positive skin test response or an allergic reaction to a second course of penicillin, then the positive predictive value of a patient's history of a penicillin allergy is 15%.

With the skin test result as a reference standard, about 10% of patients with a reported reaction to penicillin will have a positive reaction. A negative skin test result among these patients makes a subsequent anaphylactic reaction unlikely (<1%). However, patients with a negative skin test result have at least a 10% rate of subsequent reactions (most are skin reactions). The retrospective nature of this study design probably means that the rate is higher because patients with less severe skin reactions might not have sought care and some may have received care from other providers.

Reviewed by David L. Simel, MD, MHS

This page intentionally left blank

CHAPTER 40

Does This Adult Patient Have Community-Acquired Pneumonia?

Joshua P. Metlay, MD, PhD

Wishwa N. Kapoor, MD, MPH

Michael J. Fine, MD, MSc

CLINICAL SCENARIO

Does This Patient Have Pneumonia?

A 53-year-old woman comes to your office with a cough of more than 1 week's duration. She was in excellent health until 7 days ago, when she developed a nonproductive cough, mild sore throat, and myalgia. She recalls no history of asthma or chronic obstructive pulmonary disease, and she does not smoke. Despite staying home from work for the last 2 days, she has noted increasing sputum production with her cough and worsening fatigue. She has felt warm but has not documented any fever or night sweats. On physical examination, her oral temperature is 38.3°C (101°F), her heart rate is 110/min, and auscultation of her chest reveals inspiratory crackles on the left side.

WHY IS THIS AN IMPORTANT QUESTION TO ANSWER WITH A CLINICAL EXAMINATION?

Physicians commonly encounter patients with respiratory complaints similar to those in the clinical scenario. In 1994, there were more than 10 million visits to primary care physicians by adults with a chief complaint of cough, representing more than 4% of all visits to physicians that year. Pneumonia represented only 5% of all causes for these visits and was the fifth leading diagnosis, after bronchitis, upper respiratory tract infection, asthma, and sinusitis.¹ Though pneumonia may represent a small proportion of all acute respiratory illnesses, the accurate identification of this subgroup is important because of the distinct therapeutic and prognostic features of this illness.

In the preantibiotic era, mortality from pneumococcal pneumonia was consistently higher than 20% for all cases, increasing to more than 60% for bacteremic cases.² Since the introduction of antibiotics, no one has reported results from large-scale studies comparing antibiotic therapy to nonantibiotic therapy for patients with pneumonia. As a result, such therapy is universally recommended and has become a standard of care for all patients with pneumonia. No such standard exists for alternative respiratory infections such as bronchitis³ or the common cold.⁴ Moreover, inappropriate use of antibiotics for these alternative respiratory infections may be an important determinant of the increase in antibiotic resistance among common respiratory pathogens.^{5,6}

In terms of prognosis, patients with pneumonia continue to have an overall high mortality from this illness, ranging from as low as 5% in studies of hospitalized and ambulatory patients to as high as 37% in studies of patients requiring admission to intensive care units.⁷ This persistently high mortality underscores the need for physicians to choose carefully between home or hospital therapy for all patients with pneumonia.⁸ For these reasons, physicians need to know how

to use their clinical examination optimally to identify patients at suitable risk for pneumonia to require further, definitive diagnostic testing.

Chest radiography is the reference standard for diagnosing community-acquired pneumonia and provides additional information on the prognosis of patients with this illness,⁹ as well as the presence of coexisting conditions such as bronchial obstruction or pleural effusions.¹⁰ Moreover, chest radiography is highly reliable,¹¹ safe, generally available, and relatively inexpensive, so that it is a standard part of the evaluation of any patient with suspected pneumonia. It is possible that some physicians continue to diagnose and treat patients with pneumonia without the aid of chest radiography, whereas other physicians routinely obtain chest radiographs for all patients suspected of having pneumonia. We do not know the proportion of physicians who choose these alternative strategies. Therefore, the aims of this article are both to assess the validity of the former approach (diagnosing pneumonia without chest radiography, using medical history and physical examination alone) and to identify elements of the clinical examination that might improve the efficiency of the latter approach (ordering chest radiographs for all patients with suspected pneumonia).

PATOPHYSIOLOGY OF COMMUNITY-ACQUIRED PNEUMONIA

In patients with community-acquired pneumonia, the site of infection can involve the pulmonary interstitium, alveoli, or both. This provides the physiologic basis for the principal chest examination findings in pneumonia, which include dullness to percussion, changes in the intensity of tactile fremitus and breath sounds, and inspiratory crackles. Dullness to percussion and local changes in the intensity of tactile fremitus and breath sounds are the result of diffuse replacement of the pulmonary parenchyma with inflammatory tissue, leading to pulmonary consolidation or the presence of pleural effusions.¹² In a patient with pneumonia, crackles (formerly called "rales") are caused by the delayed opening of alveoli in deflated regions of pathologically inflamed lung.¹³ Crackles refer to any discontinuous adventitious lung sounds and can therefore be heard in a variety of pulmonary diseases that cause lung stiffening, including congestive heart failure, pulmonary fibrosis, and obstructive lung disease.¹²

HOW TO ELICIT THESE SYMPTOMS AND SIGNS

Patients with community-acquired pneumonia present with a large number of possible symptoms. In a study of more than 1800 patients with community-acquired pneumonia, these presenting symptoms ranged from typical respiratory complaints, including productive cough, dyspnea, and pleuritic chest pain, to predominantly systemic complaints of fatigue, anorexia, and myalgias. Moreover, the pattern of presenting symptoms varied considerably among patients, particularly among elderly patients with pneumonia, who less frequently reported a wide range of symptoms.¹⁴ As a result, careful history-taking in a

patient suspected of having community-acquired pneumonia should consider a broad range of possible symptoms, including respiratory and nonrespiratory symptoms.

In contrast, the examination of the chest in patients with suspected pneumonia is traditionally carried out in a structured manner, proceeding through the 4 steps of inspection, palpation, percussion, and auscultation. The chest is inspected for signs of asymmetric chest expansion, defined as a visible difference in excursion between the 2 sides of the chest. The chest wall is palpated while the patient speaks to assess the transmission of sound, or tactile fremitus. Percussion over symmetric areas of the anterior and posterior chest wall detects diminution in the resonance of the percussion note, or dullness to percussion. Finally, auscultation of the lung assesses the intensity of normal breath sounds, the transmission of spoken words, and the presence of adventitious sounds. Auscultation in the peripheral lung fields may detect the replacement of the normal vesicular breath sounds with tubular or bronchial breath sounds, which are normally heard only over the trachea. Increased transmission of speech may be detected as the increased clarity of whispered phrases, known as whispered pectoriloquy, or as the change in timbre of vowel sounds in the form of "e" to "a," known as egophony.¹² The principal abnormal sounds in community-acquired pneumonia are known as crackles, which are nonmusical, discontinuous sounds and should be detected with the patient in the upright position. It has been suggested that auscultation of each lung in the lateral dependent position is a more sensitive technique for crackles, but this has not been independently validated.¹⁵ Auscultation should occur with the patient breathing at normal tidal volumes because inspiration from lower lung volumes (ie, residual volume) can yield abnormal auscultatory findings in as many as 50% of normal subjects.¹⁶ Finally, both percussion and auscultation of the chest should proceed in a systematic fashion, with an examination of symmetric areas on the anterior and posterior chest wall.

METHODS

Literature Search

We searched English-language medical literature to determine the precision of the clinical examination in patients with community-acquired pneumonia and the accuracy of the examination in diagnosing patients suspected of having this illness. We searched MEDLINE from 1966 through October 1995 according to an initial search strategy similar to that used by other authors in this series. The initial retrieval of titles ($n = 7$ for precision; $n = 140$ for diagnostic accuracy) was reviewed by 2 of us (J.P.M. and M.J.F.). Articles that focused on hospital-acquired pneumonia, pediatric pneumonia, or AIDS-related pneumonia were excluded. The remaining articles were retrieved, as well as any potentially eligible articles identified through review of the article reference lists ($n = 7$ for precision; $n = 52$ for diagnostic accuracy).

A set of explicit inclusion and exclusion criteria was applied to each retrieved article. Inclusion criteria required

that the study be an original study of the accuracy or precision of the medical history or physical examination in determining the diagnosis of community-acquired pneumonia. Exclusion criteria consisted of studies of patients younger than 16 years, patients with known immunosuppression, or patients with nosocomial infections. In addition, case series (<10 observations) or review articles without original data were excluded.

Quality Review of Articles

The remaining eligible articles were each evaluated by one of us (J.P.M.) according to a methodologic quality filter that assigned a level of evidence from 1 to 5 according to the internal validity of the study. Level 1 evidence refers to a primary, prospective study of the accuracy or precision of the clinical examination in community-acquired pneumonia. For studies dealing with accuracy, this requires independent, blind comparisons of clinical findings with a criterion standard (or gold standard) of diagnosis or etiology among a large number (>50) of consecutive patients suspected of having community-acquired pneumonia. For studies dealing with precision, this requires 2 or more independent blinded raters of symptoms or signs in a large number of patients suspected of having community-acquired pneumonia. Level 2 studies were analogous to level 1 studies but with smaller numbers of patients (10-50), widening the confidence limits of the resulting calculations. Level 3 studies were based on a retrospective design (ie, clinical findings determined by chart review). Level 4 studies included nonconsecutive patients, generally selected because of their definitive results for the findings under study, or a nonblinded comparison of clinical findings with a gold standard. Level 5 studies included studies with an uncertain gold standard or a poorly defined study population (ie, may not even have community-acquired pneumonia). For the purposes of this study, only studies of level 1 quality, also called grade A evidence, were considered for the main analyses and tables. Summaries of relevant level 2 through 5 studies are provided in the text.

Data Analysis

Likelihood ratios (LRs) were calculated for the presence (positive LR [LR+]) or absence (negative LR [LR-]) of individual clinical findings.^{17,18} Only those findings significantly associated with the presence or absence of pneumonia in at least 1 study, based on a 2-tailed χ^2 or Fisher exact test with P less than .05, were included in the results. However, the actual diagnostic value of statistically significant findings still depends on both the prior probability of pneumonia and how much the LR moves the posterior probability from the prior probability.¹⁹

RESULTS

Precision of Symptoms and Signs of Community-Acquired Pneumonia

Interobserver variation in the recording of the presence of symptoms in patients with community-acquired pneumonia

has not been directly examined. However, analogous work in assessing symptom prevalence in large-scale epidemiologic studies has revealed considerable interobserver variation in the recording of symptoms.^{20,21} This has led to the adoption of standardized respiratory questionnaires in epidemiologic studies of chronic respiratory illnesses. However, no such standardized questionnaires exist for recording symptoms in patients with acute respiratory infections.²²

It has also been appreciated for some time that the physical examination of the chest is hampered by a high degree of interobserver error. Although no study has specifically addressed the reliability of the physical examination in patients with community-acquired pneumonia, Spiteri et al²³ measured reliability among 24 physicians in the examination of 24 patients with a variety of respiratory conditions, 4 of whom had radiographic evidence of pneumonia. Table 40-1 presents the calculated interobserver reliability among the physicians for several chest signs. The results are presented in the form of both mean pair observer agreement rates and κ values, which account for rates of chance agreement ranging from 0, when agreement is no better than chance, to 1, when there is perfect agreement. In fact, 2 of the most reliable findings, dullness to percussion and wheezes on auscultation, had only fair to good κ values of 0.52 and 0.51, corresponding to agreement rates of 77% and 79%, respectively. Crackles had a κ value of 0.41 (agreement rate of 72%), and several findings such as whispered pectoriloquy and increased tactile fremitus had κ values indicating poor agreement (range, 0.01-0.11), in part explained by the rarity of these findings overall.

Similarly poor interobserver reliability has been observed in the chest examination of other respiratory diseases. For example, Schilling et al²⁴ observed an agreement rate of 76% for abnormal chest sounds in the examination of 187 men with interstitial lung disease and 88 controls; this yields a κ value of 0.25. Smyllie et al²⁵ measured agreement rates among 9 physicians who examined 20 patients with a variety of chronic lung diseases. Agreement rates were generally midway between chance and perfect agreement for a number of chest examinations.

Table 40-1 Precision of Physical Examination Findings in Examination of the Chest^a

Physical Examination Finding	Agreement, % ^b	κ Value
Tachypnea	63	0.25
Reduced chest movement	70	0.38
Increased tactile fremitus	85	0.01
Dullness to percussion	77	0.52
Decreased breath sounds	... ^c	0.43
Wheezes	79	0.51
Crackles	72	0.41
Bronchial breath sounds	... ^c	0.32
Whispered pectoriloquy	... ^c	0.11

^aAdapted from Spiteri et al.²³

^bCalculated according to data provided in Table 1 by Spiteri et al.²³

^cEllipses indicate mean pair agreement rates were not calculated for the signs for which 2 or more physicians in a group failed to report the presence or absence of the sign.

tion findings, including diminished breath sounds, decreased percussion note, and crackles. Though the basis for the relatively low interobserver reliability in chest examination is unknown, at least 1 group has suggested that deficiencies in the teaching of the clinical examination are to blame.²³

Accuracy of the Clinical History in the Diagnosis of Community-Acquired Pneumonia

For this review, 4 studies were judged to have level 1 evidence on the test characteristics of individual items in the clinical history in the diagnosis of community-acquired pneumonia.²⁶⁻²⁹ In each of these studies, the reference standard for the diagnosis of pneumonia was a new infiltrate on a chest radiograph. Table 40-2 summarizes the value of findings from the medical history, including respiratory symptoms, nonrespiratory symptoms, and information on medical history.

Though all 4 studies were based in emergency departments, variations in the patterns of the results reflect, in part, variation in the selection criteria for each study. For example, in the study by Diehr et al,²⁶ chest radiographs were obtained for all patients presenting with acute cough, whereas the other studies obtained chest radiographs only when the primary physician previously determined a need for them, often to confirm or exclude a suspected diagnosis of pneumonia. The latter approach provides a more highly selected population of patients with acute respiratory complaints that may alter the measured test characteristics

of individual clinical findings. This selection bias is reflected in the fact that the prevalence of pneumonia in the study populations ranged from as low as 2.6%²⁶ to as high as 38%.²⁷

Still, certain patterns emerge. For example, there are no individual items from the clinical history whose presence or absence would reduce the odds of disease sufficiently to exclude pneumonia and eliminate the need to obtain a chest radiograph. The one exception to this is the presence of a medical history of asthma, which reduces the odds of pneumonia with an LR of 0.1, though this has been demonstrated in only 1 study.²⁹

Similarly, the presence of no single item in the clinical history increases the odds of pneumonia high enough to confirm the diagnosis without a chest radiograph. Though the presence of findings with an LR+ ranging from 2 (fever or immunosuppression) to 3 (history of dementia) may be helpful, they are not confirmatory, particularly given the typically low prevalence of pneumonia in the study populations. For example, in the study by Diehr et al,²⁶ the presence of subjective fever (LR+, 2.1; 95% confidence interval [CI], 1.4-2.9) had a positive predictive value of only 5.5%, reflecting the low prevalence of pneumonia in the population.

Accuracy of Physical Examination Findings in the Diagnosis of Community-Acquired Pneumonia

Table 40-3 summarizes the accuracy of 10 potential findings (3 vital signs and 7 abnormal findings on chest examination)

Table 40-2 Likelihood Ratios for Pneumonia, Given the Presence or Absence of Individual Medical History Findings^a

Findings ^a	Source, Year							
	LR+ ^b				LR- ^c			
	Diehr et al, ²⁶ 1984	Gennis et al, ²⁷ 1989	Singal et al, ²⁸ 1989	Heckerling et al, ²⁹ 1990	Diehr et al, ²⁶ 1984	Gennis et al, ²⁷ 1989	Singal et al, ²⁸ 1989	Heckerling et al, ²⁹ 1990
Respiratory symptoms								
Cough	... ^d	NS	1.8	NS	...	NS	0.31	NS
Dyspnea	...	1.4	NS	NS	...	0.67	NS	NS
Sputum production	1.3	NS	...	NS	0.55	NS	...	NS
Nonrespiratory symptoms								
Fever	2.1	NS	...	1.7	0.71	NS	...	0.59
Chills	1.6	1.3	...	1.7	0.85	0.72	...	0.70
Night sweats	1.7	0.83
Myalgias	1.3	NS	0.58	NS
Sore throat	0.78	NS	1.6	NS
Rhinorrhea	0.78	NS	2.4	NS
Medical history								
Asthma	0.10	3.8
Immunosuppression	2.2	0.85
Dementia	3.4	0.94

Abbreviations: LR+, positive likelihood ratio; LR-, negative likelihood ratio; NS, result not significant.

^aOnly those findings significantly associated with the presence or absence of pneumonia in at least 1 study are included ($P < .05$ in a 2-tailed χ^2 or Fisher exact test).

^bLR+ for pneumonia when symptom present, sensitivity/(1 - specificity).

^cLR- for pneumonia when symptom absent, (1 - sensitivity)/specificity.

^dEllipses indicate result is not available.

Table 40-3 Likelihood Ratios for Pneumonia, Given the Presence or Absence of Physical Examination Findings^a

Findings	Source, Year							
	LR+ ^b				LR- ^c			
	Diehr et al, ²⁶ 1984	Gennis et al, ²⁷ 1989	Singal et al, ²⁸ 1989	Heckerling et al, ²⁹ 1990	Diehr et al, ²⁶ 1984	Gennis et al, ²⁷ 1989	Singal et al, ²⁸ 1989	Heckerling et al, ²⁹ 1990
Vital signs								
Respiratory rate, breaths/min								
>20	...	1.2	0.66
>25	3.4	...	NS ^e	1.5	0.78	...	NS	0.82
>30	...	2.6	0.80
Heart rate, beats/min								
>100	NS	1.6	NS ^e	2.3	NS	0.73	NS	0.49
>120	...	1.9	0.89
Temperature > 37.8°C (100°F)	4.4	1.4	2.4	2.4	0.78	0.63	0.68	0.58
Any abnormal vital sign	...	1.2	0.18
Chest examination								
Asymmetric respiration	∞	0.96
Dullness to percussion	NS	2.2	...	4.3	NS	0.93	...	0.79
Decreased breath sounds	...	2.3	...	2.5	...	0.78	...	0.64
Crackles	2.7	1.6	1.7	2.6	0.87	0.83	0.78	0.62
Bronchial breath sounds	3.5	0.90
Rhonchi	NS	1.5	...	1.4	NS	0.85	...	0.76
Egophony	8.6	2.0	...	5.3	0.96	0.96	...	0.76
Any chest finding	...	1.3	0.57

Abbreviations: LR+, positive likelihood ratio; LR-, negative likelihood ratio; NS, result not significant.

^aOnly those findings that were significantly associated with the presence or absence of pneumonia in at least 1 study are included ($P < .05$ in a 2-tailed χ^2 or Fisher exact test).

^bLR+ for pneumonia when finding present, sensitivity/(1 – specificity).

^cLR- for pneumonia when finding absent, (1 – sensitivity)/specificity.

^dEllipses indicate result is not available.

^eActual cut points not specified in this study.

from the physical examination in patients with suspected pneumonia according to results from the 4 previously identified studies. LRs for the presence of any individual vital sign abnormality (LR+), including tachypnea, tachycardia, or fever, ranged from 2 to 4. Moreover, various cut points for these abnormalities did not have a substantial effect on the calculated LRs.²⁷ Similarly, LRs for the absence of any individual vital sign abnormality (LR-) ranged from 0.5 to 0.8. However, Gennis et al²⁷ demonstrated an LR- of 0.18 (95% CI, 0.07–0.46) for the diagnosis of pneumonia according to the absence of all 3 vital sign abnormalities (ie, respiratory rate $< 30/\text{min}$, heart rate $< 100/\text{min}$, and temperature $< 37.8^\circ\text{C}$ [100°F]). According to this finding, if the baseline prevalence of pneumonia among ambulatory patients with respiratory illnesses is assumed to be 5%, a patient without any vital sign abnormalities would have a predicted probability of pneumonia of less than 1%.

The presence of several findings on chest examination significantly raised the likelihood of pneumonia. For example, in one study the presence of asymmetric respirations essentially guaranteed the diagnosis of pneumonia (LR+, ∞ ; 95% CI, 3.2– ∞).²⁶ However the usefulness of this finding was limited because only 4% of patients with pneumonia

had asymmetric respirations. The presence of other findings, including egophony and dullness to percussion, significantly increased the likelihood of pneumonia. However, given the low prevalence of pneumonia in the overall study populations, the effect of observing these findings on estimating the probability of pneumonia was only modest. For example, the presence of egophony had a positive predictive value ranging from as low as 20%²⁶ to no higher than 56%.²⁷

Finally, all 4 studies support the conclusion that the presence or absence of crackles on examination would not be sufficient to rule in or rule out the diagnosis. For example, with a prevalence of pneumonia of 5%, the absence of crackles reduces the probability to 3%, at the lowest, and the presence of crackles increases the probability to 10%, at the highest. Moreover, the absence of any abnormality on chest examination yielded an LR- of 0.57 (95% CI, 0.39–0.83),²⁷ which is too close to the indeterminate LR value of 1.0 to substantially reduce the probability of pneumonia.

The low accuracy of individual findings on chest examination for detecting pneumonia has also been supported by studies that relied on retrospective data gathering^{30,31} or incomplete application of chest radiography to all study

patients.³² In one study, the absence of crackles yielded an LR $-$ of only 0.71 (95% CI, 0.47-0.90), and the absence of any abnormal auscultatory finding yielded an LR $-$ of only 0.68 (95% CI, 0.44-0.89), both of which would translate into small effects on the probability of pneumonia.³² In contrast, another study found that the absence of any abnormality on chest auscultation resulted in an LR $-$ of 0.13 (95% CI, 0.07-0.24),³¹ which might substantially reduce the probability of pneumonia. However, this result has not been replicated in prospective studies, which would be subject to less bias in the recording of physical examination findings.

Evaluating Algorithms to Predict Pneumonia

Because the accuracy of individual symptoms or signs for predicting pneumonia is low, several studies developed prediction rules that incorporate the presence or absence of several medical history or physical examination findings. **Table 40-4** summarizes the features of 3 such rules. Though initially designed as aids in the ordering of chest radiographs for patients with suspected pneumonia, they are reasonably considered as prediction rules for the diagnosis of pneumonia in these patients and yield probabilities of pneumonia after

completion of the clinical examination. For the rule by Diehr et al,²⁶ points are assigned for each clinical finding and summed to yield a discriminant score. For example, a threshold score of -1 (ie, all patients with scores ≥ -1 are considered to have pneumonia) yields an LR $+$ of 1.5 and an LR $-$ of 0.22, a threshold score of +1 yields an LR $+$ of 5.0 and an LR $-$ of 0.47, and a threshold score of +3 yields an LR $+$ of 14 and an LR $-$ of 0.82, according to the original study data.²⁶ The rule by Singal et al²⁸ is a logistic function that can yield probabilities of pneumonia, ranging from 4% (no findings present) to 49% (all 3 findings present).²⁸

The final prediction rule, by Heckerling et al,²⁹ is based on the presence or absence of 5 clinical findings. The performance of this prediction rule depends on the pretest probability of pneumonia in the population. In most ambulatory care settings, this probability will be relatively low. For example, as observed earlier, in a national survey, only 5% of all patients visiting primary care physicians for cough were diagnosed as having pneumonia.¹ In this setting, the presence of 2, 3, or 4 predictors would result in predicted probabilities of pneumonia of 3%, 10%, or 25%, respectively, according to a nomogram provided by Heckerling et al.²⁹ The rule would yield a maximum probability of pneumonia of 50% if all 5 of its clinical predictors were present. These findings emphasize the inaccuracy in diagnosing pneumonia clinically, in the absence of confirmatory chest radiography.

The 3 scores summarized in Table 40-4, along with the decision rule suggested by Gennis et al²⁷ (ie, obtaining chest radiographs only for patients suspected of having pneumonia, with at least 1 vital sign abnormality), were compared for their ability to predict correctly the results of chest radiography in an independent study by Emerman et al.³³ Patients presenting to an emergency department or outpatient medical clinic with a complaint of cough were enrolled prospectively, and chest radiographs were obtained for all patients regardless of the primary physician's clinical impression.

Overall, the prevalence of pneumonia among the study patients was 7%. In the absence of an explicit guideline, physician judgment that the patient did not need chest radiography reduced the probability of pneumonia to just less than 2% (LR $-$, 0.25; 95% CI, 0.09-0.61), which exceeded all 4 prediction rules. In contrast, physician judgment that the patient needed chest radiography to diagnose pneumonia increased the probability of pneumonia to only 13% (LR $+$, 2.0; 95% CI, 1.5-2.4), which meant that reliance on implicit physician judgment alone would have led to many unnecessary chest radiographs.

In comparison, the simple decision rule by Gennis et al²⁷ ordering chest radiographs only for patients with abnormal vital signs yielded the highest overall LR $+$ for predicting pneumonia, but the LR $+$ was a modest 2.6 (95% CI, 1.6-3.7). With this rule, 40% fewer radiographs would have been ordered compared with unaided physician judgment. However, excluding pneumonia according to the absence of any vital sign abnormalities would have missed 38% of patients subsequently shown to have pneumonia on chest radiography (LR $-$, 0.50 [95% CI, 0.27-0.78], compared with LR $-$, 0.18 in the original study by Gennis et al²⁷).

Table 40-4 Predictive Rules for Pneumonia Diagnosed by Chest Radiography^a

Diehr et al ²⁶	
Add points when present ^b	
Rhinorrhea	-2 Points
Sore throat	-1 Point
Night sweats	1 Point
Myalgias	1 Point
Sputum all day	1 Point
Respiratory rate > 25/min	2 Points
Temperature $\geq 37.8^{\circ}\text{C}$ (100°F)	2 Points
Singal et al ²⁸	
Probability ^c = $1/(1 + e^{-Y})$	
$Y = -3.095 + 1.214 \text{ (Cough)} + 1.007 \text{ (Fever)} + 0.823 \text{ (Crackles)}$	
Each variable = 1 if present	
Heckerling et al ²⁹	
Determine the number of findings present ^d	
Absence of asthma	
Temperature $> 37.8^{\circ}\text{C}$ (100°F)	
Heart rate $> 100/\text{min}$	
Decreased breath sounds	
Crackles	

Abbreviations: LR $+$, positive likelihood ratio; LR $-$, negative likelihood ratio.

^aAdapted from Emerman et al.³³

^bFor example, a threshold score of -1 (ie, all patients with scores ≥ -1 are considered to have pneumonia) yields an LR $+$ of 1.5 and an LR $-$ of 0.22; a threshold score of +1 yields an LR $+$ of 5.0 and an LR $-$ of 0.47; and a threshold score of +3 yields an LR $+$ of 14.0 and an LR $-$ of 0.82, according to the original study data.²⁶

^cFirst calculate Y and then calculate the predicted probability of pneumonia.

^dFor example, according to a prevalence of pneumonia of 5%, the presence of 0, 1, 2, 3, 4, or 5 findings yields probabilities of pneumonia of <1%, 1%, 3%, 10%, 25%, and 50%, respectively, according to a nomogram provided by Heckerling et al.²⁹

An algorithm that is less than perfect, that is, not all ordered chest radiographs demonstrate a new infiltrate, will still be acceptable, given the relatively low cost and risk associated with this test. Ultimately, optimum yields for chest radiography in the evaluation of patients with suspected pneumonia will need to be determined, balancing the costs of the test with the costs of missed diagnoses. Additional factors, such as illness severity and patient preferences, will also play a role in determining the appropriate threshold for ordering chest radiographs for patients with acute respiratory illnesses. For example, thresholds may be lower for patients who appear severely ill or who express strong desires to have a definitive diagnosis. We suggest that an algorithm that yields less than a 100% negative predictive value may still be acceptable, assuming that patients with missed cases of pneumonia continue to have good clinical outcomes. However, this hypothesis will need to be tested.

CLINICAL SCENARIO—RESOLUTION

The patient presents with typical symptoms of community-acquired pneumonia, including a productive cough and fever. Physical examination reveals fever and crackles on chest auscultation. In particular, the patient has 4 of the 5 clinical pneumonia predictors identified by Heckerling et al²⁹ (absence of asthma, presence of fever, tachycardia, and crackles). With the nomogram by Heckerling et al,²⁹ a 5% prevalence of pneumonia among outpatients yields a 25% probability of pneumonia. Similarly, the patient is at the threshold score of +3 points on the prediction rule by Diehr et al²⁶ (presence of sore throat, sputum, myalgias, and fever), yielding an LR for pneumonia of 14 (according to the original study data) and a calculated probability of pneumonia of 42%. Finally, the patient has all 3 criteria of Singal et al,²⁸ yielding a probability of pneumonia of 49%, according to their logistic formula. We conclude that none of these combinations of findings can be said to “rule in” the diagnosis, yet the possibility of pneumonia remains high enough that further diagnostic testing, in particular chest radiography, is warranted.

THE BOTTOM LINE

Physicians frequently disagree about the presence or absence of individual findings on chest examinations of patients with respiratory illnesses, including community-acquired pneumonia.

Individual symptoms and signs have inadequate test characteristics to rule in or rule out the diagnosis of pneumonia. Decision rules that use the presence or absence of several symptoms and signs to modify the probability of pneumonia are available, the simplest of which requires the absence of any vital sign abnormalities to exclude the diagnosis. There are no combinations of medical history and physical examination findings that confirm the diagnosis of pneumonia. If diagnostic certainty is required in the treatment of a patient

with suspected pneumonia, then chest radiography should be performed.

Future research should examine ways to improve the precision of the clinical examination in patients with suspected pneumonia, as well as to determine the accuracy of the clinical examination in these patients in settings outside the emergency department. In addition, studies should address appropriate thresholds for obtaining chest radiographs and treating accordingly vs empirical antimicrobial therapy vs clinical observation in the treatment of patients with suspected community-acquired pneumonia.

Author Affiliations at the Time of the Original Publication

General Internal Medicine Unit, Department of Medicine, Massachusetts General Hospital and Harvard Medical School, Boston (Dr Metlay); Division of General Internal Medicine, Department of Medicine, University of Pennsylvania Medical Center, Philadelphia (Dr Metlay); Division of General Internal Medicine, Department of Medicine, Center for Research on Health Care, University of Pittsburgh, Pittsburgh, Pennsylvania (Drs Kapoor and Fine).

Acknowledgments

Dr Metlay was supported by a General Medicine Research Fellowship, National Research Service Award 5 T32 PE11001-08. Dr Fine was supported as a Robert Wood Johnson Foundation Generalist Physician Faculty Scholar.

The authors gratefully acknowledge the contribution of Melanie E. Smith, MPIA, to the development of the MEDLINE search strategy and initial retrieval of articles. We also thank Daniel E. Singer, MD, for editorial suggestions on an earlier version of the manuscript.

REFERENCES

- Metlay JP, Stafford RS, Singer DE. National trends in the management of acute cough by primary care physicians. *J Gen Intern Med*. 1997;12 (suppl):77.
- Bullowa JGM, Wilcox C. Incidence of bacteremia in the pneumonias and its relation to mortality. *Arch Intern Med*. 1935;55:558-573.
- Orr PH, Scherer K, Macdonald A, Moffatt MEK. Randomized placebo-controlled trials of antibiotics for acute bronchitis: a critical review of the literature. *J Fam Pract*. 1993;36(5):507-512.
- Mainous AG, Hueston WJ, Clark JR. Antibiotics and upper respiratory infection: do some folks think there is a cure for the common cold? *J Fam Pract*. 1996;42(4):357-361.
- Jernigan DB, Cetron MS, Breiman RF. Minimizing the impact of drug-resistant *Streptococcus pneumoniae* (DRSP): a strategy from the DRSP Working Group. *JAMA*. 1996;275(3):206-209.
- Gonzales R, Steiner JF, Sande M. Antibiotic prescribing for adults with colds, upper respiratory tract infections, and bronchitis by ambulatory care physicians. *JAMA*. 1997;278(11):901-904.
- Fine MJ, Smith MA, Carson CA, et al. Prognosis and outcomes of patients with community-acquired pneumonia: a meta-analysis. *JAMA*. 1996;275(2):134-141.
- Fine MJ, Auble TE, Yealy DM, et al. A prediction rule to identify low-risk patients with community-acquired pneumonia. *N Engl J Med*. 1997; 336(4):243-250.
- Hasley PB, Albaum MN, Li YH, et al. Do pulmonary radiographic findings at presentation predict mortality in patients with community-acquired pneumonia? *Arch Intern Med*. 1996;156(19):2206-2212.
- American Thoracic Society. Guidelines for the initial management of adults with community-acquired pneumonia: diagnosis, assessment of severity, and initial antimicrobial therapy. *Am Rev Respir Dis*. 1993;148(5):1418-1426.

11. Albaum MN, Hill LC, Murphy M, et al. Interobserver reliability of the chest radiograph in community-acquired pneumonia. *Chest*. 1996;110(2):343-350.
12. Sapira JD. *The Art and Science of Bedside Diagnosis*. Baltimore, MD: Williams & Wilkins; 1990.
13. Forgacs P. Lung sounds. *Br J Dis Chest*. 1969;63(1):1-12.
14. Metlay JP, Schulz R, Li YH, et al. Influence of age on symptoms at presentation in community-acquired pneumonia. *Arch Intern Med*. 1997;157(13):1453-1459.
15. Gilbert VE. Detection of pneumonia by auscultation of the lungs in the lateral decubitus positions. *Am Rev Respir Dis*. 1989;140(4):1012-1016.
16. Thacker RE, Kraman SS. The prevalence of auscultatory crackles in subjects without lung disease. *Chest*. 1982;81(6):672-674.
17. Koopman PAR. Confidence intervals for the ratio of two binomial proportions. *Biometrics*. 1984;40:513-517.
18. Centor RM. Estimating confidence intervals of likelihood ratios. *Med Decis Making*. 1992;12(3):229-233.
19. Sackett DL. A primer on the precision and accuracy of the clinical examination. *JAMA*. 1992;267(19):2638-2644.
20. Cochrane AL, Chapman PJ, Oldham PD. Observers' errors in taking medical histories. *Lancet*. 1951;1(6662):1007-1009.
21. Fletcher CM. The problem of observer variation in medical diagnosis with special reference to chest diseases. *Method Inform Med*. 1964;3:98-103.
22. Graham NMH. The epidemiology of acute respiratory infections in children and adults: a global perspective. *Epidemiol Rev*. 1990;12:149-178.
23. Spiteri MA, Cook DG, Clarke SW. Reliability of eliciting physical signs in examination of the chest. *Lancet*. 1988;1(8590):873-875.
24. Schilling RSF, Hughes JPW, Dingwall-Fordyce I. Disagreement between observers in an epidemiological study of respiratory disease. *BMJ*. 1955;1(4905):65-68.
25. Smylie HC, Blendis LM, Armitage P. Observer disagreement in physical signs of the respiratory system. *Lancet*. 1966;2(7409):412-413.
26. Diehr P, Wood RW, Bushyhead J, Krueger L, Wolcott B, Tompkins RK. Prediction of pneumonia in outpatients with acute cough: a statistical approach. *J Chronic Dis*. 1984;37(3):215-225.
27. Gennis P, Gallagher J, Falvo C, Baker S, Than W. Clinical criteria for the detection of pneumonia in adults: guidelines for ordering chest roentgenograms in the emergency department. *J Emerg Med*. 1989;7(3):263-268.
28. Singal BM, Hedges JR, Radack KL. Decision rules and clinical prediction of pneumonia: evaluation of low-yield criteria. *Ann Emerg Med*. 1989;18(1):13-20.
29. Heckerling PS, Tape TG, Wigton RS, et al. Clinical prediction rule for pulmonary infiltrates. *Ann Intern Med*. 1990;113(9):664-670.
30. Osmur JC, Cole BK. The stethoscope and roentgenogram in acute pneumonia. *South Med J*. 1966;59(1):75-77.
31. Heckerling PS. The need for chest roentgenograms in adults with acute respiratory illness: clinical predictors. *Arch Intern Med*. 1986;146(7):1321-1324.
32. Melbye H, Straume B, Aasebo U, Dale K. Diagnosis of pneumonia in adults in general practice: relative importance of typical symptoms and abnormal chest signs evaluated against a radiographic reference standard. *Scand J Prim Health Care*. 1992;10(3):226-233.
33. Emerman CL, Dawson N, Speroff T, et al. Comparison of physician judgment and decision aids for ordering chest radiographs for pneumonia in outpatients. *Ann Emerg Med*. 1991;20(11):1215-1219.

Prepared by David L. Simel, MD, MHS

Reviewed by Joshua P. Metlay, MD

CLINICAL SCENARIO

A 36-year-old man with no underlying medical illness developed a cough 3 days ago. In the past 24 hours, his cough became productive of darkened sputum and he observed some wheezing for the first time. He decided to try to go to work, but an episode of chills made him realize he needed to see an urgent care physician. On examination, you find that his temperature is 38.2°C. He does not have tachypnea or tachycardia, although he is wheezing. You do not hear any areas of decreased breath sounds or pulmonary rales. On hearing the wheezing, you inquire to find that he has no history of asthma and that he is not a smoker.

UPDATED SUMMARY ON ADULT COMMUNITY-ACQUIRED PNEUMONIA

Original Review

Metlay JP, Kapoor, Fine WN. Does this patient have community-acquired pneumonia? diagnosing pneumonia by history and physical examination. *JAMA*. 1997;278(17):1440-1445.

UPDATED LITERATURE SEARCH

Our literature search combined the search terms “community-acquired infections or pneumonia” with the parent search strategy for The Rational Clinical Examination series, “meta-analysis,” or “roc curve,” limited to English-language publications in the MEDLINE database from 1995 to November 2004. The search strategy yielded 162 articles. We searched the abstracts for articles that used prospective data collection of clinical variables and compared them with a chest radiograph reference standard. Two categories of articles were identified: (1) those that evaluated the clinical examination for its ability to identify patients with community-acquired pneumonia, and (2) those that used the clinical examination to establish a prognosis for individuals with community-acquired pneumonia. To explore the possibility that we might be missing articles, we entered the original Rational Clinical Examination article into Citation Index (ISI Web of Knowl-

edge, Science Citation Index Expanded) to capture any articles not identified in the MEDLINE search.

A systematic review of the diagnosis of community-acquired pneumonia updated the previous Rational Clinical Examination article for new information published up to December 2000.¹ For clinical diagnosis, we used that systematic review and articles identified in the literature search published since 2000. We did not update the information on pneumonia care guidelines and risk stratification. We identified only 1 new article assessing the clinical examination for clinical diagnosis that was not cited in the systematic review.

NEW FINDINGS

- Approximately 1 of every 10 patients who are sick enough to be admitted with a clinical diagnosis of community-acquired pneumonia, but who have a normal initial chest radiograph result, will develop radiographic evidence of pneumonia by 72 hours.
- Among patients admitted to the hospital, an acute onset of illness is more likely observed in pneumonia caused by pyogenic bacteria (streptococcal, staphylococcal, or Enterobacteriaceae), but the clinical examination alone should not be used to select a patient’s antibiotic coverage.

IMPROVEMENTS IN THE DATA PRESENTED IN THE ORIGINAL PUBLICATION

No new data change the results for the 3 predictive models for community-acquired pneumonia, displayed in Table 40-4 of the original publication. The predictive models are redisplayed in Tables 40-5 and 40-6 to show the LR or probability of pneumonia associated with each.

CHANGES IN THE REFERENCE STANDARD

The pragmatic reference standard for usual clinical care is the chest radiograph. However, some patients will initially have a normal chest radiograph result in the early course of their illness. One study compared patients admitted with a clinical diagnosis of pneumonia and an abnormal radiograph result vs those with the clinical diagnosis and no radiographic

pneumonia shown by the initial radiograph result.² By 72 hours, a random sample of those admitted with a normal radiograph result showed that 7% (95% confidence interval [CI], 3%-13%) had developed pneumonia. High-resolution chest computed tomography (CT) picks up opacification or consolidation not observed on conventional chest radiographs. However, the chest CT has not been as extensively validated with microbiologic results as chest radiographs.¹

RESULTS OF LITERATURE REVIEW

A study with careful microbiologic characterization of community-acquired pneumonia for patients admitted to the hospital (75% of patients with pneumonia in the patient sample) showed that the finding of “acute” onset was the only symptom with a statistically significant diagnostic odds ratio (31) for

pneumonia caused by pyogenic organisms.³ When the patient has acute onset, the positive likelihood ratio (LR) is 3.6; when the onset of the patient’s illness is not acute, the negative LR is 0.31 and makes infection caused by atypical bacteria or viral illness more likely. However, because current guidelines for the treatment of community-acquired pneumonia in adults always include antibiotics for pyogenic bacteria, the results of the clinical examination should not be used to select the antibiotic.

In the model of Diehr et al,⁴ the score is calculated based on the clinical findings and the LR depends on the threshold that you want to consider positive. At scores of 1 or higher, the likelihood of pneumonia increases (Table 40-5). The Singal et al⁵ model is a logistic function (Table 40-5). Once the findings are recorded and the score calculated, the probability of adult pneumonia can be derived. As the number of findings increases for the Heckerling et al⁶ model, the probability of disease increases (Table 40-6); access to a nomogram is required, which makes this less practical to use. Nonetheless, the findings in all the models overlap and the physician can appropriately deduce that increased numbers of findings in these models make pneumonia more likely.

Table 40-5 Multivariate Findings for Adult Pneumonia

Diehr et al ⁴	
Add points for the presence of findings as follows:	
rhinorrhea = -2 points	
sore throat = -1	
night sweats = 1	
myalgias = 1	
sputum all day = 1	
respiratory rate > 25/min = 2	
temperature ≥ 37.8°C (100°F) = 2	
Threshold Score	LR
≥3	14
≥1	5.0
≥-1	1.5
<-1	0.22
Singal et al ⁵	
Score = $-3.095 + 1.214 \times (\text{cough}) + 1.007 \times (\text{fever}) + 0.823 \times (\text{crackles})$	
Each variable is coded as 1 if present, 0 if absent	
Probability of pneumonia = $1/(1 + e^{\text{score}})$	
Abbreviation: LR, likelihood ratio.	

Table 40-6 As the Number of Findings Increases, the Probability of Pneumonia Increases

Heckerling et al ⁶	
Count the number of findings present: absence of asthma; temperature ≥ 37.8°C (100°F); heart rate > 100/min; decreased breath sounds; crackles	
Findings Present	Probability, % (Baseline Prevalence 5%)
5	50
4	25
3	20
2	3
1	1
0	<1

EVIDENCE FROM GUIDELINES

No new federal agency recommendations address the diagnosis of community-acquired pneumonia. The role of the Pneumonia Severity Index for prognostication has been summarized and supported for treatment decisions.⁷

CLINICAL SCENARIO—RESOLUTION

The patient’s presenting symptoms and signs require clinical judgment to decide whether to obtain a chest radiograph. The new onset of wheezing might complicate your decision making because wheezing does not enter any of the predictive models. He has a cough with fever, so that the Singal et al⁵ model gives him a 29% probability of pneumonia. However, the Heckerling et al⁶ model shows the presence of only 2 findings that give him a probability of only 3% (absence of asthma and fever). Your judgment should consider the epidemiology of acute cough illness in your community (eg, are you in the middle of influenza season?), the clinical gestalt of how ill he appears, and the ability for him to return to consult you should he suddenly worsen. He does not appear ill enough for hospital admission, and he should do well with outpatient management. You think influenza is the most likely diagnosis, but the episode of chills concerns you. If you wish to treat this patient with antibiotics for pneumonia, a chest radiograph is required.

COMMUNITY-ACQUIRED PNEUMONIA, ADULT—MAKE THE DIAGNOSIS

PRIOR PROBABILITY

Using cough as a requirement for considering pneumonia, the baseline probability of radiographic-proven pneumonia in patients with acute cough illness is about 5%.

POPULATION FOR WHOM COMMUNITY-ACQUIRED PNEUMONIA SHOULD BE CONSIDERED

- Patients with symptoms of acute respiratory illness, primarily cough.
- Patients with comorbid illnesses, older patients, and those with immunocompromised status have a much higher risk for community-acquired pneumonia.

DETECTING THE LIKELIHOOD OF COMMUNITY-ACQUIRED PNEUMONIA

Individual clinical symptoms or signs have low utility for identifying patients with pneumonia. Combinations of findings are required, including cough, fever, tachypnea, and abnormalities on auscultation (decreased breath sounds or crackles). The clinical decision that a patient has a low enough likelihood of pneumonia that a chest radiograph is not required lowers the probability of pneumonia to less than 5%. Rather than recommending one particular

prediction model over the other for selecting patients who should have a chest radiograph, clinicians should use their own clinical judgment and the presence of increasing numbers of clinical signs and symptoms from the prediction models. The detection of pneumonia requires a chest radiograph, and the presence of appropriate findings on the chest radiographs is part of the case definition for pneumonia.

REFERENCE STANDARD TESTS

There is no practical reference standard test that allows correct categorization of the patient who has a pulmonary infection that will respond to antibiotics vs those that do not need antibiotics. The reference standard for pneumonia is the identification of a microbiologic pathogen from lung tissue. Because this is infrequently obtained, the pragmatic reference standard is the combination of clinical findings with appropriate abnormalities on a chest radiograph. A follow-up chest radiograph is often required to demonstrate improvement of the initial findings consistent with pneumonia or to identify findings not present on the first radiograph. The role of high-resolution CT for patients with a nondiagnostic initial chest radiograph result requires studies comparing the results to microbiologic outcomes.

REFERENCES FOR THE UPDATE

1. Metlay JP, Fine MJ. Testing strategies in the initial management of patients with community-acquired pneumonia. *Ann Intern Med.* 2003; 138(2):109-118.^a
2. Basi SK, Marrie TJ, Huang JQ, Majumdar SR. Patients admitted to hospital with suspected pneumonia and normal chest radiographs: epidemiology, microbiology, and outcomes. *Am J Med.* 2004;117(5):305-311.
3. Ruiz-González A, Falguera M, Vives M, Nogués A, Porcel JM, Rubio-Caballero M. Community-acquired pneumonia: development of a bedside predictive model and scoring system to identify the aetiology. *Respir Med.* 2000;94(5):505-510.^a
4. Diehr P, Wood RW, Bushyhead J, Krueger L, Wolcott B, Tompkins RK. Prediction of pneumonia in outpatients with acute cough: a statistical approach. *J Chronic Dis.* 1984;37(3):215-225.
5. Singal BM, Hedges JR, Radack KL. Decision rules and clinical prediction of pneumonia: evaluation of low-yield criteria. *Ann Emerg Med.* 1998; 18(1):13-20.
6. Heckerling PS, Tape TG, Wigton RS, et al. Clinical prediction rule for pulmonary infiltrates. *Ann Intern Med.* 1990;113(9):664-670.
7. Stanton MW. *Improving Treatment Decisions for Patients With Community-Acquired Pneumonia.* Rockville, MD: Agency for Healthcare Research and Quality; 2002. Research in Action Issue 7. AHRQ Publication No. 02-0033. <http://www.ahcpr.gov/clinic/pneumonia/pneumonia.pdf>. Accessed June 8, 2008.

^aFor the Evidence to Support the Update for this topic,
see <http://www.JAMAEvidence.com>.

This page intentionally left blank

EVIDENCE TO SUPPORT THE UPDATE:

40

Community-Acquired Pneumonia, Adult

TITLE Testing Strategies in the Initial Management of Patients With Community-Acquired Pneumonia.

AUTHORS Metlay JP, Fine MJ.

CITATION *Ann Intern Med.* 2003;138(2):109-118.

QUESTIONS Do clinical findings allow the physician to establish the diagnosis of community-acquired pneumonia? Once a chest radiograph confirms community-acquired pneumonia, do clinical and laboratory results allow identification of individuals who can be safely treated as outpatients?

DESIGN Formal systematic review.

DATA SOURCES MEDLINE database.

STUDY SELECTION AND ASSESSMENT

The authors used the same selection criteria as that in the original Rational Clinical Examination article on adult pneumonia: published from January 1996 to December 2000, English language, but excluding studies of children or inpatients. Studies had to report the reference standard (chest radiography) for all patients suspected of having community-acquired pneumonia. The references from retrieved articles were reviewed.

For studies of short-term prognosis and the need for hospitalization, the authors updated a previous meta-analysis.¹ The search strategy excluded nosocomial, nursing home-acquired, noninfectious, pediatric, and human immunodeficiency virus-associated pneumonia.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

A chest radiograph was required as the reference standard.

OUTCOME MEASURES

Sensitivity, specificity, and likelihood ratios (LR) of clinical findings for pneumonia. Ranges for the LR were reported rather than summary measures.

MAIN RESULTS

The authors found no additional clinical diagnosis of pneumonia studies not referenced in the original Rational Clinical Examination article. The authors found 134 cohort studies of the short-term prognosis of community-acquired pneumonia.

Once pneumonia is established (by clinical examination and chest radiograph), the results of the medical history, physical examination, and laboratory tests are used to establish a prognosis. The Pneumonia Patient Outcomes Research Team Severity Index (PSI) accurately identified low-risk patients who could be treated as outpatients for community-acquired pneumonia.² Although the PSI allows risk stratification for all patients with pneumonia, its primary purpose was the accurate identification of low-risk patients. For those not in the lowest risk class, the clinical history, physical examination, and laboratory findings should be used to establish a risk class. The following clinical variables predict a poorer prognosis, so they should be the focus of the evaluation²:

Demographic variables:

1. Age
2. Male
3. Nursing home residence

Comorbid illness:

1. Neoplastic disease
2. Liver disease
3. Congestive heart failure
4. Cerebrovascular disease
5. Renal disease

Physical examination findings:

1. Altered mental status
2. Respiratory rate greater than or equal to 30/min
3. Systolic blood pressure less than 90 mm Hg
4. Temperature less than 35°C or greater than or equal to 40°C
5. Pulse greater than or equal to 125/min

In addition to these variables, several common laboratory tests further modify the clinical variables.

Laboratory findings:

1. pH less than 7.35
2. Blood urea nitrogen level greater than 64.5 mg/dL

3. Sodium level less than 130 mEq/L
4. Glucose level greater than 250 mg/dL
5. Hematocrit level less than 30%
6. PO₂ less than 60 mm Hg (or oxygen saturation < 90%)
7. Pleural effusion

CONCLUSIONS

LEVEL OF EVIDENCE Systematic review.

STRENGTHS Formal systematic review that used the same methods as the original article in The Rational Clinical Examination series on adult pneumonia.

LIMITATIONS None.

As of 2001, no additional data on the clinical findings for the diagnosis of adult pneumonia were identified.

The PSI was selected as a validated prognostic model with the highest methodologic criteria.³

Reviewed by David L. Simel, MD, MHS

REFERENCES FOR THE EVIDENCE

1. Fine MJ, Smith MA, Carson CA, et al. Prognosis and outcomes of patients with community-acquired pneumonia: a meta-analysis. *JAMA*. 1996;275(2):134-141.
2. Fine MJ, Auble TE, Yealy DM, et al. A prediction rule to identify low-risk patients with community-acquired pneumonia. *N Engl J Med*. 1997;336(4):243-250.
3. Auble TE, Realy DM, Fine MJ. Assessing prognosis and selecting an initial site of care for adults with community-acquired pneumonia. *Infect Dis Clin North Am*. 1998;12(3):741-759.

TITLE Community-Acquired Pneumonia: Development of a Bedside Predictive Model and Scoring System to Identify the Aetiology.

AUTHORS Ruiz-González A, Falguera M, Vives M, Nogués A, Porcel JM, Rubio-Caballero M.

CITATION *Resp Med*. 2000;94(5):505-510.

QUESTION Among patients admitted to the hospital with a clinical diagnosis of community-acquired pneumonia, do clinical features identify those likely to have bacterial pneumonia from pyogenic organisms?

DESIGN Prospective, consecutive enrollment with clinical data recorded before microbiologic results obtained.

SETTING University hospital in Spain.

PATIENTS Patients older than 14 years, admitted during a 15-month period, with a principal diagnosis of pneumonia. Patients admitted within the previous 7 days or transferred to the hospital were excluded. Of the potentially eligible patients with pneumonia treated in the emergency department, 75% were admitted to the hospital.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

Clinical data were recorded prospectively by the examining clinician. The criteria for inpatient vs outpatient treatment were not specified. The microbiologic reference standard was used to classify patients with bacterial pneumonia caused by *Streptococcus pneumoniae* and other pyogenic bacteria (streptococci, *Haemophilus influenzae*, *Staphylococcus aureus*, Enterobacteriaceae) vs atypical pneumonia (*Mycoplasma pneumoniae*, *Chlamydia pneumoniae*, *Chlamydophila psittaci*, *Coxiella burnetii*, *Legionella pneumophila*, or virus) vs pneumonia of unknown etiology. The microbiologic diagnosis was based on blood culture results, microbiologic results, or polymerase chain reaction test results on samples from trans-thoracic needle aspiration of the lung in patients without contraindications to the procedure, sputum culture for legionella, or 4-fold titer increase for atypical organism or virus.

MAIN OUTCOME MEASURES

A total of 85 patients (82%) of the sample had an etiologic diagnosis with the microbiologic standards. A logistic model was created to see which clinical variables predicted pyogenic bacterial pneumonia (see Tables 40-7 and 40-8).

MAIN RESULTS

The diagnostic odds ratio (OR) for acute onset (OR, 31; 95% confidence interval [CI], 6-150) and age greater than 65 years

Table 40-7 Likelihood Ratios for Pyogenic Bacterial Pneumonia

Test	LR+ (95% CI)	LR- (95% CI)
Age > 65 y or comorbidity ^a	2.7 (1.6-4.6)	0.43 (0.26-0.71)
Acute onset	3.6 (2.0-6.5)	0.31 (0.17-0.55)
Chills	1.60 (0.73-3.40)	0.86 (0.67-1.10)
Pleuritic chest pain	1.40 (0.97-2.00)	0.62 (0.36-1.10)
Purulent sputum	1.20 (0.56-2.50)	0.95 (0.74-1.20)
Signs of consolidation on auscultation ^a	1.10 (0.79-1.50)	0.86 (0.48-1.60)
Leukocytosis or leukopenia ^b	2.0 (1.3-2.8)	0.32 (0.16-0.66)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

^aThe specific comorbidities or signs of consolidation were not described.

^bLeukocytosis defined as white blood cell count $\geq 11\,000/\mu\text{L}$ and leukopenia defined as white blood cell count $\leq 4000/\mu\text{L}$.

or comorbidity (OR, 6.9; 95% CI, 2-23) were the only findings that did not include 1 in the OR CI.

CONCLUSIONS

LEVEL OF EVIDENCE Level 3.

STRENGTHS Microbiologic proof of infection in most patients, including results from lung parenchyma samples.

LIMITATIONS Study includes only patients admitted to the hospital. Radiographic results not provided. Patient population not well described in terms of comorbid illness.

Table 40-8 Likelihood Ratio of Bacterial Pneumonia From a Scoring System

Test	Sensitivity (95% CI)	Specificity (95% CI)	LR+ ^a	LR- ^a
Bacterial pneumonia score $\geq 5^b$	0.89 (0.78-0.96)	0.63 (0.54-0.81)	2.4	0.17

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

^aThe LRs are estimated from the sensitivity and specificity. CIs cannot be calculated without the raw values.

^bAge > 65 years or comorbidity = 3 points; acute onset = 5 points; leukocytosis or leukopenia = 2 points.

Among admitted patients with community-acquired pneumonia, an acute onset of disease is the variable that most increases the likelihood of bacterial pneumonia attributed to pyogenic bacteria. An onset that is not acute decreases the likelihood of pyogenic bacterial pneumonia the most. The lack of significance (diagnostic OR not statistically different from 1) for chills, pleurisy, purulent sputum, and auscultatory signs of consolidation is also important.

We do not know whether the results can be applied to the 25% of patients receiving ambulatory treatment, because those patients did not have the same microbiologic studies.

Reviewed by David L. Simel, MD, MHS

This page intentionally left blank

CHAPTER 41

Does This Infant Have Pneumonia?

Peter Margolis, MD, PhD

Anne Gadomski, MD, MPH

CLINICAL SCENARIO

A mother brings her 8-month-old infant to your office in midwinter with a cough. She reports that the illness began 4 days ago with a runny nose. Two days ago, the baby developed a fever. Now the baby's symptoms are getting worse. The baby has become more irritable, is eating less, and seems to be having more difficulty breathing. This is the third child you have treated today with a cough. While the first two children were treated for acute upper respiratory tract infections, you wonder if the findings in this infant suggest pneumonia.

WHY IS THIS AN IMPORTANT QUESTION TO ANSWER WITH A CLINICAL EXAMINATION?

Acute respiratory illnesses are among the most common conditions of infants treated in primary care offices. Although the majority of respiratory illnesses involve infections of the upper respiratory tract, most infants will experience a lower respiratory tract illness (LRI) in the first year of life. Of those with LRIs, about 30% visit a physician,^{1,2} and about 2% are hospitalized.³

LRIs can be defined simply as infections at an anatomic level below the vocal cords. The majority of LRIs in infants are caused by viruses; only a small proportion is due to bacteria. The differential diagnosis for cough is long ([Table 41-1](#)).

Therapies are available to treat a variety of manifestations of lower respiratory tract disease, so it is important to diagnose these complaints accurately and estimate their severity to deliver the appropriate treatment. Identifying infants at lower risk of bacterial disease may help clinicians avoid the unnecessary use of antibiotics, which may reduce the risk of subsequent bacterial infection and slow the emergence of resistant strains of bacteria within the population.⁴ Greater certainty about the presence of a viral LRI may also help clinicians avoid additional testing such as radiography or blood culture.

This overview focuses on the medical history and physical examination findings of infants that distinguish pneumonia from other LRIs.

METHODS

We conducted a MEDLINE search from 1982 to 1995 to identify articles about the diagnosis of pneumonia in children. We searched for articles with any of the following Medical Subject Heading terms: "pneumonia," "diagnostic tests," "sensitivity and specificity," "reproducibility of results," "physical examination," or "medical history taking." This search was further limited to studies published in English about humans and that involved children. This search strategy identified 38 articles. Four more articles were identified by reviewing a compendium of references prepared by the World Health Organization.⁵ Etiologic studies, which did not include a chest

Table 41-1 Differential Diagnosis of Cough in Infants

Anatomic
Foreign body
Congenital malformation (eg, vascular ring, cystic adenomatous malformation, bronchogenic cyst, tracheomalacia)
Inflammatory
Reactive airway disease
Infectious
Viral
Croup
Laryngotracheobronchitis
Bronchitis
Viral pneumonia
Bacterial
Epiglottis
Tracheitis
Bronchitis
Bacterial pneumonia
Chlamydia
Tuberculosis
Other
Cystic fibrosis
Congestive heart failure
Gastroesophageal reflux

radiograph examination as part of the gold standard, involved only inpatients. Studies of illness in families' homes, rather than in clinical settings, were excluded ($n = 29$).

All the articles were reviewed by the authors, and disagreements were resolved by discussion. We used the methods developed for this series to assess the quality of the articles. The highest-quality studies are emphasized in the "Results" section. We did not aggregate statistically the results of the studies because of differences in the ages of the study samples and differences in cutoff points of key variables, such as respiratory rate. Confidence intervals (CIs) were calculated according to the method suggested by Koopman⁶ and Centor and Keightley.⁷

Reference Standard for Diagnosing Pneumonia

The reference standard for diagnosing pneumonia is an aspirate from the lower respiratory tract obtained by bronchoalveolar lavage or lung puncture. The use of bronchial lavage is appropriate in guiding antibiotic choice in patients with refractory or complicated pneumonia. In general practice, chest radiographs are readily obtained and can be considered a pragmatic reference standard for pneumonia.

A number of studies evaluated the accuracy of the chest radiograph in differentiating viral from bacterial disease in children.⁸⁻¹³ It is difficult to determine the accuracy of the chest radiograph from these studies because of methodologic limitations, as well as problem with study design introduced by the biology of pneumonia. It is not possible to obtain cultures from a lung in most patients. Therefore, investigators

have had to use combinations of other clinical features as a proxy for bacterial pneumonia. Reliance on less than perfect gold standards for diagnosing bacterial pneumonia may produce over- or underestimates of the association of a positive chest radiographic finding with bacterial pneumonia. Two studies used the same definition of bacterial pneumonia (duration of symptoms < 2 days, temperature > 39.5°C, total white blood cell count > 15 000/ μ L).^{8,9} Both found the sensitivity of the chest radiograph for diagnosing bacterial vs viral pneumonia to be approximately 75%. However, one reported a specificity of 100%; the other, a specificity of 63%. The reported sensitivity for studies with varying definitions ranges from 42% to 80% and the specificity from 42% to 100%. Studies of the accuracy of chest radiographs have also been compromised by other methodologic problems, such as interobserver variability in the interpretation of the radiograph, oversampling patients with relatively severe disease, and the relatively small numbers of patients with bacterial pneumonia. Such problems make estimates of chest radiographic accuracy unreliable.

Variation in the biologic manifestations of bacterial pneumonia also presents challenges in the interpretation of published studies. For example, bacterial pneumonia is classically associated with lobar consolidation on the radiograph. However, studies report that bacterial pneumonia may be associated with infiltrates that are lobar, perihilar, segmental, interstitial, or nodular infiltrates.¹⁴⁻¹⁶ Consolidation can also be observed with viral pneumonia, but it is unclear whether this radiologic appearance is due to segmental consolidation, atelectasis, or bacterial coinfection. Such variability in the radiographic appearance of bacterial pneumonia may produce over- or underestimates of the association of a positive chest radiographic finding with bacterial pneumonia.

Clinicians should be aware that the chest radiographic results may be negative in patients with early bacterial pneumonia.¹⁷ The sensitivity of the chest radiograph will be reduced in this group. The implications of this observation are important for studies of the clinical examination. For the purposes of this systematic review, we included studies that used the chest radiograph as the reference standard. Studies that combined the clinical diagnosis with the chest radiographic results as the reference standard were excluded because inclusion of the diagnostic test in the reference standard may overestimate the accuracy of clinical findings. The significance of clinical findings of pneumonia in the absence of a positive chest radiographic finding remains to be studied.

Normal Anatomy and Pathophysiology of Pneumonia

Lower respiratory tract infections occur at or below the larynx and include epiglottitis, laryngitis, laryngotracheobronchitis (croup), bronchiolitis, and pneumonia (Figure 41-1). Pneumonia typically follows an upper respiratory tract illness in which the lower respiratory tract is invaded by bacteria, viruses, or other pathogens that trigger the immune response and produce inflammation. Histamines, leukotrienes, and chemotactic factors are released that recruit white blood cells to the area. This response fills the air spaces

of the lower respiratory tract with white blood cells, fluid, and cellular debris. This process reduces lung compliance, increases resistance, obstructs smaller airways, and possibly results in collapse of distal air spaces.

The resultant physical findings vary with the site of infection, ranging from coarse breath sounds or rhonchi in bronchopneumonia to crackles in the alveoli in cases of pneumonia or bronchiolitis. Crackles are the result of the explosive equalization of gas pressure between the terminal bronchiole and the alveoli.¹⁸ Wheezes result from the oscillation of air through a narrowed airway that produces a musical sound likened to a vibrating reed.¹⁹ Decreased breath sounds may also be observed in areas of consolidation.

How to Elicit the Relevant Symptoms and Signs

The physician's first goal when taking the medical history and performing the physical examination in a child who presents with a cough is identification of the clinical syndrome and level of involvement, as shown in Figure 41-1. The second goal is to estimate the severity of the illness. The physician should ask the parent about symptoms associated with pneumonia, as well as those that may discriminate pneumonia from other lower respiratory tract diseases. In addition to cough, symptoms that may increase the likelihood of pneumonia include trouble breathing, rattling in the chest, noisy breathing, trouble feeding, fever, rapid breathing, anxiety, or restlessness. Clinicians working in different regions or with different cultures need to familiarize themselves with local terminology for lower respiratory tract symptoms. It may also be useful to ask about previous episodes of these chest symptoms because recurrent bouts of pneumonia or bronchitis may suggest reactive airway disease. In early infancy (<2 months), infants of mothers who had chlamydia during pregnancy may develop afebrile pneumonia. Infants only rarely produce sputum. In older infants, foreign body ingestion and salicylate poisoning should be considered. Although clinical experience suggests that the history of pneumonia may be of acute or gradual onset and that bacterial pneumonia tends to be associated with fever, we were unable to find any studies substantiating these observations.

The physical examination should include an assessment of the child's general appearance, measurement of the respiratory rate, evaluation of the work of breathing, and auscultation of the chest. The child's general appearance may provide important clues about the presence of bacterial illness and its severity. Infants can exhibit a wide range of behaviors and mood changes during the parental interview, while being undressed, and during the physical examination. Therefore, it is important to take a nonthreatening approach with the young child. Infants should be observed initially at a distance, while they are comfortable, usually in the caretaker's lap. The assessment of general appearance should include an evaluation of a number of factors: attentiveness to the environment, ability to breast-feed or drink, ability to sustain sucking, vocalization, smiling, movement, color, and consolability. If there is uncertainty about particular findings, it may be helpful to try to elicit them; for example, encouraging the child to smile, having the mother offer the breast, or showing the child a toy to engage his or her attention.

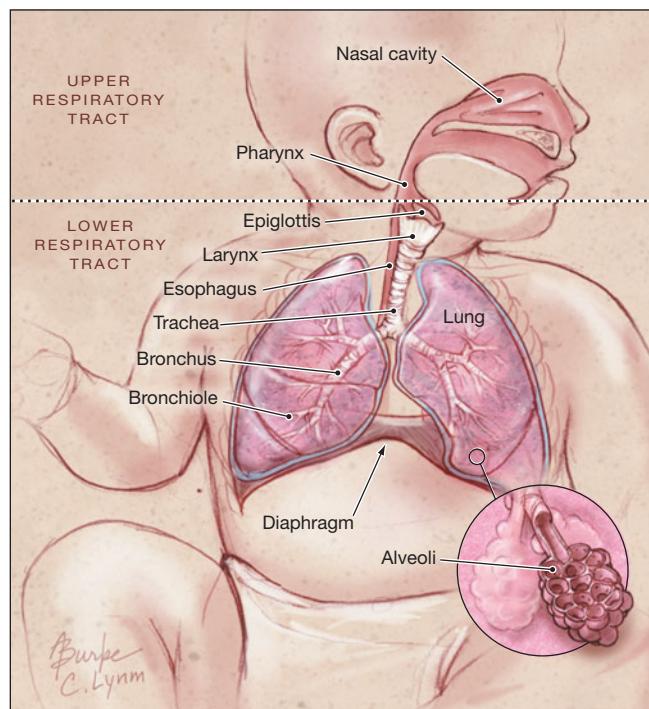


Figure 41-1 Clinical Syndromes of Acute Respiratory Infections

Acute upper respiratory tract infections include cold, otitis media, and pharyngitis, which are all located above the dashed line in the figure. Acute lower respiratory tract infections causing stridor include epiglottitis, laryngitis, and laryngotracheitis. Anatomically, lower respiratory tract infections include bronchitis, bronchiolitis, and pneumonia.

Respiratory rates change considerably in the first year of life, decreasing from a mean in awake babies of about 50/min at 1 week of age to about 40/min at 6 months of age.²⁰⁻²³ The respiratory rate in children can also vary during brief intervals as the child's level of interest in the environment changes or while the child is asleep or feeding.²⁴ Polygraphic studies of infants younger than 6 months have demonstrated that mean respiratory rates were 4/min to 13/min higher in active sleep (rapid eye movement) than in quiet sleep.^{23,25} Fever can also increase an infant's respiratory rate by 10/min per degree centigrade in children without pneumonia.²⁶ However, the effects of fever in the presence of pneumonia have not been studied.

The respiratory rate is best measured by observing chest wall movements during 1 minute.²⁷⁻²⁹ Listening to the chest with a stethoscope may stimulate the child and lead to a falsely elevated measurement. Measurement errors in counting the respiratory rate are greater when children are agitated or crying compared with when they are calm, feeding, or sleeping. The examiner should count the respiratory rate before conducting other parts of the examination. Respiratory rate cutoffs that are commonly used to indicate an elevated rate are greater than 60/min in infants younger than 2 months, greater than 50/min in infants 2 through 12 months of age, and greater than 40/min in children older than 12 months.³⁰

Assessing an infant's work of breathing is important to estimate the severity of LRI. This assessment includes evaluation of chest wall movements, nasal flaring, and grunting.

Chest wall movements include retractions or chest indrawing, best observed with the chest fully exposed. SuprACLAVICULAR retractions may be observed as indrawing of the soft tissue above the clavicle or above the sternal notch. Intercostal retractions are seen as indrawing of skin between the ribs. Subcostal retractions occur on or just below the costal margin. Many experts suggest that these types of retractions, involving only the soft tissue, should be distinguished from chest wall indrawing, defined as an inward movement of the lower chest wall (ie, ribs) when the child breathes in. Chest indrawing is more likely to be observed in infants younger than 18 to 24 months, whose chest walls are more pliant. The finding may be appreciated best by viewing the chest laterally and looking for indrawing of the ribs or lower sternum with inspiration, relative to a fixed point beyond the child's chest that is set as a mental reference point (Figure 41-2). Normally, the costal margin moves little during quiet breathing. If it does, it moves up and outward because the normal diaphragm lifts the costal margin outward. In disease states, the depressed diaphragm may apply an inward traction on the chest, resulting in paradoxical movement of the chest wall during inspiration.³¹ Therefore, in airway obstruction, the costal margin tends to move paradoxically (ie, down and inward). Sometimes, the abdomen moves outward while the chest moves inward during inspiration. This has also been called Hoover sign³² or paradoxical or seesaw breathing.

Nasal flaring is enlargement of both openings of the nose during inspiration. It is due to constriction of anterior and posterior dilators naris muscles. Grunting is a repetitive short upper respiratory tract sound produced by partial vocal cord closure during expiration.³³ Grunting slows expiratory flow and increases lung volume and alveolar pressures.³⁴ It can be a sign of severe disease and suggests impending respiratory failure. Examiners should be aware that the presence of signs of increased work of breathing may change with the state of the child. For example, chest wall indrawing may be present only when the child is awake or more active.

Adventitious sounds that can be appreciated on auscultation include discontinuous or popping sounds, sounds that occur throughout the inspiratory or expiratory phase, or continuous sounds.³⁵ Discontinuous sounds have been called *crackles*, *rales*, or *crepitations*. They typically occur at the end of inspiration. Continuous sounds include wheezes and rhonchi and can be musical, high or low pitched, inspiratory

or expiratory, short or long, or monophonic or polyphonic.¹⁸ Clinicians should try to distinguish whether sounds are continuous or discontinuous before applying a name. Many clinicians differentiate continuous sounds that are whistling or high pitched (usually called wheezes) from low-pitched, snoring, or rattling sounds (usually called rhonchi). Many experts consider wheezes to reflect small airway obstruction (ie, bronchioles), whereas rhonchi reflect obstruction of the large airways (ie, bronchi).

The language used to describe auscultatory findings can be a source of confusion. For example, *rhonchi* and *rales* are, respectively, the Latin and French words for crackles. Indeed, Laënnec (the inventor of the stethoscope) distinguished 6 types of crackles.³⁶ He believed that only 1 of these was associated with pneumonia.

Auscultation of the chest is often more difficult in infants when they are crying. For this reason, it should be performed after the visual inspection of the child. It is important to listen to the front, back, and sides of the infant's chest because adventitious sounds may only be heard in one location. Even when the infant is crying, adventitious sounds may be heard at the end of inspiration when the infant is quiet and about to take a breath. Examiners should also be aware that wheezes can often be appreciated by listening to the sounds of breaths from infants' mouths (audible wheezing). Finally, infants may have several types of adventitious sounds present (although this is more common in reactive airway disease or viral LRI). Textbooks do not recommend percussion of the chest in infants because it is difficult to get infants to cooperate with this maneuver.

Are These Symptoms or Signs Ever Normal?

Premature infants and neonates may appear to have chest indrawing during normal breathing or exertion. Grunting and groaning noises occur from time to time in normal healthy infants. An infant who is playful may demonstrate increased respiratory rate, intercostal retractions, and increased work of breathing.

RESULTS

The Precision of Symptoms and Signs

A total of 56 patients with lower respiratory tract symptoms were examined by pairs of general pediatricians from a group that included academic pediatric generalists, pediatric residents, and pediatricians in community practice.³⁷ Agreement was good for most signs on physical examination that could be observed by inspection, including the social interaction markers of attentiveness (κ , 0.49), smiling (κ , 0.51), quality of cry (κ , 0.63), physical appearance and movement (κ , 0.54), color (κ , 0.66), respiratory effort retractions (κ , 0.48), and use of accessory muscles (κ , 0.59). There was only fair agreement about most auscultatory findings: prolonged expiratory phase (κ , 0.22), adventitious sounds (κ , 0.3), and inspiratory wheezing (κ , 0.29). Agreement was good for audible wheezing (κ , 0.7) and for expiratory wheezing (κ , 0.63). In

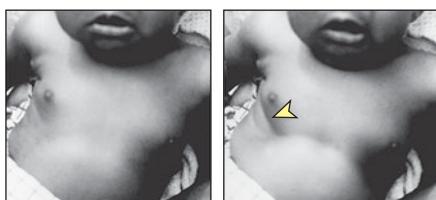


Figure 41-2 Chest Indrawing

Inward movement of the lower chest wall (ie, ribs) when the child breathes in. Chest indrawing is best appreciated by viewing the chest laterally and looking for indrawing of the ribs or lower sternum with inspiration. Reproduced with permission from the World Health Organization.⁵

general, physicians agreed more often that a finding was present than when it was absent. A second study indicated that observers are less likely to agree about the severity of findings than about their presence or absence.³⁸

Several studies of the precision of the respiratory rate suggest that respiratory rates counted over 30 seconds average 2/min to 4/min faster than respiratory rates counted during 60 seconds.²⁹ Counting the respiratory rate over 30 seconds will lead to more abnormal rates and may spuriously increase the number of children diagnosed as having pneumonia. More accurate results are obtained if the average of two 30-second counts is taken or one 60-second count is taken.

Observer agreement is good for most signs on the physical examination. There is better agreement about signs that can be observed than signs that require auscultation of the chest.

The Accuracy of Signs of Pneumonia

The reported accuracy of clinical findings varies considerably among studies because of methodologic limitations and differences in the spectrum of illness severity among sites in which the studies were conducted. In most reports, chest radiographs were used as the gold standard and children who had clinical findings suggestive of pneumonia were more likely to have had a radiographic examination than those who did not (Table 41-2). Although this approach makes sense clinically, it introduces verification bias that tends to overestimate a test's sensitivity and underestimate its specificity.⁴⁷

Two studies, both of which were conducted in developing countries, attempted to overcome the problem of selective ordering of the gold standard by obtaining chest radiographs on all children with abnormal clinical findings (eg, elevated respiratory rate), as well as a sample of children without abnormal findings.³⁹⁻⁴¹ The reported accuracy was then adjusted statistically for the fraction of patients sampled in each group. These 2 studies found that there was no single sign that could be used to rule in or rule out pneumonia definitively. In these studies, children with elevated respiratory rates were about twice as likely to have pneumonia (positive likelihood ratio [LR+], 1.5-2.1) as children without elevated respiratory rates (Table 41-3). Conversely, those without elevated respiratory rates were only about 0.36 to 0.5 times as likely to have pneumonia. These studies also found that the presence of chest indrawing (retractions) increased the likelihood of pneumonia (LR+, 2.4-2.5). However, normal chest movements did not rule out pneumonia (negative likelihood ratio [LR-], 0.7-0.78). Other useful findings that increased the likelihood of pneumonia included nasal flaring (LR+, 3.0) and crepitations (LR+, 3.5). The absence of nasal flaring (LR, 0.71) and crepitations (LR, 0.69) did not effectively lower the likelihood of pneumonia. Other studies in developing countries, even though less methodologically sound, found the accuracy of clinical signs to be more or less in the same range as that found in the 2 more well-designed investigations (Table 41-3).^{43,48-50}

The lower prevalence of bacterial disease and severe pneumonia found in developed countries^{51,52} might suggest that

Table 41-2 Characteristics of Studies Included in Systematic Review

Source, y	Country, Setting	Quality Level ^a	Inclusion and Exclusion Criteria	Age Range	Pneumonia Prevalence, % (No./Total)	Definition of Pneumonia
Redd et al, ^{39,40} 1994	Lesotho outpatient department	1	Children with cough, upper respiratory tract infection, trouble breathing, and ear pain	3 mo to 5 y	17 (65/382)	Parenchymal infiltrate on chest radiograph
Harari et al, ⁴¹ 1991	New Guinea outpatient department	1	Cough exclusion: wheeze, stridor, measles, and pertussis	8 wk to 6 y	30 (56/185)	Radiographic pneumonia
Crain et al, ⁴² 1991	US emergency department	1	Infants with temperature > 38°C	1 d to 2 mo	12 (27/228)	Positive chest radiographic examination result
Lozano et al, ⁴³ 1994	Columbia emergency department	4	Respiratory signs and symptoms, cough < 7 d	1 wk to 3 y	65 (130/200)	Radiographic pneumonia
Leventhal, ⁴⁴ 1982	US emergency department	4	Children with fever or respiratory symptoms for whom chest radiograph was ordered Excluded major chronic disease, asthma, croup, and trauma	3 mo to 15 y	19 (26/136)	Abnormal chest radiographic examination result
Taylor et al, ⁴⁵ 1995	US emergency department	4	Temperature > 38°C Excluded infants with chronic lung disease, bronchopulmonary dysplasia, wheezing, and stridor	1 d to 2 y	7.3 (42/572)	Positive chest radiographic examination result
Zukin et al, ⁴⁶ 1986	US emergency department	4	Children with chest radiographic examination as part of emergency department evaluations	1 d to 17 y	14 (18/125)	Radiographic pneumonia

^aSee Table 1-7 for a summary of Evidence Grades and Levels.

Table 41-3 Operating Characteristics of Selected Clinical Findings

Source, y	Item	LR+ (95% CI) ^a	LR- (95% CI) ^a
Description of Breathing (Time or Explanation)			
Redd et al, ^{39,40} 1994	Respiratory rate ≥ 50/min (3-11 mo)	1.9 (...)	0.36 (...)
Harari et al, ⁴¹ 1991	Respiratory rate ≥ 50/min	2.2 (...)	0.52 (...)
Crain et al, ⁴² 1991	Respiratory rate ≥ 60/min (<8 wk)	8.0 (5.3-12)	0.55 (0.4-0.8)
Lozano et al, ⁴³ 1994	Respiratory rate ≥ 50/min (0-11 mo)	1.7 (1.2-2.3)	0.52 (0.4-0.7)
Leventhal, ⁴⁴ 1992	Tachypnea (clinician judgment of fast breathing)	2.0 (1.5-2.7)	0.32 (0.1-0.7)
Taylor et al, ⁴⁵ 1995	Tachypnea (maximal sensitivity and specificity in different age strata)	3.2 (2.5-4.1)	0.34 (0.2-0.6)
Zukin et al, ⁴⁶ 1986	Tachypnea (≥standard deviation for age)	1.6 (0.9-2.6)	0.75 (0.5-1.2)
Work of Breathing			
Redd et al, ^{39,40} 1994	Chest indrawing	2.4 (...)	0.70 (...)
Harari et al, ⁴¹ 1991	Chest indrawing	2.5 (...)	0.78 (...)
Lozano et al, ⁴³ 1994	Chest indrawing	1.3 (1.0-1.5)	0.53 (0.3-0.9)
Crain et al, ⁴² 1991	Chest indrawing	26 (5.7-119)	0.75 (0.6-0.9)
Redd et al, ^{39,40} 1994	Nasal flaring (3-11 mo)	6.6 (...)	0.71 (...)
Lozano et al, ⁴³ 1994	Nasal flaring	1.2 (0.9-1.6)	0.83 (0.6-1.1)
Leventhal, ⁴⁴ 1992	Nasal flaring	1.9 (1.0-3.8)	0.79 (0.6-1.1)
Lozano et al, ⁴³ 1994	Grunting	1.2 (0.8-1.8)	0.89 (0.7-1.1)
Leventhal, ⁴⁴ 1992	Grunting	3.2 (1.1-9.2)	0.86 (0.7-1.0)
Temperature			
Harari et al, ⁴¹ 1991	>38°C	1.1 (...)	0.95 (...)
Zukin et al, ⁴⁶ 1986	Fever ^b	1.5 (1.3-1.7)	0.17 (0.02-1.1)
Auscultation			
Leventhal, ⁴⁴ 1992	Crepitations	2.1 (1.2-3.8)	0.73 (0.5-1.0)
Lozano et al, ⁴³ 1994	Crepitations	1.8 (1.4-2.3)	0.36 (0.2-0.5)
Crain et al, ⁴² 1991	Crepitations	15 (2.9-78)	0.86 (0.7-1.0)
Zukin et al, ⁴⁶ 1986	Crepitations	2.9 (1.4-3.7)	0.57 (0.3-0.97)
Lozano et al, ⁴³ 1994	Wheezes	0.63 (0.4-1.1)	1.12 (1.0-1.3)
Crain et al, ⁴² 1991	Wheezes	4.0 (0.4-37)	0.97 (0.9-1.1)
Zukin et al, ⁴⁶ 1986	Wheezes	0.19 (0.03-1.3)	1.30 (1.2-1.5)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

^aEllipses indicate CIs could not be calculated because insufficient information was reported.

^bFever defined as 2 SD for age.

the accuracy of physical examination signs would be lower than that reported in studies from developing countries. However, the few studies performed in developed countries reported results similar to those cited above.^{42,44,45} These studies may have overestimated the accuracy of clinical findings because chest radiographs were more likely to be obtained in patients with signs and symptoms of disease. In a study by Leventhal,⁴⁴ the absence of tachypnea, as observed by the clinician examining the patient, was useful for ruling out pneumonia (LR-, 0.32), whereas the presence of tachypnea somewhat increased the odds of pneumonia (LR+, 2.0). Grunting and crepitations were more useful in ruling in disease (LR+, 3.2 and 2.1, respectively). Their absence did not appreciably decrease the likelihood of disease (LR-, 0.86 and 0.73). The study by Taylor et al⁴⁵ reported a somewhat higher LR+ for tachypnea (LR+, 3.2), but this study included only

febrile children, and chest radiographs were not obtained for all study patients.

A study by Crain et al⁴² included only infants with fever and younger than 8 weeks who were treated in an emergency department. The authors reported that tachypnea (LR+, 8.0; 95% CI, 5.3-12) and chest indrawing (LR+, 26; 95% CI, 2.7-119) substantially increased the likelihood of pneumonia. Although these likelihood ratios are high, the number of patients with pneumonia in this study was small and the reported estimates are imprecise (as indicated by the wide 95% CIs). In addition, the high likelihood ratios also reflect the high specificity of tachypnea and indrawing in a particular group of patients (early infants). The value of the clinical examination may differ in this group of children. As in other studies, the absence of these findings did not dramatically decrease the likeli-

hood of disease for tachypnea (LR $-$, 0.55) or for indrawing (LR $-$, 0.75).

Accuracy of Combinations of Findings

Clinicians typically evaluate the presence of many findings simultaneously to rule in or rule out pneumonia. Despite the large number of studies, few have examined the value of clinical findings when they are used together. Two studies assessed the value of combinations of clinical findings. Leventhal⁴⁴ found that the absence of pulmonary findings defined as respiratory distress (nasal flaring, grunting, retractions), tachypnea, rales, or decreased breath sounds ruled out pneumonia (LR $-$, 0.0; 95% CI, 0.0-0.4). When present, these findings raised the likelihood of pneumonia to 1.6 (95% CI, 1.3-31). In this study, information about the presence or absence of respiratory symptoms was used in the decision to obtain the gold standard examination (a chest radiographic examination). Thus, the reported data are likely to overestimate the diagnostic accuracy of these combinations of findings so that the true LR $-$ is not as good as reported and the LR $+$ is better than reported.

In a study of children younger than 2 months, Crain et al⁴² found that the absence of any respiratory findings (rhinorrhea, cough, adventitious sounds, or retractions) decreased substantially the likelihood of a positive chest radiographic finding (LR $-$, 0.10; 95% CI, 0.03-0.4). The presence of any of these findings increased the likelihood of pneumonia to 3.4 (95% CI, 2.6-4.3). Because this study included only infants younger than 8 weeks, it is not clear how well the results apply to older age groups. Crain et al⁴² also found that as the number of positive respiratory findings increased, so did the probability of an abnormal chest radiographic finding.

To summarize, physical examination findings can help primary care physicians be more certain that an infant does or does not have pneumonia. In developed countries, where the prevalence of bacterial pneumonia is lower, pneumonia is unlikely if all signs are negative. The presence of a positive sign will be more useful in increasing clinicians' certainty that an infant has pneumonia in developing countries compared with developed countries because the prevalence of bacterial pneumonia is higher. In developed countries, clinicians will be more certain if multiple findings are positive. Further studies are needed to examine the diagnostic accuracy of the chest radiographic examination, the value of certain signs (such as fever and toxic appearance), and how to best take advantage of combinations of clinical findings.

THE BOTTOM LINE

First, the initial observation of the infant may be the most critical component of the examination. Observation is important before interacting with a child.

Second, because of its moment-to-moment variability, the respiratory rate should be counted by observing the chest while the child is quiet during two 30-second intervals or during a full minute. Clinicians need to be especially aware of the variability of the examination as the child's level of activity changes.

Third, auscultation is relatively unreliable for examination of infants. Clinicians need better training and better terminology to describe abnormal chest sounds. The overall clinical appearance may be accurate but the delineation of its value needs more study.

Fourth, the best individual finding for ruling out pneumonia is the absence of tachypnea. Chest indrawing and other signs of increased work of breathing (eg, nasal flaring) and abnormal auscultatory findings are better for ruling in pneumonia. In developed countries, multiple findings must be present for more certainty about the presence of pneumonia.

Fifth, if all clinical signs (respiratory rate, auscultation, and work of breathing) are negative, the chest radiographic finding is unlikely to be positive.

Author Affiliations at the Time of the Original Publication

Division of Community Pediatrics, University of North Carolina at Chapel Hill, Chapel Hill (Dr Margolis); Bassett Health Care, Cooperstown, New York (Dr Gadomski).

Acknowledgments

Dr Margolis is supported as a Robert Wood Johnson Generalist Faculty Physician Scholar.

We greatly appreciate the thoughtful comments of the following individuals who reviewed the manuscript: David Simel, MD, MHS, Joshua Metlay, MD, PhD, Dennis Clemmons, MD, PhD, and Jeffery Baker, MD, PhD. Debbie Sears provided assistance in the preparation of the manuscript and tables.

REFERENCES

- Henderson F, Reid H, Morris R, et al. Home air nicotine levels and urinary cotinine excretion in preschool children. *Am Rev Respir Dis*. 1989;140(1):197-201.
- Henderson F, Clyde W, Collier A, Denny F. The etiologic and epidemiologic spectrum of bronchiolitis in pediatric practice. *J Pediatr*. 1979;95(2):183-190.
- McConnochie K, Hall C, Barker W. Lower respiratory tract illness in the first two years of life: epidemiologic patterns and costs in a suburban pediatric practice. *Am J Public Health*. 1988;78(1):34-39.
- Hall C, Powell K, Schnabel K, Gala C, Pincus P. Risk of secondary bacterial infection in infants hospitalized with respiratory syncytial viral infection. *J Pediatr*. 1988;113(2):266-271.
- World Health Organization. Pneumonia and related infections in young children: an annotated selective bibliography, volume 1: articles published before 1991. In: *Programme for the Control of Acute Respiratory Infections*. Geneva, Switzerland: World Health Organization; 1993:173.
- Koopman P. Confidence intervals for the ratio of two binomial proportions. *Biometrics*. 1984;40:513-517.
- Centor R, Keightley J. *The Two by Two Analyzer*. Birmingham, AL: Center-Soft Statistical Programs for Medical Decision Making Research; 1992.
- Swischuk L, Hayden C. Viral vs bacterial pulmonary infections in children (is roentgenographic differentiation possible?). *Pediatr Radiol*. 1986;16(4):278-284.
- Bettenay F, Campo J, Crossin D. Differentiating bacterial from viral pneumonia in children. *Pediatr Radiol*. 1988;18(6):453-454.
- McCarthy P, Spiesel S, Stashwick C, et al. Radiographic findings and etiologic diagnosis in ambulatory childhood pneumonia. *Clin Pediatr (Phila)*. 1981;20(11):686-691.
- Isaacs D. Problems in determining the etiology of community-acquired childhood pneumonia. *Pediatr Infect Dis J*. 1989;8(3):143-148.
- Korppi M, Kiekara O, Heiskanen-Kosma T, Soimakallio S. Comparison of radiological findings and microbial aetiology of childhood pneumonia. *Acta Paediatr*. 1993;82(4):360-363.

13. Courtoy I, Lande A, Turner R. Accuracy of radiographic differentiation of bacterial from nonbacterial pneumonia. *Clin Pediatr (Phila)*. 1989; 28(6):261-264.
14. Simpson W, Hacking P, Court S, Gardner P. The radiological findings in respiratory syncytial virus infections in children, II: the correlation of radiological categories with clinical and virological findings. *Pediatr Radiol*. 1974;2(3):155-160.
15. Kirchner S, Horev G. Diagnostic imaging in children with acute chest and abdominal disorders. *Pediatr Clin North Am*. 1985;32(6):1363-1382.
16. Wildin S, Chonmaitree T, Swischuk L. Roentgenographic features of common viral respiratory tract infections. *Am J Dis Child*. 1988;142(1):43-46.
17. Mulholland E, Simoes E, Costales M, McGrath E, Manalac E, Gove S. Standardized diagnosis of pneumonia in developing countries. *Pediatr Infect Dis J*. 1992;11(2):77-81.
18. Lehrer S. *Understanding Breath Sounds*. Philadelphia, PA: WB Saunders Co; 1984:85.
19. Forgacs P. *Lung Sounds*. London, England: Bailliere Tindall; 1978:44-49.
20. Richards J, Alexander J, Shinebourne E, de Swiet M, Wilson A, Southall D. Sequential 22-hour profiles of breathing patterns and heart rate in 110 full-term infants during their first 6 months of life. *Pediatrics*. 1984;74(5):763-777.
21. Ashton R, Connolly K. The relation of respiration rate and heart rate to sleep states in the human newborn. *Dev Med Child Neurol*. 1971;13(2):180-187.
22. Hoppenbrouwers T, Harper R, Hodgeman J, Sterman M, McGinty D. Polygraphic studies of normal infants during the first six months of life, II: respiratory rate and variability as a function of state. *Pediatr Res*. 1978;12:120-125.
23. Iliff A, Lee V. Pulse rate, respiratory rate and body temperature of children between two months and eighteen years of age. *Child Dev*. 1952; 23(4):237-245.
24. Kendig E, Chernick V. *Disorders of Respiratory Tract in Children*. 4th ed. Philadelphia, PA: WB Saunders Co; 1983:62.
25. Finer N, Abroms I, Taeusch H. Influence of sleep state on the control of ventilation. *Pediatr Res*. 1975;9:365.
26. Gadomski A, Permutt T, Stanton B. Correcting respiratory rate for the presence of fever. *J Clin Epidemiol*. 1994;47(9):1043-1049.
27. Simoes E, Roark R, Berman S, Esler L, Murphy J. Respiratory rate: measurement of variability over time and accuracy at different counting periods. *Arch Dis Child*. 1991;66(10):1199-1203.
28. Gadomski A, Khallaf N, El A, Black R. Control of acute respiratory infections: respiratory rate and chest indrawing assessment by primary care physicians in Egypt. *Bull World Health Organ*. 1993;71(5):523-527.
29. Berman S, Simoes E, Lanata C. Respiratory rate and pneumonia in infancy. *Arch Dis Child*. 1991;66(1):81-84.
30. World Health Organization. *Acute Respiratory Infections in Children: Case Management in Small Hospitals in Developing Countries*. Geneva, Switzerland: World Health Organization, Programme for the Control of Acute Respiratory Infections; 1990 (WHO/ARI/90.5).
31. Campbell E. Physical signs of diffuse airways obstruction and lung distension. *Thorax*. 1969;24(1):1-3.
32. Hoover C. Definitive percussion and inspection in estimating size and contour of the heart. *JAMA*. 1920;75:1626-1630.
33. Poole S, Chetham M, Anderson M. Grunting respirations in infants and children. *Pediatr Emerg Care*. 1995;11(3):158-161.
34. Davis G, Bureau M. Pulmonary and chest wall mechanics in the control of respiration in the newborn. *Clin Perinatol*. 1987;14(3):551-579.
35. Murphy R, Holford S. Lung sounds. In: *Basics of Respiratory Disease*. New York, NY: American Thoracic Society; 1980:1-6.
36. Forbes J. *A Treatise on the Diseases of the Chest and on Mediate Auscultation*. Philadelphia, PA: Desilver Thomas & Co; 1835.
37. Margolis P, Ferkol T, Marsocci S, et al. Accuracy of the clinical examination in detecting hypoxemia in infants with respiratory illness. *J Pediatr*. 1994;124(4):552-560.
38. Wang E, Milner R, Navas L, Maj H. Observer agreement for respiratory signs and oximetry in infants hospitalized with lower respiratory infections. *Am Rev Respir Dis*. 1992;145(1):106-109.
39. Redd S, Patrick E, Vreuls R, et al. Comparison of the clinical and radiographic diagnosis of paediatric pneumonia. *Trans R Soc Trop Med Hyg*. 1994;88(3):307-310.
40. Redd S, Vreuls R, Metsing M, et al. Clinical signs of pneumonia in children attending a hospital outpatient department in Lesotho. *Bull World Health Organ*. 1994;72(1):113-118.
41. Harari M, Shann F, Spooner V, et al. Clinical signs of pneumonia in children. *Lancet*. 1991;338(8772):928-930.
42. Crain E, Bulas D, Bijur P, et al. Is a chest radiograph necessary in the evaluation of every febrile infant less than 8 weeks of age? *Pediatrics*. 1991;88(4):821-824.
43. Lozano J, Steinhoff M, Ruiz J, et al. Clinical predictors of acute radiological pneumonia and hypoxaemia at high altitude. *Arch Dis Child*. 1994;71(4):323-327.
44. Leventhal J. Clinical predictors of pneumonia as a guide to ordering chest roentgenograms. *Clin Pediatr*. 1982;21(12):730-734.
45. Taylor J, Del Beccaro M, Done S, Winters W. Establishing clinically relevant standards for tachypnoea in febrile children younger than 2 years. *Arch Pediatr Adolesc Med*. 1995;149(3):283-287.
46. Zukin D, Hoffman J, Cleveland R, et al. Correlation of pulmonary signs and symptoms with chest radiographs in the pediatric age group. *Ann Emerg Med*. 1986;15(7):792-796.
47. Greenes R, Begg C. Assessment of diagnostic technologies: methodology for unbiased estimation from samples of selectively verified patients. *Invest Radiol*. 1985;20(7):751-756.
48. Dai Y, Foy H, Zhu Z, et al. Respiratory rate and signs in roentgenographically confirmed pneumonia among children in China. *Pediatr Infect Dis J*. 1995;14(1):48-50.
49. Wafula E, Tindyebewa D, Onyango F. The diagnostic value of various features for acute lower respiratory infection among under fives. *East Afr Med J*. 1989;66(10):678-685.
50. Mulholland E, Ollinsky A, Shann F. Clinical findings and severity of acute bronchiolitis. *Lancet*. 1990;335(8700):1259-1261.
51. Selwyn B. The epidemiology of acute respiratory tract infection in young children: comparison of findings from several developing countries. *Rev Infect Dis*. 1990;12(suppl 8):S870-S888.
52. Pio A, Lewski J, Jen Dam H. The magnitude of the problem of acute respiratory infections. In: Douglas R, Kerby-Eaton E, eds. *Acute Respiratory Infections in Childhood: Proceedings of an International Workshop*. Adelaide, Australia: University of Adelaide; 1985:3-16.

Prepared by Daniel A. Ostrovsky, MD
Reviewed by Anne Gadomski, MD, MPH

CLINICAL SCENARIO

A 15-month-old child is brought to your office in May. She had been “breathing heavy” the previous day. She was well until about 2 days ago, when she developed nasal congestion with clear rhinorrhea, cough, and a low-grade fever. Your review shows this child had a normal birth history, demonstrated normal growth and development, and has not had any significant respiratory infections or reactive airway disease. On examination, you find a temperature of 38.2°C and a respiratory rate of 45/min. She has clear rhinorrhea and mild subcostal retractions but no abnormal lung sounds on auscultation.

UPDATED SUMMARY ON PEDIATRIC PNEUMONIA

Original Review

Margolis P, Gadomski A. Does this infant have pneumonia? JAMA. 1998;279(4):308-313.

UPDATED LITERATURE SEARCH

A MEDLINE search was conducted from 1996 to 2005 to identify English-language articles about pneumonia in infants or children, using the search strategy techniques of The Rational Clinical Examination series. The search yielded 49 articles. Additionally, Scientific Citation Index was used to identify articles that cited the original publication in The Rational Clinical Examination series, yielding 18 additional articles. The abstracts of these 67 articles were reviewed and all case-control, cohort, or randomized trials that addressed clinical signs and symptoms of pneumonia were selected for further inspection. The references for these articles were also reviewed to find any other relevant articles. The focus of the original publication was on identifying symptoms and signs that help distinguish pneumonia from other types of pediatric lower respiratory tract illnesses. In this update, we shifted the focus slightly and attempted to discover the findings that help identify the pediatric patient who will have an abnormal chest radiograph result. In total, 5 articles were selected for inclusion, although we subsequently excluded one article that

had confusing likelihood ratio (LR) results, and we were unable to contact the author for verification.¹

NEW FINDINGS

- Diminished breath sounds show substantial interrater reliability (κ , 0.73).¹
- Pulse oximetry with values less than 98% has a sensitivity of only 55% for pneumonia and has no independent utility after consideration of the auscultatory findings and respiratory rate.
- The LR for pneumonia is 3.4 when the onset of a respiratory illness was equal to or greater than 6 days.

Details of the Update

Since the publication of the original review, 4 additional studies evaluated different clinical findings for predicting radiographic changes suggestive of pneumonia in pediatric patients. Overall, there remains a paucity of data that examine combinations of clinical signs. Additionally, there remains difficulty in combining data from multiple studies because of differences in the definitions of certain clinical findings such as tachypnea and respiratory distress. Finally, the broad age range of patients included in the studies makes generalization of findings to infants more difficult. For example, grunting and nasal flaring would not be typical findings in older pediatric patients with pneumonia. The studies included in this update used age-based criteria for the finding of tachypnea.

A prospective study of children presenting to an emergency department with any type of acute respiratory illness provides useful information that allows comparison between signs and the overall clinical judgment.² As a single finding, tachypnea had the best diagnostic odds ratio (DOR; 5.8) that came from its positive LR of 2.2 (95% confidence interval [CI], 1.5-3.2) and negative LR of 0.39 (95% CI, 0.22-0.70). The additional information from chest indrawing and alveolar rales did not clinically improve the diagnostic odds or LRs. Clinical judgment that factors in all items from the medical history and physical examination (DOR, 3.6; 95% CI, 1.5-8.7) had results that were slightly less efficient than the single finding of tachypnea. Clinicians should recall the age-based World Health Organization (WHO) definitions of tachypnea for infants (Table 41-4).

Table 41-4 World Health Organization Age-based Criteria for Tachypnea

Age, mo	Tachypnea, Breaths/min
<2	>60
2-12	>50
>12	>40

A case-controlled study from retrospectively collected data suggested that pulse oximetry at a threshold of 98% has no value for diagnosing pneumonia.³ Although clinical examination data reported from case-controlled studies typically provide a low level of evidence, the findings here supported the usefulness of tachypnea. Pulse oximetry added no significance to a model containing the respiratory rate and auscultatory findings. Unfortunately, the model itself was not particularly powerful for predicting pneumonia. ($R^2 = 0.072$ is a measure of how well the model predicts the outcome. The value means that the model explains only 7.2% of the variance, a statistically significant result, although one that will lead to incorrect diagnoses for many patients.)

The patient selection criteria affects the interpretation of the results. A study that included wheezing children (younger than 18 months) first determined the factors associated with the clinician ordering a chest radiograph.⁴ The presence of any typical clinical sign for pneumonia was associated with a request for chest radiograph. When confined to wheezing young children, the presence of grunting worked better than tachypnea with a respiratory rate of greater than 60/min (a rapid rate in comparison with the WHO standards noted above). The presence of grunting had an LR of 2.7 for pneumonia (95% CI, 1.6-4.4). However, when combined with a low oxygen saturation of less than 93% (much lower than the threshold in the case-control study), the combination of grunting and a low oxygen saturation in a wheezing young child had an LR of 4.0 (95% CI, 1.3-12). Unfortunately, the absence of both these findings had little effect on ruling out pneumonia (LR, 0.90 when both signs were normal; 95% CI, 0.81-1.0).

A prospective study provides some insight into the probability of pneumonia once the physician requests chest radiography (prevalence, 36%).⁵ The prevalence among all children with res-

piratory symptoms would be lower. Tachypnea, using the WHO criteria for respiratory rate, was similar in utility to that of previous studies (positive likelihood ratio [LR+], 2.8; 95% CI, 1.6-5.0; negative likelihood ratio [LR-], 0.91; 95% CI, 0.86-0.97).

Another prospective cohort study⁶ found that tachypnea had an LR+ of 1.4 (95% CI, 1.0-1.9) and an LR- of 0.67 (95% CI, 0.44-1.0). However, the age-adjusted definitions for tachypnea required a much higher respiratory rate than the WHO criteria. These poor results for tachypnea allow the inference that the WHO criteria are necessary to optimize the clinical utility of the finding.

IMPROVEMENTS IN THE DATA PRESENTED IN THE ORIGINAL PUBLICATION

New data allowed for comparison with the data from the original review. For the most part, the new data confirmed the findings of the original review. Among infants and young children with respiratory symptoms or signs, a broad range of prevalence should be considered (15%-35%) that may show seasonal and geographic variation. The data confirm that although tachypnea may be the most predictive in ruling in or ruling out pneumonia, no clinical examination finding alone is sufficiently powerful to predict the presence or absence of pneumonia.

CHANGES IN THE REFERENCE STANDARD

The reference standard for the diagnosis of pediatric pneumonia remains the chest radiograph.

RESULTS OF LITERATURE REVIEW

Selected Univariate Findings for Pediatric Pneumonia

Clinical judgment, a measure that allows the clinician to consider all findings, may not work better than individual findings (Table 41-5). When the clinician suspects pneumonia, the LR is 1.7 to 2.5; when the clinician suspects the child has no pneumonia, the LR is 0.29 to 0.46.^{2,5}

Table 41-5 Likelihood Ratios of Univariate Findings for Pediatric Pneumonia

Source	Finding	LR+ (95% CI)	LR- (95% CI)
Lynch et al ⁵	Tachypnea (WHO criteria)	2.8 (1.6-5.0)	0.91 (0.86-0.97)
Palafox et al ²	Tachypnea (WHO criteria)	2.2 (1.5-3.2)	0.39 (0.22-0.70)
Mahabee-Gittens et al ⁴	Grunting and pulse oximetry < 93%	4.0 (1.3-12)	0.90 (0.81-1.0)
Mahabee-Gittens et al ⁴	Grunting among children wheezing, < 18 mo	2.7 (1.6-4.4)	0.7 (0.55-0.89)
Lynch et al ⁵	Retractions	2.7 (1.1-6.9)	0.97 (0.93-1.0)
Palafox et al ²	Chest indrawing (retractions)	1.7 (1.2-2.4)	0.54 (0.32-0.91)
Palafox et al ²	Clinical judgment	1.7 (1.2-2.3)	0.46 (0.25-0.84)
Lynch et al ⁵	Fever	1.2 (1.1-1.3)	0.30 (0.18-0.49)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio; WHO, World Health Organization.

Multivariate Findings for Pediatric Pneumonia

In a study by Lynch et al,⁵ a multivariate model was assessed for diagnosing pediatric pneumonia. The study evaluated combinations of findings and created a pneumonia score that also supported a role for assessing tachypnea:

$$\text{Pneumonia score} = -4.71 + 1.10 \times (\text{tachypnea}) + 0.74 \times (\text{crackle}) + 0.42 \times (\text{decreased breath sound}) + 1.15 \times (\text{measured fever})$$

$$\text{Probability of pneumonia} = \exp^{\text{score}} / (1 + \exp^{\text{score}})$$

(The presence of a finding is coded as 1, whereas the absence of a finding is coded as 0. The presence of tachypnea is based on age-adjusted rates.)

The most useful finding from this model is that the absence of all 4 findings leads to a less than 1% probability of pneumonia. The presence of all 4 findings creates a probability of 21%, which suggests the need for a chest radiograph but does not establish a clinical diagnosis with a high degree of confidence. The area under the receiver operating characteristic curve was only 0.67 (a measure of accuracy), highlighting the finding that even combinations of signs lack a high level of efficiency for diagnosing pneumonia.

The model may be best for identifying signs that physicians might consider as part of their clinical judgment. However, clinicians should recognize that their overall clinical judgment and the results from a more structured approach in the form of a logistic model lack accuracy.

EVIDENCE FROM GUIDELINES

Jadavji et al⁷ published guidelines in 1997 for the diagnosis and management of pediatric pneumonia. They conducted a systematic review on the etiology, diagnosis, and management of pediatric pneumonia. The evidence from this review includes the studies that were reviewed in the original Rational Clinical Examination article, with 2 exceptions. One study focused on infants younger than 4 months and therefore not as easily generalized to the overall pediatric population.⁸ Overall, the data from this guideline are consistent with the findings of the original Rational Clinical Examination article.

CLINICAL SCENARIO—RESOLUTION

This infant may have pneumonia. According to WHO criteria, she has tachypnea, although she is febrile, which could explain her mildly increased respiratory rate. Tachypnea should raise your suspicion for pneumonia, with its best LR+ of about 2.8. Although she has mild retractions that would seem to further increase the likelihood of pneumonia, the additional information provided by this sign is less accurate than the information from tachypnea alone. The clinical history and time of year would make you less suspicious of other entities such as asthma or infection from respiratory syncytial virus (RSV). From the original Rational Clinical Examination article and this Update, you estimate a prevalence range of 15% to 35% at the lower end of this range. The posttest probability of pneumonia with an LR of 2.8 for tachypnea is 33%. From the multivariate model, she has tachypnea and fever, making the probability of pneumonia 7.9%. In this infant, it would be reasonable to check a chest radiograph to confirm or exclude pneumonia.

REFERENCES FOR THE UPDATE

1. Rothrock S, Green S, Fanelli JM, Cruzen E, Costanzo K, Pagane J. Do published guidelines predict pneumonia in children presenting to an urban ED? *Pediatr Emerg Care*. 2001;17(4):240-242.
2. Palafox M, Guiscafre H, Reyes H, Munoz O, Martinez H. Diagnostic value of tachypnea in pneumonia defined radiographically. *Arch Dis Child*. 2000;82(1):41-45.^a
3. Tanen DA, Trocinski DR. The use of pulse oximetry to exclude pneumonia in children. *Am J Emerg Med*. 2002;20(6):521-523.
4. Mahabee-Gittens EM, Dowd MD, Beck JA, Smith SZ. Clinical factors associated with focal infiltrates in wheezing infants and toddlers. *Clin Pediatr*. 2000;39(7):387-393.^a
5. Lynch T, Platt R, Gouin M, Larson C, Patenaude Y. Can we predict which children with clinically suspected pneumonia will have the presence of focal infiltrates on chest radiographs? *Pediatrics*. 2004(3 pt 1);113:186-189.^a
6. Grossman L, Caplan S. Clinical, laboratory, and radiological information in the diagnosis of pneumonia in children. *Ann Emerg Med*. 1988;17(1):87-90.^a
7. Jadavji T, Law B, Lebel M, Kennedy W, Gold R, Wang E. A practical guide for the diagnosis and treatment of pediatric pneumonia. *CMAJ*. 1997;156(5 suppl):S703-S711.
8. Berman S, Simoes EA, Lanata C. Respiratory rate and pneumonia in infancy. *Arch Dis Child*. 1991;66(1):81-84.

^aFor the Evidence to Support the Update for this topic, see <http://www.JAMAevidence.com>.

PNEUMONIA, INFANT AND CHILD—MAKE THE DIAGNOSIS

PRIOR PROBABILITY

Given cough or respiratory symptoms, the prevalence of pneumonia is approximately 15% to 35%. However, prevalence of pneumonia may be lower during RSV season. Prevalence may also be slightly higher in children younger than 3 years.

POPULATION FOR WHOM PEDIATRIC PNEUMONIA SHOULD BE CONSIDERED

Patients with symptoms of acute respiratory illness, primarily cough, respiratory distress, or tachypnea, need to have pneumonia considered as part of the differential diagnosis.

DETECTING THE LIKELIHOOD OF PEDIATRIC PNEUMONIA

The individual clinical symptoms used to identify patients with pneumonia have relatively poor predictive value. Tachypnea, respiratory distress, and abnormal lung sounds (rales) have the best operating characteristics, although the data from different sources conflict on their significance ([Table 41-6](#)). Additionally, the clinician's overall clinical judgment/impression may have operating characteristics similar to individual signs and symptoms in diagnosing pneumonia, but the overall judgment is admittedly a complex and difficult "finding" to quantify. To date, there are no randomized controlled studies to validate any proposed multivariate model for predicting pneumonia.

Table 41-6 Likelihood Ratios of Symptoms and Signs for Pediatric Pneumonia

Symptom or Sign	LR+ (95% CI) or Range	LR- (95% CI) or Range
Grunting among children with wheezing, < 18 mo	2.8 (1.6-4.4)	0.7 (0.55-0.89)
Grunting	2.8-3.2	0.70-0.86
Retractions	2.7 (1.1-6.9)	0.97 (0.93-1.0)
Rales	1.8-15	0.69-0.86
Tachypnea (use WHO age-adjusted criteria)	1.6-8.0	0.32-0.91
Fever	1.2-1.5	0.17-0.30

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio; WHO, World Health Organization.

REFERENCE STANDARD TESTS

The reference standard for pediatric pneumonia remains the chest radiograph. Sputum production is not a frequent finding in pediatric patients, and therefore, isolation of sputum for microbiologic correlation with pneumonia remains both difficult and impractical. The development of rapid antigen detection of common viruses such as RSV and influenza will help the clinician rule out causes of respiratory symptoms other than bacterial pneumonia. As of now, there is still no way to differentiate bacterial vs viral pneumonia by chest radiograph.

EVIDENCE TO SUPPORT THE UPDATE:

41

Pneumonia, Infant and Child

TITLE Clinical, Laboratory, and Radiological Information in the Diagnosis of Pneumonia in Children.

AUTHORS Grossman L, Caplan S.

CITATION *Ann Emerg Med*. 1988;17:(1)43-46.

QUESTION In pediatric patients with suspected pneumonia who undergo chest radiograph, are there signs or symptoms that predict radiographic pneumonia?

DESIGN This is a prospective nonconsecutive cohort study of 155 patients during 7 months.

SETTING Two pediatric emergency departments.

PATIENTS Pediatric patients younger than 19 years in whom pneumonia was considered and a chest radiograph was ordered. None of the patients had a history of pneumonia, chronic lung disease, chronic heart disease, or immunodeficiency. Sixty-two percent of the study patients were younger than 2 years. Eleven potential subjects were not enrolled because a decision to treat was made without a chest radiograph being performed.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

The presenting signs and symptoms of patients before chest radiography were systematically recorded. Clinicians (pediatricians, pediatric nurse practitioners, and medical students) recorded their overall clinical impression of pneumonia and what their treatment plan would be if radiography were not available before performance of the chest radiograph. Chest radiograph was the reference standard for the diagnosis of pneumonia.

MAIN OUTCOME MEASURES

Sensitivity and specificity of recorded signs, symptoms, and clinical impression for predicting radiographic pneumonia. Assessing accuracy of clinical impression in predicting pneumonia. Assessing combinations of signs and symptoms in predicting pneumonia.

MAIN RESULTS

Cough, tachypnea, moderate/severe degree of illness, and fever were the only symptoms and signs that were present in more than 50% of patients enrolled in the study (66%, 52%, 62%, and 55%, respectively).

Clinician accuracy in the diagnosis of pneumonia was 77%, and both the positive and negative likelihood ratios (LRs) were more promising than the individual findings ([Table 41-7](#)).

Despite the results for clinical judgment, regression analysis did not find any combination of signs or symptoms that adequately predicted the presence of pneumonia.

CONCLUSIONS

LEVEL OF EVIDENCE Level 4.

STRENGTHS This was a prospective cohort study. All patients enrolled underwent the reference standard of chest radiograph. There was quantification of the different signs and symptoms that led clinicians to order chest radiographs.

LIMITATIONS The definition of tachypnea used in this study is different from the World Health Organization (WHO) criteria. If WHO criteria had been used, there might have been a

Table 41-7 Likelihood Ratios of Findings for Pediatric Pneumonia

Findings	Sensitivity, %	Specificity, %	LR+ (95% CI)	LR- (95% CI)
Clinical judgment	80	68	2.5 (1.8-3.4)	0.29 (0.17-0.52)
Rales	43	77	1.9 (1.2-3.0)	0.74 (0.57-0.96)
Tachypnea ^a	64	54	1.4 (1.0-1.9)	0.67 (0.44-1.0)
Decreased breath sounds	23	84	1.4 (0.74-2.8)	0.92 (0.77-1.1)
Degree of illness ^b	67	40	1.1 (0.87-1.4)	0.83 (0.52-1.3)
Sudden onset of illness ^c	17	84	1.1 (0.50-2.2)	0.99 (0.85-1.1)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

^a>80/min < 1 year; >40/min > 1 year; >30/min > 2 years; >25/min > 5 years; >22/min > 10 years; >20/min > 15 years.¹ (This differs from the World Health Organization definition for tachypnea.)

^bDegree of illness not further clarified in article.

^cLess than 12 hours of symptoms before presentation.

much higher number of patients who would have been classified as having tachypnea, which might have increased the positive likelihood ratio (LR+) of tachypnea.

There was no explanation of what “degree of illness” means, and therefore, it has limited clinical utility.

There was no mention of blinding of radiologists to the clinical presentation of study patients. There was no description of what qualified a radiograph as being diagnostic of pneumonia.

The results provided did not include 95% confidence interval. Additionally, it is unclear how many “observations” were made for each patient because there were multiple examiners for each patient.

CONCLUSIONS

The focus of this study was to determine whether there were any signs or symptoms that were helpful in diagnosing pneumonia in children younger than 18 years and presenting with symptoms sufficient to warrant a chest radiograph. Additionally, it sought to assess how the results of the radiograph influenced management decisions by the ordering clinician. Finally, it attempted to assess the clinician’s overall impression as a predictor of pneumonia. It is useful to know that cough, tachypnea, “moderate/severe degree of illness,” and fever are the most common symptoms and signs for which a radiograph is ordered. This may give a hint as to what goes into the clinician’s overall clinical impression when he or she considers the diagnosis of pneumonia. In this study, the overall clinical impression performed better (LR+, 2.5) than any individual sign or symptom in diagnosing pneumonia. Physician accuracy of diagnosing pneumonia was 77%. Obtaining radiographs is useful because they changed management plans for 22% of study patients. In this study, only rales and tachypnea reached statistical significance in predicting pneumonia. However, both of these signs had only marginal diagnostic power, with LR+s of only 1.9 and 1.4, respectively.

Reviewed by Daniel Ostrovsky, MD

REFERENCE FOR THE EVIDENCE

- Johnson TR. Development of the lungs. In: Johnson TR, Moore WM, Jeffries JE, eds. *Children Are Different*. 2nd ed. Columbus, OH: Ross Laboratories; 1978:128-129.

TITLE Can We Predict Which Children With Clinically Suspected Pneumonia Will Have the Presence of Focal Infiltrates on Chest Radiographs?

AUTHORS Lynch T, Platt R, Gouin M, Larson C, Patenaude Y.

CITATION *Pediatrics*. 2004;113(3 pt 1):186-189.

QUESTION In patients presenting to the emergency department with “clinically suspected pneumonia,” what clinical factors predict an infiltrate on radiograph?

DESIGN Prospective nonconsecutive cohort study of 570 patients.

SETTING Tertiary-referral-center pediatric emergency department.

PATIENTS Children aged 1 to 16 years who were suspected by the pediatric emergency physician to have pneumonia and who were receiving a chest radiograph. Children with chronic respiratory disease, congenital or complex heart disease, gastroesophageal reflux, sickle cell anemia, malignancy, spastic quadriplegia, acute asthma exacerbation, or recent pneumonia treatment with antibiotics were excluded.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

Baseline demographic data and a study questionnaire were obtained prospectively. Physicians filled out a clinical questionnaire about symptoms at evaluation.

All subjects had a posterior-anterior and lateral chest radiograph evaluated by 3 radiologists for the presence or absence of infiltrates.

MAIN OUTCOME MEASURES

Six symptoms (fever history, cough, coryza, shortness of breath, wheezing, and pleurisy), 6 signs (decreased breath sounds, crackles, bronchial sounds, wheezing, retractions, and grunting), and 3 vital signs (measured temperature for fever, age-adjusted tachypnea, and tachycardia) were entered into a logistic model to determine their independent significance.¹ Previously recommended guidelines were assessed for their sensitivity and specificity. The interobserver agreement for the chest radiographs was assessed.

MAIN RESULTS

Five hundred seventy patients were enrolled, of whom 204 (36%) had pneumonia. The agreement between 7 radiologists for the presence of pneumonia was moderate (weighted κ , 0.57).

Only 2 findings, when present, had a likelihood ratio (LR) that was greater than 2 and that excluded 1 in its 95% confidence interval (CI): the presence of tachypnea (8% of children) had an LR of 2.8 (95% CI, 1.6-5.0), whereas retractions (3% of children) had an LR of 2.7 (95% CI, 1.1-6.9). For decreasing the likelihood of pneumonia, the absence of a fever (LR, 0.30; 95% CI, 0.18-0.49) or cough (LR, 0.35; 95% CI, 0.54-0.81) was the only finding with an LR less than 0.6 and that excluded 1.0 from the 95% CI.

A logistic model (Box 41-1) identified 4 findings that were independently useful. However, the model was not highly accurate because of its poor specificity (area under the receiver operating characteristic curve of 0.67).

CONCLUSIONS

LEVEL OF EVIDENCE

Level 3.

STRENGTHS Multiple radiologists who were blinded to the clinical presentation evaluated the reference standard. There was a large sample size. Multiple clinical predictors were assessed with regression analysis.

LIMITATIONS The enrollment process may have caused selection bias, but only fever history and cough occurred in more than half the patients, which suggests that the remaining findings might have valid results because they were not used to preferentially identify children.

When the likelihood of focal opacities is predicted in children clinically suspected of having pneumonia, only 4 signs and symptoms are independently statistically significant. The model is highly sensitive but poorly specific. To highlight this, the probability of pneumonia can be contrasted for the child with no tachypnea, crackles, decreased breath sounds, or fever (probability, 0.9%) vs a child with all 4 findings present (probability, 21%). Thus, a child with no findings has a less than 1% chance of having pneumonia. On the other hand, even with the presence of all 4 findings, most children will not have a radiographic infiltrate.

Reviewed by Daniel Ostrovsky, MD

REFERENCES FOR THE EVIDENCE

- Chamberlain JM, Patel KM, Ruttimann UE, Pollack MM. Pediatric risk of admission (PRISA): a measure of severity of illness for assessing the risk of hospitalization from the emergency department. *Ann Emerg Med*. 1998;32(2):161-169.
- Leventhal JM. Clinical predictors of pneumonia as a guide to ordering chest roentgenograms. *Clin Pediatr*. 1982;21(12):730-734.

Box 41-1 Logistic Model for Calculating the Pneumonia Score

$$\text{Pneumonia score} = -4.71 + 1.10 \times (\text{tachypnea}) + 0.74 \times (\text{crackle}) + 0.42 \times (\text{decreased breath sound}) + 1.15 \times (\text{measured fever})$$

The presence of a finding is coded as 1, whereas the absence of a finding is coded as 0. The presence of tachypnea is based on age-adjusted rates.

$$\text{Probability of pneumonia} = \exp^{\text{score}} / (1 + \exp^{\text{score}})$$

The investigators attempted to validate the guidelines by Leventhal² (respiratory distress, tachypnea, rales, and decreased breath sounds) but found these yielded a sensitivity of only 81% and a specificity of 37%.

TITLE Clinical Factors Associated With Focal Infiltrates in Wheezing Infants and Toddlers.

AUTHORS Mahabee-Gittens EM, Dowd MD, Beck JA, Smith SZ.

CITATION *Clin Pediatr*. 2000;39(7):387-393.

QUESTION In wheezing infants presenting to an emergency department, are there clinical factors that can predict focal infiltrates on chest radiograph?

DESIGN Prospective cohort of infants up to 18 months of age.

SETTING The study took place during October and April at the Children's Hospital Medical Center in Cincinnati, Ohio, a tertiary-care hospital pediatric emergency department.

PATIENTS Infants aged 18 months or younger and presenting to the emergency department. Inclusion was a convenience sample of patients with documented wheezing on physical examination by a physician.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

Patients selected for inclusion had baseline information collected prospectively. Physical examination findings were documented at evaluation. The reference standard was a chest radiograph for the presence of focal infiltrates. The reference standard was applied at the discretion of the evaluating physician. A radiologist who was not masked to the clinical presentation interpreted the radiographs. A report result was considered positive if it recorded "focal infiltrate," "pneumonia," "consolidation," or "atelectasis vs infiltrate."

MAIN OUTCOME MEASURES

The authors collected data on all potential eligible patients and compared the odds ratios for physical examination signs for individuals selected for chest radiographs vs those who did not undergo radiography. These odds ratios describe the factors associated with requesting a radiograph.

Sensitivity, specificity, and odds ratios of the clinical findings for diagnosing pneumonia were calculated for children who underwent radiography. Interobserver variability was assessed in 12% of the children.

MAIN RESULTS

Among 471 children who made a visit to the emergency department with wheezing and were potentially eligible, 212 had chest radiographs. Twenty-three percent (49/212) had a focal infiltrate. Except for localized wheezing, each sign in Table 41-8 was more likely present in a child receiving a chest radiograph than one who did not (odds ratio with lower 95% confidence interval ≥ 1.0).

In patients who did not undergo chest radiograph, follow-up telephone calls and searches of admission databases were made 48 hours after presentation to look for patients who may have incorrectly been classified as not having pneumonia. Only 3 patients who did not undergo chest radiography

were hospitalized within 2 days of presentation. All 3 patients had a chest radiograph on representation and none of them had an infiltrate. Seventeen other patients who did not initially undergo chest radiography had a subsequent chest radiograph in the following 48 hours. Three of these patients had an infiltrate.

Oxygen saturation, nasal flaring, grunting, crackles, and retractions were all reliable and had $\kappa > 0.70$.

CONCLUSIONS

LEVEL OF EVIDENCE Level 3.

STRENGTHS The study design was prospective. A uniform entrance criterion (wheezing) identified potentially eligible patients with a narrower age range than was present in some of the other studies. The authors compared the signs present in children who underwent radiographs vs those who did not.

LIMITATIONS The enrollment process created a convenience sample that leads to selection bias. Radiologists who were not masked to the clinical presentation interpreted the outcome measure. There was only 1 radiologist per case, which could lead to accuracy issues in interpretation. This study was done in a population during a period when respiratory syncytial virus bronchiolitis typically has a high prevalence.

Children younger than 18 months who wheezed were the focus of this study. A number of clinical signs were more likely present in children who were referred for chest radiography. However, most of these signs were not particularly useful when either present or absent. As a single finding, the presence of grunting (present in 60 children overall and in 45 referred for radiography [21%]) was the most useful finding, with a likelihood ratio of 2.7. The absence of any of these findings was clinically not useful. When combined with low oxygen saturation, a logistic model selected grunting with low oxygen saturation as useful. The likelihood ratio increased to 4.0 for the presence of these 2 signs.

Clinicians should recognize that the prior probability of pneumonia has seasonal variation in the pediatric population. Bronchiolitis, an illness that may “look like pneumonia,” is more common in the winter and is associated with tachypnea and abnormal lung findings. Thus, the complex relationship between changing prevalence of disease and seasonal variation in signs affects the interpretation of the predictive power of these findings.

Reviewed by Daniel A. Ostrovsky, MD

Table 41-8 Likelihood Ratios of Findings for Pediatric Pneumonia

Sign (No. With the Finding)	LR+ (95% CI)	LR- (95% CI)	DOR (95% CI)
Vital Signs			
Temperature $\geq 100.4^{\circ}\text{F}$ (38.0°C) (115)	1.12 (0.95-1.6)	0.76 (0.51-1.1)	1.6 (0.8-3.1)
Respiratory rate $> 60/\text{min}$ (61)	1.1 (0.67-1.8)	0.97 (0.78-1.2)	1.1 (0.6-2.2)
Oxygen saturation $\leq 93\%$ (41)	2.0 (1.1-3.4)	0.82 (0.67-1.1)	2.4 (1.1-5.0)
Physical Examination			
Nasal flaring (82)	1.3 (0.90-1.9)	0.70 (0.55-0.89)	1.6 (0.8-3.0)
Grunting (45)	2.7 (1.6-4.4)	0.90 (0.79-1.0)	3.8 (1.9-7.8)
Crackles (67)	1.6 (1.1-2.4)	0.76 (0.58-1.0)	2.1 (1.1-4.1)
Decreased breath sounds (19)	2.4 (1.0-5.7)	0.90 (0.79-1.0)	2.7 (1.0-7.2)
Localized wheezing (21)	1.0 (0.4-2.7)	1.0 (0.90-1.1)	1.0 (0.4-3.0)
Retractions (202)	1.0 (0.94-1.1)	0.83 (0.18-3.8)	1.2 (0.2-5.9)
I:E $\geq 1:2$ (166)	1.2 (1.0-1.3)	0.43 (0.18-1.0)	2.8 (1.0-7.5)
Combination of Findings			
Grunting and oxygen saturation $\leq 93\%$ (48) ^a	4.0 (1.3-12)	0.90 (0.81-1.0)	4.4 (1.3-15)

Abbreviations: CI, confidence interval; DOR, diagnostic odds ratio; I:E, length of time in inspiration in proportion to time in expiration; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

^aVariables selected as independently useful in a logistic model.

TITLE Diagnostic Value of Tachypnoea in Pneumonia Defined Radiologically.

AUTHORS Palafox M, Guiscafre H, Reyes H, Munoz O, Martinez H.

CITATION *Arch Dis Child.* 2000;82(1):41-45.

QUESTION In children presenting with acute respiratory infection, what are the sensitivity and specificity of tachypnea for diagnosing pneumonia?

DESIGN This study is a prospective cohort study of children presenting to an emergency department with an acute respiratory tract infection. All children had chest radiography. Baseline characteristics and physical examination findings were obtained prospectively.

SETTING A general hospital in Mexico that is a referral center for sick children.

PATIENTS Eligible children were between 3 days and 5 years of age, required medical care during the 6-month study period, had been clinically diagnosed with pneumonia, and had the disease for fewer than 2 weeks. Each child in the study had a matched control who was the next child treated in the clinical unit and had a diagnosis of an acute respiratory infection without pneumonia. Exclusion criteria were children with chronic diseases, genetic abnormalities, neurologic diseases, asthma, or sepsis.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

The respiratory rate was measured for 1 minute, with the child lying down, not crying, and without fever. Tachypnea was defined by age-based World Health Organization (WHO) criteria (Table 41-9).

A chest radiograph, evaluated by a single radiologist blinded to the clinical diagnosis, served as the reference standard.

MAIN OUTCOME MEASURES

Sensitivity and specificity of tachypnea, chest indrawing, alveolar rales, and combinations of these findings. Ten radiographs were reassessed to determine the intraobserver variation.

MAIN RESULTS

Thirty-five children (32%) had pneumonia. There were 7 signs and symptoms or combinations that had significant sensitivity and specificity for predicting pneumonia on chest radiograph. Tachypnea had the best sensitivity of the signs studied (74%), followed by chest indrawing (71%). Although combining signs did slightly improve specificity, it decreased sensitivity. Alveolar rales had the best specificity but had poor sensitivity (Table 41-10).

Table 41-9 World Health Organization Age-based Criteria for Tachypnea

Age, mo	Tachypnea, Breaths/min
<2	>60
2-12	>50
>12	>40

A discriminant analysis, using all the recorded symptoms and signs, was 71% accurate but not appreciably different from the accuracy of tachypnea alone (69% accurate). The discriminant analysis performed better than clinical judgment (62% accurate).

If a patient had disease at least 6 days, the likelihood ratio was 3.4. A discriminant analysis revealed that duration of disease correctly classified 83.3% of patients.

The κ statistic for intraobserver variability of the radiologist was 0.68.

CONCLUSIONS

LEVEL OF EVIDENCE Level 2.

STRENGTHS Whereas other studies included children according to whether they had a chest radiograph, this study included a broader population of patients for whom the diagnosis of pneumonia was a reasonable consideration. The “control” patients were patients with some type of respiratory illness (cough or rhinorrhea with systemic signs of infection). These patients were not really “controls,” but rather patients at risk for pneumonia and in whom pneumonia could have been part of the differential diagnosis (although at a lower likelihood than the for case patients). All included study patients underwent the reference standard. The radiologists were masked to the clinical presentation. Intraobserver variability of the radiologist reading the radiographs was tested.

LIMITATIONS Among all children with respiratory illnesses, the “case patients” were oversampled, which can lead to an overestimation of sensitivity (and underestimation of specificity). A single radiologist performed the interpretation of the radiographs, although there was an attempt to account for this by measuring intraobserver variability and masking the radiologist to the clinical presentation.

Of the presenting clinical signs, all except chest indrawing (51%) occurred in less than half the patients, which allows us to make inferences about the utility of the findings because no one finding was required in each patient. It is remarkable that the overall clinical judgment had a diagnostic odds ratio (a measure of accuracy) that was not quite as good as the single finding of tachypnea. Tachypnea, defined by WHO criteria, was the most accurate finding as evidenced by its diagnostic odds ratio.

In subgroup analysis using tachypnea as the clinical sign being evaluated, there was no significant difference in the sensitivity and specificity generated for children of differing

Table 41-10 Likelihood Ratios of Findings for Pediatric Pneumonia

Test (No. With Finding)	Sensitivity	Specificity	LR+ (95% CI)	LR- (95% CI)	DOR (95% CI)
Tachypnea, chest indrawing, and alveolar rales (27)	0.43	0.84	2.7 (1.4-5.1)	0.68 (0.50-0.92)	4.0 (1.6-9.8)
Tachypnea and alveolar rales (29)	0.46	0.83	2.6 (1.4-4.9)	0.70 (0.48-0.91)	4.1 (1.7-10)
Tachypnea (51)	0.74	0.67	2.2 (1.5-3.2)	0.39 (0.22-0.70)	5.8 (2.4-14)
Tachypnea and chest indrawing (47)	0.68	0.69	2.1 (1.4-3.2)	0.50 (0.31-0.80)	4.7 (2.0-11)
Alveolar rales (32)	0.46	0.79	2.1 (1.2-3.8)	0.69 (0.50-0.96)	3.2 (1.3-7.6)
Chest indrawing and alveolar rales (30)	0.42	0.80	2.1 (1.2-3.9)	0.71 (0.53-0.97)	1.2 (2.9-7.0)
Clinical judgment (59)	0.74	0.56	1.7 (1.2-2.3)	0.46 (0.25-0.84)	3.6 (1.5-8.7)
Chest indrawing (56)	0.71	0.59	1.7 (1.2-2.4)	0.54 (0.32-0.91)	3.5 (1.5-8.3)

Abbreviations: CI, confidence interval; DOR, diagnostic odds ratio; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

age groups. There was a significant difference in the sensitivity and specificity of tachypnea when disease duration was considered. Sensitivity increased from 55% to 93% if disease was fewer than 3 days' duration or more than 6 days' duration, respectively. Specificity increased from 64% to 73% as well.

Reviewed by Daniel A. Ostrovsky, MD

CHAPTER 42

Is This Patient Pregnant?

Lori A. Bastian, MD, MPH

Joanne T. Piscitelli, MD

CLINICAL SCENARIOS

Are These Patients Pregnant?

For each of the following cases, the clinician may need to determine the probability that the patient is pregnant.

CASE 1 A 36-year-old woman telephones her primary care physician, complaining of symptoms consistent with uncomplicated sinusitis. Before treating her with an antibiotic, you ask her about the possibility of pregnancy; she states her last menstrual period was 3 weeks ago and she is not pregnant.

CASE 2 A sexually active 16-year-old girl requests birth control pills and asks during the pelvic examination, when her mother has stepped out of the room, if you can tell whether she is pregnant. Her last menstrual period was 8 weeks ago, her home pregnancy test result was negative, and findings on her pelvic examination were normal.

CASE 3 A 41-year-old woman presents with breast tenderness, and her last menstrual period was 6 weeks ago. She wants to know whether she is “going through the change.”

WHY IS THIS AN IMPORTANT QUESTION TO ANSWER WITH A CLINICAL EXAMINATION?

Frequent laboratory analyses are performed in the outpatient clinic and emergency department to rule in or to rule out the possibility of pregnancy. Generally accepted clinical indicators of pregnancy include amenorrhea, morning sickness, tender or tingling breasts, and, after 8 weeks' gestational age (defined as weeks since the last menstrual period), an enlarged uterus with a soft cervix. Standard textbooks of obstetrics do not indicate the value (ie, sensitivity and specificity) of these symptoms and signs as predictors of the diagnosis of early pregnancy.

In the outpatient clinical setting, there are many reasons to determine whether the patient is pregnant, including avoiding nonurgent radiographs; avoiding teratogenic drugs, such as anticonvulsants; initiating early prenatal care; reassuring the patient; and explaining the multiple nonspecific complaints easily confused with the early symptoms of pregnancy.

We are reviewing a common problem facing the primary care physician: When treating or evaluating a woman of childbearing years, what is the value of historical or physical examination features in determining the probability of early pregnancy? We will focus on the patient's medical history and physical examination findings that help the clinician rule in or rule out early pregnancy. We intend to answer the following questions: (1) What is the value of history and symptoms in determining the probability of early pregnancy? (2) How accurate are home pregnancy tests (often part of the patient's medical history) for determining early pregnancy?

- (3) What is the value of physical examination findings in determining the probability of early pregnancy?

ANATOMIC AND PHYSIOLOGIC ORIGINS OF THE SIGNS AND SYMPTOMS OF PREGNANCY DURING THE FIRST TRIMESTER

Pregnancy is suspected whenever a woman of childbearing years who has had regular menstrual cycles notices abrupt cessation of her menses. However, cessation of menses is a difficult symptom to evaluate in patients with previously irregular bleeding patterns. Occasionally, women have unexplained cyclic bleeding during pregnancy, especially in the first few months, and thus lack the symptom of amenorrhea. About 8% of pregnant women have a small amount of bleeding on or before the 40th day, which is thought to be related to implantation.¹

The term *morning sickness* refers to the tendency of many women (approximately 50%) to develop nausea, often with vomiting, between 6 and 12 weeks' gestational age.¹ Usually the nausea is worse when the pregnant woman awakens in the morning, whereas it tends to diminish as the day progresses.

Shortly after missing her first period, the pregnant woman may notice a heavy sensation in her breasts, accompanied by tingling and soreness. These symptoms relate to hormone stimulation of the ducts and alveoli of the breast parenchyma, but may occur in identical form just before a menstrual period. As early as 6 weeks' gestational age, there may be noticeable enlargement of the breasts, with engorgement of the superficial veins in the breasts.² During the first trimester, the nipples darken and become more sensitive. The areolar areas darken and become puffy. These symptoms and signs are thought to be of more value in primigravida because in multigravida women, areolar and nipple changes often remain from previous pregnancies.³

A few weeks after implantation (6 weeks' gestational age), distinct enlargement of the uterus may be felt on bimanual palpation. In early pregnancy, the uterus becomes softened and changes from a pear-shaped configuration to a globular contour.¹ The congestive hyperemia of the pelvis in early pregnancy is manifested by a softening of the vagina and cervix, as well as a change in the color of these tissues. A significant increase in uterine artery pulsatile activity may occur as blood flow to the pregnant uterus increases.⁴ In early pregnancy, the enlarging uterus exerts pressure on the bladder. Some patients note an increase in urinary frequency and nocturia during the first trimester.

HOW TO ELICIT THESE SYMPTOMS AND SIGNS

Medical History

Although patients may give a simple description such as "I may be pregnant," the examiner should seek a more complete medical history. Histories that indicate an increased likelihood of pregnancy include amenorrhea, morning sickness,

breast symptoms (swelling, tingling, or tenderness), sexual activity, not using or inconsistent use of contraception, patient suspects she is pregnant, and a positive home pregnancy test result. Specific questions to ask include the following: (1) When was your last menstrual period, and was it normal? (2) Do you use any form of contraception? (3) Do you have any symptoms of pregnancy? (4) Is there a chance you are pregnant?

Frequently, the patient may report, "My home pregnancy test was positive, and I want to know whether I am pregnant." Important questions regarding this type of history would be these: (1) How many days or weeks after your last menstrual period did you perform the test? (2) Did you feel comfortable performing the test? (3) Did the instructions seem complicated to you? (4) What kind of home pregnancy test did you use? (5) Did you repeat the test and get a similar result?

Physical Examination

To diagnose pregnancy, the clinician might examine the patient's breasts, as well as the vaginal wall, cervix, and uterus, by bimanual examination. The breasts may become engorged and enlarged, with darkening of the areolar area. The venous pattern over the breasts becomes increasingly visible as pregnancy progresses.⁵

Vaginal examination can be performed to elicit the Chadwick sign associated with early pregnancy. As early as 8 to 12 weeks' gestational age, the mucous membranes of the vulva, vagina, and cervix become congested and take on a bluish-violet hue (Chadwick sign).¹ This hue is especially well defined in the anterior vaginal wall but is also present to some extent throughout the vagina and on the cervix. The Chadwick sign is rarely seen before 7 weeks' gestational age.⁶

On bimanual examination, softening of the cervix (Goodell sign) may be detected by 8 weeks' gestational age.⁷ The cervix of a nonpregnant woman is fibrous and normally feels like the tip of the nose. By contrast, the progressive edema that develops during pregnancy softens the consistency of the cervix tip to approximate that of the lips (Goodell sign).

Examination of the uterus on bimanual examination can be performed to detect changes in uterine consistency and size. A palpable softening of the lowermost portion of the corpus occurs at about 6 weeks' gestational age (Hegar sign).⁷ To elicit this sign, when the uterus is anteverted, the examiner places two fingers in the anterior vaginal fornix (or the posterior fornix in the presence of a retroverted uterus) and then compresses behind the fundus at the lower uterine segment with the other hand, using suprapubic pressure (Figure 42-1). In this way, a distinct area of uterine softening is observed between 2 firmer structures: the fundus above and the cervix below.⁵ Occasionally, the softening at the isthmus is so marked that the cervix and the body of the uterus seem to be separate organs.³

Another early sign of pregnancy is the uterine artery pulsation that can be palpated on a bimanual examination.⁴ During a bimanual examination, the second and

third digits of the examining hand can be placed in the lateral vaginal fornix, and the presence of uterine artery pulsations can often be palpated with minimal pressure on the parametrium.⁴

A few weeks after the embryo has become implanted, a distinct enlargement of the uterus may be felt on bimanual examination. The uterus remains confined in the pelvis until 12 weeks' gestational age, when the fundus becomes palpable above the pubic symphysis (Figure 42-2).

The identification of the fetal heart rate distinct from the maternal heart rate establishes a diagnosis of pregnancy. Transvaginal ultrasonography can detect fetal heart activity as early as 5 weeks' gestational age, and transabdominal ultrasonography can detect this activity as early as 6 weeks' gestational age. Instruments that use the Doppler effect can detect fetal cardiac activity at 10 to 12 weeks' gestational age. The fetal heart can usually be auscultated with a fetoscope by 20 weeks' gestational age.

Reference Standard for Diagnosing Early Pregnancy

In this review, the detection of the β subunit of human chorionic gonadotropin (HCG) in urine or serum is the routine reference standard (or gold standard) for diagnosing early pregnancy. The diagnostic reliability of both the serum and urine HCG tests is comparable. The sensitivity and specificity for the diagnosis of pregnancy for both tests are between 97% and 100% when performed in the laboratory.⁸ In this review, we also report the results of studies conducted before the development of the HCG test. These earlier studies used delivery as the reference standard.

METHODS

Search Strategy

We searched the MEDLINE database for English-language articles concerning the diagnosis of pregnancy that were published between 1966 and 1996. The key words used were "pregnancy," "diagnosis," and "pregnancy tests." Additional articles listed in the bibliographies of standard obstetric texts and references cited in articles included in our study were also included among the articles considered.

Articles were systematically reviewed by authors and given a grade of A, B, or C according to the study design and level of evidence (see Table 1-7 for a summary of Evidence Grades and Levels).⁹ Articles were excluded if the results of the symptom or sign being investigated were not compared with the gold standard or the results could not be classified into a contingency table (attempts were made to reach authors of potential articles to obtain additional information needed to create contingency tables).

Through the MEDLINE, textbook reference, and bibliography searches, we initially identified 55 articles, 40 of which were rejected because the test was not compared with the gold standard (urine or serum HCG test) or a pregnancy outcome. The remaining 15 articles were then analyzed by us, and 6 more were excluded because the reported data were

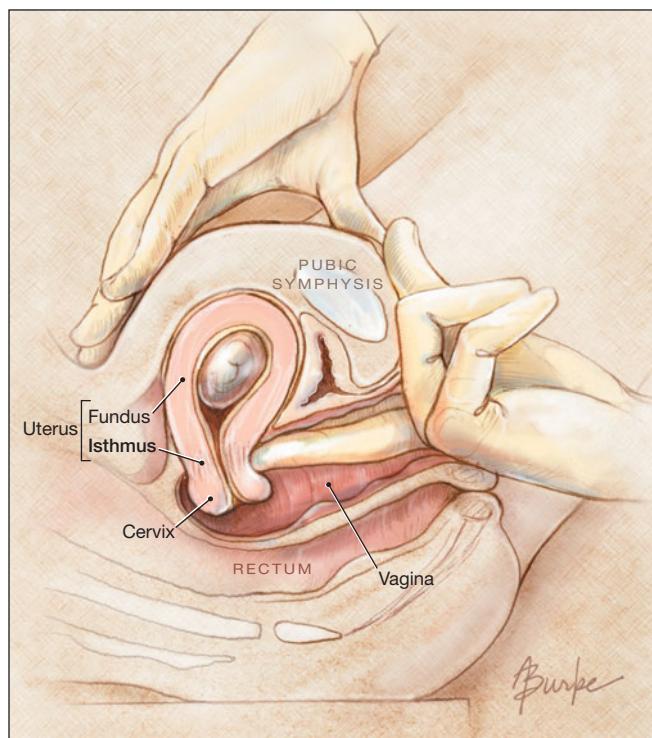


Figure 42-1 Examination Eliciting the Hegar Sign

The Hegar sign is a softening of the lower uterine segment that can be appreciated during a bimanual examination.

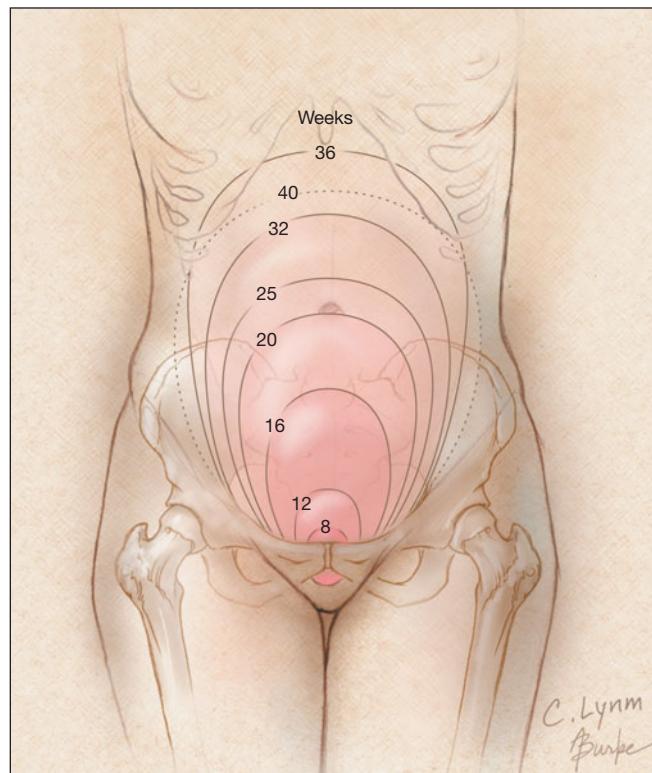


Figure 42-2 Uterine Height at Different Gestational Weeks

The height of the fundus at comparable gestational dates varies greatly among patients. Those shown are the most common. A convenient rule of thumb is that, at 20 weeks of gestation, the fundus is usually at or slightly above the umbilicus.

not sufficient to permit construction of contingency tables. Therefore, the results of 9 studies form the basis for this review.

We used data from contingency tables to calculate sensitivity and specificity. Likelihood ratios were also calculated to characterize the behavior of the diagnostic tests. The positive likelihood ratio ($LR+$) is defined as sensitivity/(1 – specificity) and expresses the change in odds favoring a disease, given a positive test result ($LR+$ values are ≥ 1), whereas the negative likelihood ratio ($LR-$) is defined as (1 – sensitivity)/specificity and expresses the change in odds favoring disease, given a negative test result ($LR-$ values are 0 to 1).¹⁰ Data were sufficiently similar in design to assess for statistical similarity. The data were pooled when the Breslow-Day test for homogeneity was not significant ($P > .05$).¹¹

Accuracy of History and Symptoms for Pregnancy Diagnosis

Several studies have been performed to evaluate the value of patient history in ruling in or ruling out early pregnancy compared with the gold standard HCG test (Tables 42-1, 42-2, 42-3, and 42-4). Among 208 consecutive patients for whom a qualitative serum HCG determination is ordered, emergency department physicians recorded the date of the patient's last menstrual period, whether her menstrual period was on time, if birth control had been used, and whether the patient suspected she was pregnant.¹² The main indication for ordering a pregnancy test in this study was abdominal pain (138 patients). Sixty-eight women (33%) were pregnant. Three historical variables were statistically

less likely to be associated with pregnancy: a last menstrual period that was on time, the patient thinking that she was not pregnant, and the patient stating that there was no chance that she could be pregnant ($P < .001$). Combinations of historical criteria were unsuccessful at ruling out pregnancy; there was still a 10% chance of pregnancy's being overlooked using any combination of these historical variables.

Women may not associate symptoms with early pregnancy. Investigators measured the effectiveness of a standardized patient history questionnaire in detecting unrecognized pregnancies.¹³ Consecutive fertile women ($n = 191$) presenting to the emergency department for any reason completed a menstrual and sexual history questionnaire and had a pregnancy test. This study reports a 6.3% prevalence of unrecognized pregnancy, defined as a "pregnancy not definitely known to exist" when the patient presented to the emergency department.¹³ Among those with abdominal pain or pelvic complaints (70 patients), the prevalence of unrecognized pregnancy was found to be 13%. Historical factors were analyzed for correlation with positive pregnancy test results. Two factors were found to be statistically significant correlates: the patient thought there was a chance she could be pregnant and an abnormal last menstrual period ($P < .001$). One factor, the delayed menstrual period, was not found to be significant ($LR+, 1.0$). Among the historical factors analyzed, "Is there any chance that you could be pregnant now?" was the most sensitive for pregnancy (92%), with a specificity of 71% (David Seaberg, MD, University of Pittsburgh, Pennsylvania, unpublished data, June 1995).

Unlike women who do not associate symptoms with early pregnancy, others self-diagnose pregnancy and request medical confirmation. Women ($n = 283$) with late menstrual periods who requested evaluation in a health center completed a structured contraception and sexual history questionnaire that included questions on whether the woman believed she was pregnant and whether subjective symptoms of pregnancy were present.¹⁴ The patient sealed her answers to the questionnaire in an envelope before the results of the pregnancy tests were available. One hundred eighteen women (42%) were pregnant. Women were better at ruling out pregnancy (sensitivity, 92%) than ruling in pregnancy (specificity, 42%).

In another study,¹⁵ general practitioners performed a study to determine the value of pregnancy symptoms (presence or absence of amenorrhea and morning sickness) in determining the probability of pregnancy. Information was collected prospectively about women who consulted their general practitioner for a diagnosis of pregnancy; the gold standard was a positive pregnancy test result. General practitioners throughout Scotland ($n = 155$) participated in the study, which was restricted to women between the ages of 16 and 45 years. Of the 1592 women enrolled, 979 (62%) were pregnant. The symptom of amenorrhea was 63% sensitive and 60% specific for pregnancy. Morning sickness as a symptom of pregnancy had a sensitivity of 39% and a specificity of 86%. This study did not ask the participants whether they thought they were pregnant.

Table 42-1 Does a Delayed Menstrual Period Predict Pregnancy?^a

Study	Evidence Grade ^b	Characteristics	Pregnant		
			Yes	No	LR (95% CI)
Robinson and Barber ¹⁵	A	Delayed menses	618	248	1.6 (1.4-1.7)
		Menses on time	361	365	0.62 (0.56-0.69)
Ramoska et al ¹²	A	Delayed menses	58	58	2.1 (1.6-2.6)
		Menses on time	10	82	0.25 (0.14-0.45)
Stengel et al ^{13c}	B	Delayed menses	3	43	1.0 (0.38-2.9)
		Menses on time	9	136	0.99 (0.70-1.4)
Zabin et al ¹⁶	A	Delayed menses	703	1078	1.1 (1.0-2.9)
		Menses on time	331	707	0.81 (0.68-0.76)

Abbreviations: CI, confidence interval; LR, likelihood ratio.

^aIn testing for homogeneity, $\chi^2 = 37$ and $P = .001$. Therefore, data were not pooled.

^bSee Table 1-7 for a summary of Evidence Grades and Levels.

^cUnpublished data from this study provided by David Seaberg, MD, University of Pittsburgh, Pennsylvania, June 1995.

In 1996, Zabin et al¹⁶ performed a similar study in a population of adolescents (younger than 17 years) to determine historical predictors of pregnancy. They performed a cross-sectional study of 2926 adolescents who presented to 52 clinics in the United States and requested a pregnancy test. The girls were asked to complete an anonymous questionnaire (98% response rate) while they waited for the results of their pregnancy test. Thirty-six percent of adolescents in this study were pregnant. A late menstrual period was the most frequent reason (63%) for the visit (for pregnancy: sensitivity, 68%; specificity, 40%).

Although a delayed menstrual period yields statistically significant results for predicting pregnancy, with an LR+ of 1.1 to 2.1 (Table 42-1), the results are inconsistent and, therefore, not a reliable symptom of pregnancy. Typical early symptoms of pregnancy provide more consistent results across studies and serve to increase slightly the likelihood of pregnancy (LR+, 2.4) (Table 42-2). Unfortunately, the absence of early symptoms of pregnancy, such as morning sickness, does not rule out pregnancy (LR-, 0.71). Likewise, the patient's use of birth control decreases the likelihood of pregnancy (LR-, 0.29), but not enough to efficiently rule it out (Table 42-3). Even the patient's suspicion of pregnancy statistically alters the likelihood of pregnancy, but not enough to be reliable (Table 42-4).

Accuracy of Home Pregnancy Tests

It has been reported that one-third of women who think they may be pregnant have used a home pregnancy test.¹⁷ A recent study of teenagers requesting pregnancy tests in health departments revealed that 28% of adolescents had used an in-home pregnancy test before their visit.¹⁶ In-home pregnancy test kits became available in 1976 and used the hemagglutination-inhibition method of detecting HCG. Currently, most test kits use monoclonal HCG antibodies, which can produce test results that can be read as a color change. The accuracy of these tests is claimed to be 97% to 99% by the manufacturers.¹⁸ Studies have shown that accuracy depends on several factors, such as whether the woman read the instructions carefully and the number of days beyond the missed menstrual period.¹⁹

In 1986, Doshi²⁰ published a study measuring the accuracy of 3 in-home tests for early pregnancy. The author studied 109 women of childbearing age whose menses were late by at least 6 days, but not more than 20 days. Volunteers for the study were obtained from 3 sites; the majority were white and educated. Participants brought to the study site their first morning urine, which was then divided in half. One portion of the sample was returned to the participant to use in performing a pregnancy test at home. Using 1 of 3 study kits (Answer [Carter Products; Carter-Wallace, Inc, New York, New York]; Daisy 2 [Boehringer-Mannheim Corp, Ingelheim, Germany]; and e.p.t. [Warner-Lambert Co, Morris Plains, New Jersey]), the participants were instructed to follow the package directions in performing the test, call the site with results, and complete and return the data collection survey to the investigator. The investigator performed an identical test using the other portion of the urine sample. Despite

Table 42-2 Probability of Pregnancy if Patient Reports Symptoms of Pregnancy^a

Study	Evidence Grade ^b		Pregnant			LR (95% CI)
			Yes	No		
Robinson and Barber ¹⁵	A	Morning sickness	380	88	2.7 (2.2-3.3)	
		No morning sickness	599	525	0.71 (0.67-0.76)	
Bachman ¹⁴	A	Any pregnancy symptoms	59	34	2.4 (1.7-3.4)	
		No pregnancy symptoms	59	131	0.63 (0.52-0.77)	

Abbreviations: CI, confidence interval; LR, likelihood ratio.

^aPregnancy symptoms defined as morning sickness, breast tenderness and fullness, urinary frequency, or fatigue.

^bSee Table 1-7 for a summary of Evidence Grades and Levels.

Table 42-3 Probability of Pregnancy if Patient Reports Not Using Birth Control

Study	Evidence Grade ^a		Pregnant			LR (95% CI)
			Yes	No		
Ramoska et al ¹²	A	No birth control	61	96	1.3 (1.1-1.5)	
		Birth control	7	44	0.33 (0.16-0.69)	
Stengel et al ^{13,b}	B	No birth control	9	88	1.5 (1.1-2.2)	
		Birth control	3	91	0.49 (0.18-1.3)	
Pooled ^c		No birth control	70	184	1.5 (1.3-1.7)	
		Birth control	10	135	0.29 (0.16-0.53)	

Abbreviations: CI, confidence interval; LR, likelihood ratio.

^aSee Table 1-7 for a summary of Evidence Grades and Levels.

^bUnpublished data from this study provided by David Seaberg, MD, University of Pittsburgh, Pennsylvania, June 1995.

^cIn testing for homogeneity, $\chi^2 = 0.097$ and $P = .76$. Therefore, data were pooled.

Table 42-4 Probability of Pregnancy if Patient Thinks There Is a Chance She Is Pregnant

Study	Evidence Grade ^a	Patient Thinks She Is	Pregnant			LR (95% CI)
			Yes	No		
Bachman ¹⁴	A	Pregnant	109	95	1.6 (1.4-1.8)	
		Not pregnant	9	70	0.18 (0.09-0.34)	
Ramoska et al ¹²	A	Pregnant	58	63	1.9 (1.5-2.3)	
		Not pregnant	10	77	0.27 (0.15-0.48)	
Stengel et al ^{13,b}	B	Pregnant	11	52	3.2 (2.4-4.2)	
		Not pregnant	1	127	0.12 (0.02-0.77)	
Zabin et al ¹⁶	A	Pregnant	789	640	2.1 (2.0-2.3)	
		Not pregnant	254	1148	0.38 (0.34-0.42)	
Pooled results ^c		Pregnant	967	850	2.1 (2.0-2.2)	
		Not pregnant	270	1422	0.35 (0.31-0.39)	

Abbreviations: CI, confidence interval; LR, likelihood ratio.

^aSee Table 1-7 for a summary of Evidence Grades and Levels.

^bUnpublished data from this study provided by David Seaberg, MD, University of Pittsburgh, Pennsylvania, June 1995.

^cIn testing for homogeneity, $\chi^2 = 4.3$ and $P = .23$. Therefore, data were pooled.

manufacturer claims of 97% overall accuracy for the test kits used, the investigator found an accuracy of 77%. The participants had a sensitivity of 80% and specificity of 68% for detecting early pregnancy with the home pregnancy tests ($LR+$, 2.5; $LR-$, 0.29), with similar diagnostic efficiency observed for all 3 kits. These results concerned Doshi²⁰ because of missed opportunities for early prenatal care and the postponement of discontinuing teratogenic substances.

In 1993, investigators from France published an extensive analysis of the reliability and feasibility of home pregnancy tests.²¹ They looked at 27 different test kits (manufacturers were not identified) and selected 11 kits for the study, which were found to have a 100% sensitivity and specificity under ideal laboratory conditions. Laywomen volunteers (aged 14–49 years; $n = 638$) were asked to test a home-use test kit for pregnancy using a coded urine specimen. They also were asked to complete a questionnaire after they performed the test. The results of the diagnostic study showed that 5 of the 11 kits had 100% specificity; the others had specificity values between 77% and 94%. Two kits had a high diagnostic sensitivity (>90%), and 2 kits were found to have a low diagnostic sensitivity (<10%). Whereas 90% of the participants claimed that the test was easy to perform, of the 478 positive (result positive for pregnancy) urine samples distributed, 230 were falsely interpreted as negative (sensitivity, 48%). The authors concluded that the main reason for the poor performance was difficulty in interpreting the instructions rather than the socioeconomic situation of the participants.

Table 42-5 Probability of Pregnancy if Physician Examination Findings Present

Study	Evidence Grade ^a	Characteristic	Pregnant		
			Yes	No	LR (95% CI)
Chadwick ⁶	C	Chadwick sign			
		Present	144	1	29 (4.1-200)
		Absent	137	55	0.50 (0.44-0.56)
Robinson and Barber ¹⁵	A	Breast signs			
		Present ^b	549	127	2.7 (2.3-3.2)
		Absent	430	486	0.55 (0.50-0.60)
Robinson and Barber ¹⁵	A	Vaginal examination signs			
		Present ^c	172	34	3.2 (2.2-4.5)
		Absent	807	579	0.87 (0.84-0.90)
Robinson and Barber ¹⁵	A	Palpable fundus			
		Present	84	19	2.8 (1.7-4.5)
		Absent	895	594	0.94 (0.92-0.97)
Meeks et al ⁴	B	Uterine artery pulsation			
		Present	19	9	11 (5.6-21)
		Absent	6	121	0.26 (0.13-0.52)

Abbreviations: CI, confidence interval; LR, likelihood ratio.

^aSee Table 1-7 for a summary of Evidence Grades and Levels.

^bBreast signs were not explicitly defined and include any abnormal findings on breast examination.

^cVaginal examination signs were not explicitly defined and include any abnormal findings on vaginal or pelvic examination.

Accuracy of the Physical Examination

Only a few studies have analyzed at the accuracy of the physical examination for pregnancy. Unfortunately, no studies have examined interobserver or intraobserver reliability. In 1887, Chadwick⁶ published a study of 337 women evaluated weekly (until delivery for those women who were pregnant) to assess the presence of the Chadwick sign. He described the coloration of the vaginal wall as no color or doubtful color, suggestive color, characteristic color, and general deep color. He classified any vaginal wall with characteristic or general deep color to be “diagnostic.” With his criteria, the sensitivity of this physical sign is 51% and the specificity is 98%. No validation studies could be found.

Robinson and Barber¹⁵ performed a study in 1977 to determine the value and reliability of the physical examination for pregnancy compared with a pregnancy test. They examined the vagina for signs of pregnancy, palpated the fundus, and assessed breast changes on physical examination. The most common feature observed was breast signs (42%), with a sensitivity of 56% and a specificity of 79%. Thirteen percent of women were observed to have “signs” on vaginal examination (signs not specified, but presumably some combination of the Goodell, Hegar, and Chadwick signs) consistent with pregnancy, with a sensitivity of 18% and a specificity of 94%. Last, 6% of women were observed to have a palpable fundus at presentation for a pregnancy test (sensitivity, 9%; specificity, 97%).

Recently, a study was performed to determine whether palpable uterine artery pulsation is a reliable clinical indicator of early pregnancy.⁴ The authors conducted the study in 2 phases. During the first phase, one of the authors examined 299 women who were less than 6 weeks from their last menstrual period for palpable uterine artery pulsation; this examination was conducted after a medical history had been obtained, and thus the examiner was not blind to the clinical situation. During the second phase, one of the authors examined 155 women for palpable uterine artery pulsation but performed only the bimanual examination and was blind to all other historical and physical examination data. With data from the second phase only, palpation of uterine artery pulsations may be a valuable tool in diagnosing early pregnancy (sensitivity, 76%; specificity, 93%). According to the results of this study, physicians were encouraged to add uterine artery pulsation to their clinical examination in diagnosing early pregnancy.

Despite descriptive articles dating back to the 1880s, no studies could be identified that measured the value of the Goodell or Hegar signs. In 1908, McDonald⁷ reported the prevalence of early pregnancy findings in 100 women known to be pregnant. He followed up women with weekly pelvic examinations during their first trimester. In this descriptive study, pregnant women were found to have the following: Hegar sign, 94%; Goodell sign, 66%; and Chadwick sign, 61%. This study is included for historical interest. Knowing the pregnancy status of patients creates expectation bias that probably overstates the value and prevalence of these signs.

As summarized in **Table 42-5**, several physical findings significantly increase the likelihood of pregnancy. The most useful findings on physical examination for making the diagnosis of early pregnancy appear to be Chadwick sign (LR+, 29) and palpable uterine artery pulsation (LR+, 11), although validation studies are needed because these 2 studies had comparatively lower methodologic quality scores. Unfortunately, if any of these signs are absent, this does not rule out pregnancy.

THE BOTTOM LINE

Clearly, to establish a diagnosis of early pregnancy, a clinician should order a urine or serum HCG test. However, there may be circumstances in which it would be useful for patients or physicians to know the value of pregnancy symptoms, home pregnancy test results, and physical examination findings for the diagnosis of pregnancy.

We can predict the likelihood of pregnancy for the patients in the clinical scenarios. For case 1, the woman with sinusitis has a prior probability of pregnancy of about 5%. Because she reports that her menses was on time (LR-, 0.62) and states that she is not pregnant (LR-, 0.35), the calculated probability of pregnancy might be from 1.7% to 3.1% for this patient. We would not order a pregnancy test for case 1. For case 2, the sexually active teenager, we can also calculate a probability that she might be pregnant. Zabin et al¹⁶ reported a pregnancy rate of 36% among teenagers presenting for a pregnancy test in their study. If we assume her prior probability of pregnancy is 36% and know her menses is late (LR+, 1.1), her home pregnancy test result was negative (LR-, 0.29), and her pelvic examination findings were normal (LR-, 0.87), her probability of pregnancy ranges from 10% to 41%, and we would recommend ordering a pregnancy test for this case. For case 3, the 41-year-old woman with a late menses and breast tenderness, the prior probability of pregnancy might be low (approximately 2%) because of decreased fecundity secondary to her age. If we consider her late menses (LR+, 1.6) and her breast tenderness (LR+, 2.4), her probability of pregnancy has increased approximately 2-fold to a range of 3.1% to 4.9%, and we would order a pregnancy test.

Patients may call their clinician asking for advice regarding a late period or symptoms of pregnancy. They may want to know whether they should perform a home pregnancy test, or they may request assistance in interpreting the test results. Evidence suggests that some historical features, when absent, are fair but not reliable for ruling out pregnancy. When diagnosing pregnancy, the patient or clinician should not rely on symptoms and signs of pregnancy or a home pregnancy test; a laboratory test should be requested.

Author Affiliations at the Time of the Original Publication

Departments of Internal Medicine, Obstetrics and Gynecology, Duke University Medical Center, and Comprehensive

Health Center and the Center for Health Services Research in Primary Care, Veterans Affairs Medical Center, Durham, North Carolina (Dr Bastian); and the Department of Obstetrics and Gynecology, Duke University Medical Center, Durham, North Carolina (Dr Piscitelli).

Acknowledgment

We thank David Seaberg, MD, University of Pittsburgh, Pennsylvania, for providing the unpublished data necessary to calculate sensitivity and specificity for the study by Stengel et al.¹³

REFERENCES

1. Scott JR, Disaia PJ, Hammond CB, Spellacy WN. *Danforth's Obstetrics and Gynecology*. 7th ed. Philadelphia, PA: JB Lippincott; 1994.
2. Hibbard BM. Pregnancy diagnosis. *BMJ*. 1971;1(5749):593-595.
3. Cunningham FG, MacDonald PC, Gant NF. *Williams Obstetrics*. 19th ed. East Norwalk, CT: Appleton & Lange; 1993.
4. Meeks GR, Casare CD, Bates GW. Palpable uterine artery pulsation as a clinical indicator of early pregnancy. *J Reprod Med*. 1995;40:194-196.
5. Bates B, Bickley LS, Hoekelman RA, Thompson JE. *A Guide to Physical Examination and History Taking*. 6th ed. Philadelphia, PA: JB Lippincott; 1995.
6. Chadwick JR. Value of the bluish coloration of the vaginal entrance as a sign of pregnancy. *Trans Am Gynecol Soc* (1886). 1887;11:399-418.
7. McDonald E. The diagnosis of early pregnancy, with report of one hundred cases and special reference to the sign of flexibility of the isthmus of the uterus. *Am J Obstet Dis Women Child*. 1908;57:323-346.
8. O'Connor RE, Bibro CM, Pegg PJ, Bouzoukis JK. The comparative sensitivity and specificity of serum and urine HCG determinations in the ED. *Am J Emerg Med*. 1993;11(4):434-436.
9. Holleman DR, Simel DL. Does the clinical examination predict pulmonary airflow obstruction [published correction appears in *JAMA*. 1995;273(4):1334]? *JAMA*. 1995;273:313-319.
10. Simel DL, Samsa GP, Matchar DB. Likelihood ratios with confidence: sample size estimation for diagnostic test studies. *J Clin Epidemiol*. 1991;44(8):763-770.
11. Stokes ME, Davis CS, Koch GG. *Categorical Data Analysis Using the SAS System*. Cary, NC: SAS Institute Inc; 1995.
12. Ramoska EA, Sacchetti AD, Nepp M. Reliability of patient history in determining the possibility of pregnancy. *Ann Emerg Med*. 1989;18(1):48-50.
13. Stengel CL, Seaberg DC, Macleod BA. Pregnancy in the emergency department: risk factors and prevalence among all women. *Ann Emerg Med*. 1994;24(4):697-700.
14. Bachman GA. Myth or fact: can women self-diagnose pregnancy? *J Med Soc N J*. 1984;81(10):857-858.
15. Robinson ET, Barber JH. Early diagnosis of pregnancy in general practice. *J R Coll Gen Pract*. 1977;27(179):335-338.
16. Zabin LS, Emerson MR, Ringers PA, Sedivy V. Adolescents with negative pregnancy test results: an accessible at-risk group. *JAMA*. 1996;275(2):113-117.
17. Jeng LL, Moore RM, Kaczmarek RG, Placek PJ, Bright RA. How frequently are home pregnancy tests used? results from the 1988 National Maternal and Infant Health Survey. *Birth*. 1991;18(1):11-13.
18. Caiola SM. The pharmacist and home-use pregnancy tests. *Am Pharm*. 1992;32(1):57-60.
19. Latman NS, Bruot BC. Evaluation of home pregnancy test kits. *Biomed Instrum Technol*. 1989;23(2):144-149.
20. Doshi ML. Accuracy of consumer-performed in-home tests for early pregnancy detection. *Am J Public Health*. 1986;76(5):512-514.
21. Daviaud J, Fournet D, Ballongue C, et al. Reliability and feasibility of pregnancy home-use tests: laboratory validation and diagnostic evaluation by 638 volunteers. *Clin Chem*. 1993;39(1):53-59.

This page intentionally left blank

Prepared by Lori A. Bastian, MD, MPH

Reviewed by Joanne T. Piscitelli, MD

CLINICAL SCENARIO

A 16-year-old adolescent who is concerned she might be pregnant calls her local Planned Parenthood clinic. She has not noticed any symptoms of pregnancy such as early morning nausea or breast tenderness. She does observe that her period is 3 weeks overdue. She purchased a home pregnancy test (HPT) kit, and the results suggested that she is not pregnant. When asked about performing the HPT kit, she observes she felt nervous and was not sure that she followed the directions correctly.

UPDATED SUMMARY ON PREGNANCY

Original Review

Bastian LA, Piscitelli J. Is this patient pregnant? can you reliably rule in or rule out early pregnancy by clinical examination? *JAMA*. 1997;278(7):586–591.

UPDATED LITERATURE SEARCH

Our literature search used the parent search strategy for The Rational Clinical Examination series combined with the subjects “pregnancy” and “pregnancy tests,” published in English between 1996 and September 2004. The results yielded 301 titles, for which we reviewed the titles and abstracts; 12 articles were selected for additional review. These articles were reviewed to identify studies that assessed the sensitivity and specificity of the medical history or physical examination features of pregnancy. Only 1 article, a meta-analysis on the diagnostic characteristics of HPT kits, was retained.¹ We included home testing, as in the original review, because it is frequently part of the patient’s medical history when evaluating for pregnancy.

IMPROVEMENTS IN THE DATA PRESENTED IN THE ORIGINAL PUBLICATION

None.

CHANGES IN THE REFERENCE STANDARD

There are no changes observed in the reference standard, which is based on laboratory testing of serum or urine human chorionic gonadotropin (HCG). Recently, Wilcox et al² used an extremely sensitive assay for HCG and found that 10% of pregnancies were undetectable on the first day of the missed period. These authors recommend waiting 1 week after the first day of the missed period to perform pregnancy testing.

RESULTS OF LITERATURE REVIEW

Home pregnancy kits have become increasingly popular and manufacturers claim these HPT kits are 99% accurate. Most studies have found that women choose to use HPT kits because of the speed of obtaining results and the convenience of testing at home. A systematic review of 5 studies reviewing 16 HPT kits found that the diagnostic performance of these kits is affected by the characteristics of the users (Table 42-6). In studies in which urine samples obtained by the investigators were tested by volunteers, sensitivity was 91% (95% confidence interval [CI], 84%-96%). However, the sensitivity was less in studies in which subjects were actual patients who used the HPT kit on their own urine samples (sensitivity, 75%; 95% CI, 64%-85%).

EVIDENCE FROM GUIDELINES

None.

CLINICAL SCENARIO—RESOLUTION

For diagnosing pregnancy, you recommend repeating the HPT kit 1 week after using the first kit. If the results remain negative, she should still present to the clinic for further testing because most physicians would recommend that this teenager be screened for sexually transmitted infections and counseled about contraception, independent of the HPT’s result.

Table 42-6 Likelihood Ratios of Commercially Available Home Pregnancy Test Kits^a

Test	Sensitivity	Specificity	LR+ (95% CI)	LR- (95% CI)
Acu Test ³	0.52	0.89	4.7 (1.5-14)	0.54 (0.36-0.81)
Advance ⁴	0.86	0.91	9.7 (3.8-25)	0.15 (0.10-0.22)
Advance ³	0.91	1.0	53 (3.4-830)	0.12 (0.04-0.32)
Answer 2 ³	1.0	0.94	13 (3.9-13)	0.02 (0-0.26)
Answer ⁵	0.78	0.64	2.2 (1.3-3.7)	0.34 (0.18-0.62)
Daisy 2 ⁵	0.82	0.64	2.3 (1.1-4.8)	0.28 (0.11-0.75)
Daisy 2 ³	0.98	0.61	2.5 (1.7-3.6)	0.04 (0.01-0.28)
e.p.t. ⁵	0.82	0.75	2.3 (1.3-4.2)	0.28 (0.15-0.55)
e.p.t. ³	0.88	1.0	53 (3.4-832)	0.15 (0.06-0.17)
e.p.t. plus ⁴	0.90	0.92	13 (4.4-40)	0.10 (0.06-0.17)
e.p.t. plus ³	0.95	1.0	63 (4.0-988)	0.07 (0.02-0.25)
Fact ³	1.0	0.94	14 (4.2-46)	0.01 (0-0.23)
First ³	0.93	1.0	47 (3.0-744)	0.10 (0.03-0.32)
Predictor ⁶	0.97	0.96	22 (8.4-57)	0.03 (0.01-0.10)
Predictor ³	1.0	0.77	4.3 (2.3-8.1)	0.02 (0-0.31)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

^aInformation on products is shown as reported by Bastian et al.¹ Testing kits may subsequently undergo changes in name, undergo product updates, or be sold between manufacturers. Acu Test: J.B. Williams Co, Cranford, New Jersey. Advance and Fact: Advanced Care Products, Ortho Pharmaceutical Corp, Raritan, New Jersey. Answer 2, Answer, and First: Carter Products, Carter-Wallace, Inc, New York, New York. Daisy 2: Boehringer-Mannheim Corp, Ingelheim, Germany. e.p.t. And e.p.t. Plus: Warner-Lambert, Morris Plains, New Jersey. Predictor: Whitehall Laboratories, New York, New York.

REFERENCES FOR THE UPDATE

- Bastian LA, Nanda K, Hasselblad V, Simel DL. Diagnostic efficiency of home pregnancy test kits: a meta-analysis. *Arch Fam Med*. 1998;7(5):465-469.^a
- Wilcox AJ, Baird DD, Dunson D, McChesney R, Weinberg CR. Natural limits of pregnancy testing in relation to the expected menstrual period. *JAMA*. 2001;286(14):1759-1761.
- Latman NS, Bruot BC. Evaluation of home pregnancy test kits. *Biomed Instrum Technol*. 1989;23(2):144-149.
- Hicks JM, Losensohn M. Reliability of home pregnancy-test kits in the hands of laypersons. *N Engl J Med*. 1989;320(5):320-321.
- Doshi ML. Accuracy of consumer performed in-home tests for early pregnancy detection. *Am J Public Health*. 1986;76(5):512-514.
- Arends J, Uldall A. The reliability of pregnancy tests performed by trained and untrained persons. *Dan Med Bull*. 1977;24(3):115-116.
- Stengel CL, Seaberg DC, MacLeod BA. Pregnancy in the emergency department: risk factors and prevalence among all women. *Ann Emerg Med*. 1994;24(4):697-700.
- Cohen DJ, Bruce KE. Sex and mortality: real risk and perceived vulnerability. *J Sex Res*. 1997;34(3):279-291.

^aFor the Evidence to Support the Update for this topic,
see <http://www.JAMAevidence.com>.

EARLY PREGNANCY—MAKE THE DIAGNOSIS

PRIOR PROBABILITY

The probability of pregnancy varies, depending on the clinical situation. In the emergency department, the prevalence of unsuspected pregnancy is 6.3%. The prevalence increased to 13% in women with abdominal or pelvic complaints.⁷ Among women trying to get pregnant, the probability of pregnancy after a single episode of unprotected sexual intercourse approximates 20% to 33%.⁸

POPULATION FOR WHOM PREGNANCY SHOULD BE CONSIDERED

- All women of childbearing years with an intact uterus who are sexually active and who have missed their last menstrual period or had an abnormal menstrual period.
- Any woman who wonders whether she might be pregnant.

DETECTING THE LIKELIHOOD OF PREGNANCY

General symptoms of early pregnancy include amenorrhea, morning sickness, and tender or tingling breasts. In the original review, the range of likelihood ratios (LRs) for women reporting a delayed menses was 1.0 to 2.1 and 0.25 to 0.99 for women reporting their menses on time. For women reporting morning sickness or any pregnancy symptoms, the LR was 2.7 or 2.4, respectively. Another indicator of early pregnancy is whether the woman thinks she is pregnant. When a woman thinks there is a chance she is pregnant, the LR for pregnancy is 2.1 (95% CI, 2.0-2.2); if she thinks she is not pregnant, the LR is 0.35 (95% CI, 0.31-0.39). Physical examination findings, such as an enlarged uterus with a soft cervix or a palpable uterine artery, have been studied and may be useful in some clinical settings.

REFERENCE STANDARD TEST

To establish a diagnosis of early pregnancy, a clinician should order a urine or serum HCG test.

EVIDENCE TO SUPPORT THE UPDATE:

42

Early Pregnancy

TITLE Diagnostic Efficiency of Home Pregnancy Test Kits: A Meta-analysis.

AUTHORS Bastian LA, Nanda K, Hasselblad V, Simel DL.

CITATION *Arch Fam Med.* 1998;7(5):465-469.

QUESTION What are the diagnostic characteristics of home pregnancy test kits?

DESIGN A systematic literature search of studies that compared home pregnancy test kits with laboratory testing of human chorionic gonadotropin using MEDLINE from 1966-1996. Two investigators extracted data independently. Five studies evaluating 16 home pregnancy test kits met the inclusion criteria.

MAIN OUTCOME MEASURES

Sensitivity, specificity, and effectiveness scores. The data were used to calculate the likelihood ratios (not reported in the original publication).

MAIN RESULTS

The authors found data for 11 different home testing kits. We dropped information on the OVA II because complete information on the manufacturer was not available,⁴ leaving the 10 shown in Table 42-7.

CONCLUSIONS

LEVEL OF EVIDENCE Systematic review.

STRENGTHS Comprehensive review of home pregnancy test (HPT) kits that were described in the original Rational Clinical Examination article.

LIMITATIONS Most studies published in 1970s and 1980s after HPT kits came on market. The most recent study was published in 1989. HPT kits currently on the market have not been reviewed.

The effectiveness of home pregnancy testing kits is dependent on the skill of the user. When taking the history from a

woman who has used a testing kit, you should confirm that she repeated the results. Newer kits can be accurate when

Table 42-7 Likelihood Ratios of Commercially Available Home Pregnancy Test Kits^a

Test	Sensitivity	Specificity	LR+ (95% CI)	LR- (95% CI)
Acu Test ¹	0.52	0.89	4.7 (1.5-14)	0.54 (0.36-0.81)
Advance ²	0.86	0.91	9.7 (3.8-25)	0.15 (0.10-0.22)
Advance ¹	0.91	1.0	53 (3.4-830)	0.12 (0.04-0.32)
Answer 2 ¹	1.0	0.94	13 (3.9-13)	0.02 (0-0.26)
Answer ³	0.78	0.64	2.2 (1.3-3.7)	0.34 (0.18-0.62)
Daisy 2 ³	0.82	0.64	2.3 (1.1-4.8)	0.28 (0.11-0.75)
Daisy 2 ¹	0.98	0.61	2.5 (1.7-3.6)	0.04 (0.01-0.28)
e.p.t. ³	0.82	0.75	2.3 (1.3-4.2)	0.28 (0.15-0.55)
e.p.t. ¹	0.88	1.0	53 (3.4-832)	0.15 (0.06-0.17)
e.p.t. Plus ²	0.90	0.92	13 (4.4-40)	0.10 (0.06-0.17)
e.p.t. Plus ¹	0.95	1.0	63 (4.0-988)	0.07 (0.02-0.25)
Fact ¹	1.0	0.94	14 (4.2-46)	0.01 (0-0.23)
First ¹	0.93	1.0	47 (3.0-744)	0.10 (0.03-0.32)
Predictor ⁵	0.97	0.96	22 (8.4-57)	0.03 (0.01-0.10)
Predictor ¹	1.0	0.77	4.3 (2.3-8.1)	0.02 (0-0.31)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

^aInformation on products is shown as originally reported in this systematic review. Testing kits may subsequently undergo changes in name, undergo product updates, or be sold between manufacturers. Acu Test: J.B. Williams Co, Cranford, New Jersey. Advance and Fact: Advanced Care Products, Ortho Pharmaceutical Corp, Raritan, New Jersey. Answer 2, Answer, and First: Carter Products, Carter-Wallace, Inc, New York, New York. Daisy 2: Boehringer-Mannheim Corp, Ingelheim, Germany. e.p.t. And e.p.t. Plus: Warner-Lambert, Morris Plains, New Jersey. Predictor: Whitehall Laboratories, New York, New York.

performed according to the manufacturer's specifications. Most physicians (and patients) may be unaware that a negative home pregnancy test result is not perfect for ruling out pregnancy.

Reviewed by Lori A. Bastian, MD

REFERENCES FOR THE EVIDENCE

1. Latman NS, Bruot BC. Evaluation of home pregnancy test kits. *Biomed Instrum Technol.* 1989;23(2):144-149.
2. Hicks JM, Iosefsohn M. Reliability of home pregnancy-test kits in the hands of laypersons. *N Engl J Med.* 1989;320(5):320-321.
3. Doshi ML. Accuracy of consumer performed in-home tests for early pregnancy detection. *Am J Public Health.* 1986;76(5):512-514.
4. Baker LD, Yert LW, Chase MC, Dale E. Evaluation of a "do-it-yourself" pregnancy test. *Am J Public Health.* 1976;66(2):166-167.
5. Arends J, Uldall A. The reliability of pregnancy tests performed by trained and untrained persons. *Dan Med Bull.* 1977;24(3):115-116.

CHAPTER 43

Does This Patient Have Pulmonary Embolism?

Sanjeev D. Chunilal, MB ChB, FRACP, FRCPA

John W. Eikelboom, MBBS, MSc, FRACP, FRCPA

John Attia, MD, PhD, FRCPC

Massimo Miniati, MD

Akbar A. Panju, MBChB, FRCPC

David L. Simel, MD, MHS

Jeffrey S. Ginsberg, MD, FRCPC

CLINICAL SCENARIOS

Do These Patients Have Pulmonary Embolism?

CASE 1 A 28-year-old woman with recently diagnosed systemic lupus erythematosus presents with 2 days of pleuritic chest pain and breathlessness. She has no leg symptoms and no personal or family history of venous thromboembolism. She is taking a second-generation oral contraceptive pill. Examination reveals a finding of mild tachypnea (20/min) and minimal tenderness over the right lateral chest wall. Examination finding of the legs is normal, and a red blood cell agglutination D-dimer test shows a negative result.

CASE 2 A 78-year-old man presents with 3 days of worsening pleuritic chest pain and breathlessness. He was discharged from the hospital 2 weeks earlier after a 14-day admission with acute cholecystitis. Surgery was not performed. His medical history includes 2 episodes of idiopathic, right-leg, deep vein thrombosis. He has controlled hypertension and previous left ventricular failure. The examination reveals tachypnea (20/min) but findings are otherwise normal. Chest radiograph and electrocardiogram (ECG) findings are normal, and a red blood cell agglutination D-dimer test shows a negative result.

Background

Pulmonary embolism occurs in 1 to 2 persons per 1000 annually in the United States.^{1,2} If untreated, it is associated with a high mortality rate, but anticoagulant therapy is highly effective in reducing mortality.^{3,4} The diagnosis of pulmonary embolism is difficult because of the wide spectrum of symptoms and signs, and most patients with suggestive symptoms do not have the disease.⁵ Typically, patients with proven pulmonary embolism present with dyspnea or acute chest pain and less frequently with cough, hemoptysis, or fainting.^{6,7} These findings often occur in association with well-defined risk factors, such as lower limb surgery or immobility (Table 43-1).⁸⁻¹⁰ Frequent findings on examination include tachycardia, tachypnea, and an accentuated pulmonary component of the second heart sound (S₂). Other features such as jugular venous distention, S₃ or S₄ (third or fourth heart sound), an audible systolic murmur at the left sternal edge, and hepatomegaly infrequently are present and may reflect right-sided ventricular compromise.

Results of arterial blood gas analysis commonly show hypoxia and hypocapnia. Chest radiography results are non-specific, and common findings include an elevated hemidiaphragm, unilateral pleural effusion, and platelike atelectasis; radiography is useful because it will sometimes provide an alternative diagnosis (eg, pneumothorax). Similarly, ECG findings are nonspecific and may show T-wave inversion across

Table 43-1 Risk Factors for Venous Thromboembolism

Risk Factors ⁸⁻¹⁰	OR (95% CI)
Surgery	21 (9.4-50)
Trauma	13 (4.1-40)
Immobility (hospital or nursing home)	8.0 (4.5-14)
Cancer	
With chemotherapy	6.5 (2.1-20)
Without chemotherapy	4.1 (1.9-8.5)
Neurologic disease with lower-extremity paresis	3.0 (1.3-7.4)
Oral contraceptive pill ¹⁰	3.0 (2.6-3.4) ^a
Hormone therapy ^b	2.7 (1.4-5.0) ^b

Abbreviations: CI, confidence interval; OR, odds ratio.

^aRelative risk from case-control studies.

^bRelative hazard.

precordial leads, the S₁Q₃/S₁Q₃T₃ pattern, or a right-sided bundle-branch block.^{6,7} Thus, although the above findings are observed in patients with objectively diagnosed pulmonary embolism, they also are common in patients without pulmonary embolism and lack specificity when considered individually. On the other hand, pulmonary embolism is uncommon in the absence of acute or worsening breathlessness or chest pain.^{6,7} Because anticoagulant therapy reduces mortality from pulmonary embolism, the threshold for considering the diagnosis should be low.³ We believe that pulmonary embolism should at least be considered whenever a patient presents with any of the above symptoms or symptom complexes, particularly in the presence of known risk factors or when there is no clear alternative.

Before the development of accurate diagnostic testing, the diagnosis of pulmonary embolism largely was based on clinical history and examination findings. Unfortunately, the clinical evaluation alone proved inaccurate in diagnosing and excluding pulmonary embolism^{7,11-13} and was virtually abandoned in the evaluation of patients with suspected pulmonary embolism. Lung scanning became routine in the 1980s and was shown to be clinically useful.⁵ However, lung scanning proved to be less than optimal because more than half of patients with suspected pulmonary embolism had nondiagnostic lung scan results and the prevalence of pulmonary embolism in such patients was approximately 25%.⁵

Once clinicians raise the possibility of pulmonary embolism, they can further define the clinical likelihood of pulmonary embolism into a pretest probability. Rather than definitively diagnosing or excluding pulmonary embolism, pretest probability assessment categorizes patients into subgroups, such as low, intermediate, and high, with ascending order of prevalences of pulmonary embolism. The potential for clinical assessment of the pretest probability to significantly influence the posttest probability of pulmonary embolism was demonstrated in the Prospective Investigation of Pulmonary Embolism Diagnosis (PIOPED) study⁵ and was confirmed in a later study by Wells et al.¹⁴ When the participating clinicians in the PIOPED study used clinical judgment

to categorize patients into low-, moderate-, or high-pretest-probability subgroups for pulmonary embolism, a moderate correlation with disease prevalence was found (9%, 30%, and 68%, respectively). In addition, in patients with a low pretest probability and a high-probability lung scan result, only about 50% had pulmonary embolism, whereas in those with a moderate or high pretest probability and a high-probability lung scan result, more than 90% had pulmonary embolism.⁵

According to the medical history and physical examination findings, clinical prediction rules that assess pretest probability for deep vein thrombosis, a closely related condition to pulmonary embolism, have been developed and shown to simplify the diagnosis.^{15,16} For example, the safety of withholding anticoagulant therapy, without additional testing, has been demonstrated in patients with a low¹⁷ or low/moderate¹⁸ pretest probability for deep vein thrombosis and a negative D-dimer test result. D-dimer is a plasmin-derived fibrin degradation product that is highly sensitive for deep vein thrombosis and pulmonary embolism.¹⁹ Elevated levels of D-dimer are observed in most patients with pulmonary embolism and deep vein thrombosis, but because the available assays have moderate specificity (30%-75%), they also show elevated results in patients with non-thrombotic disorders.¹⁹ We postulated that assessment of pretest probability of pulmonary embolism also might be useful in simplifying the diagnosis of this condition.

The objectives of this article are 2-fold: (1) to determine whether, according to their clinical impression after collecting routine data (the clinical gestalt), experienced clinicians can accurately group patients into strata distinguished by an increasing probability of pulmonary embolism; and (2) to determine whether clinical prediction rules are useful in determining the pretest probability for pulmonary embolism. For the first objective, the examiner estimates the probability of pulmonary embolism according to his or her clinical gestalt. Each examiner values the information differently in quantifying an overall impression. For the second objective, clinical prediction rules rely on an explicit prespecified list of data items, each of which is assigned a score.

METHODS

Data Sources

We searched the MEDLINE electronic database for English-language articles published between 1966 and March 2003, using the following Medical Subject Headings: "pulmonary embolism," "prospective studies," "EXP" (explode) "sensitivity and specificity," "EXP probability" and "EXP models," and "statistical." We identified studies in which clinical assessment of patients with suspected pulmonary embolism was performed routinely. The reference lists of identified articles also were examined for additional studies missed by the MEDLINE search.

Study Selection and Data Extraction

Three independent reviewers (S.D.C., J.W.E., J.A.) identified potentially eligible articles, and a senior reviewer (J.S.G.)

resolved disagreements. To be eligible, studies had to include the following: (1) an estimate of the pretest probability of pulmonary embolism, using the clinical gestalt or clinical prediction rule; (2) performance of the clinical assessment blind to the results of diagnostic testing; and (3) comparison of these assessments with validated methods of confirming or refuting the diagnosis of pulmonary embolism (Box 43-1).²⁰⁻²⁴ Additional eligibility criteria were applied to studies in which a clinical prediction rule was being derived.²⁵ These studies had to systematically collect all relevant clinical data from consecutive patients and have a sufficient number of patients with confirmed pulmonary embolism ($n > 50$) to ensure accuracy of the derived rule. For each eligible study, where possible, the pretest probability categories, corresponding disease prevalences, and likelihood ratios (LRs) (and corresponding 95% confidence intervals [CIs]) are summarized.

The clinical gestalt must have been determined according to information available from the patient's medical history and findings from physical examination and routine investigations (eg, chest radiograph, ECG, and arterial blood gas analysis) without predetermined elements or a standardized score, and most important, it must have been assessed before other diagnostic testing. A clinical prediction rule used a mathematically derived formula that combined the individual contribution of each component of the medical history, physical examination findings, and routine laboratory results before diagnostic testing.

Data Analysis

Likelihood ratios and their 95% CIs were calculated with Metstat (version 1)²⁶ and CI Analysis (version 1.1).²⁷ Summary LRs were derived with random-effects measures that provide conservative CIs around the estimates.^{28,29} Decisions to include or exclude studies were made before the analysis according to the reported methods, rather than their actual results. We determined the summary LRs to get a general sense of whether structured models performed as well as the clinical gestalt. Furthermore, we pooled data only from studies that derived a structured model and specifically did not include data from subsequent validation studies, because these studies varied substantially in their study design (retrospective assessment and concomitant use of D-dimer) from the derivative studies.

RESULTS

Our search yielded a total of 1709 articles, and after scanning the abstracts and titles, we selected 443 abstracts for detailed review. Of these, 30 articles were selected for complete review and 16 were included in the final analysis. These studies involved a total of 8306 patients.

Clinical Gestalt

In the PIOPED study, physicians used their clinical gestalt to estimate the probability of pulmonary embolism according to patient medical history and physical examination findings, together with the results of a chest radiograph, an ECG, and

Box 43-1 Criteria for Diagnosis and Exclusion of Pulmonary Embolism

POSITIVE RESULT FOR PULMONARY EMBOLISM

Positive pulmonary angiogram result.²⁰

High-probability lung scan (≥ 1 segmental perfusion defect²¹ or ≥ 2 large [$>75\%$ of a segment] segmental perfusion defects⁵ with corresponding normal ventilation).

Nondiagnostic lung scan with either a positive venogram result²² or a compression ultrasonogram diagnostic for deep vein thrombosis.

Positive lung perfusion scan²³ (single or multiple wedge-shaped defect with or without matching chest radiograph abnormalities; wedge-shaped areas of overperfusion usually exist).

NEGATIVE RESULT FOR PULMONARY EMBOLISM

Normal perfusion lung scan result²³ and a normal 3-month follow-up result.

Negative pulmonary angiogram result²⁰ and a normal 3-month follow-up result.

Nondiagnostic lung scan and negative venogram result,²² serial leg compression ultrasonography,¹⁴ or impedance plethysmography²⁴ and a normal 3-month follow-up result.

Negative spiral computed tomographic scan result and negative venogram or negative serial compression ultrasonographic result and a normal 3-month follow-up result.

Negative D-dimer test result and a normal 3-month follow-up result, provided anticoagulants were withheld.

an arterial blood gas analysis (Table 43-2).^{5,23,30-34} The results of this study showed that the prevalence of pulmonary embolism correlated reasonably well with the pretest probability estimates of pulmonary embolism.

The Prospective Investigative Study of Acute Pulmonary Embolism Diagnosis (PISA-PED) tested the accuracy of perfusion scan alone compared with pulmonary angiography.²³ In this study, experienced clinicians estimated the probability of pulmonary embolism from their clinical gestalt according to patient symptoms, signs, and risk factors, together with the results of a chest radiograph, an ECG, and an arterial blood gas analysis.

Perrier et al³⁰⁻³² reported the clinical gestalt from 3 separate studies, using a diagnostic strategy in which a ventilation/perfusion lung scan, a D-dimer assay, and compression ultrasonography followed the clinical evaluation. In the first 2 studies,^{30,31} all patients underwent a ventilation/perfusion scan and then were treated according to the pretest probability assessment, D-dimer assay result, and compression ultrasonographic finding. In the third study,³² patients were assessed initially with a highly sensitive (but nonspecific) enzyme-linked immunosorbent assay D-dimer laboratory analysis. The results of these studies are consistent with those reported in the PISA-PED²³ and PIOPED⁵ studies.

Table 43-2 Accuracy of Pretest Probability Assessment for Pulmonary Embolism With Clinical Gestalt

Source, y	No. of Patients	Prevalence of Pulmonary Embolism, %	Category	Probability Estimate, %	No. of Patients	Actual Probability, %	LR (95% CI) ^a
PIOPED, ⁵ 1990	887	28	Low	0-19	228	9	0.26 (0.17-0.4)
			Moderate	20-79	569	30	1.1 (0.96-1.2)
			High	80-100	90	68	5.3 (3.5-8.0)
Minati et al, ²³ 1996	783	44	Unlikely	10	349	8	0.13 (0.09-0.18)
			Possible	50	179	47	1.1 (0.86-1.4)
			Very likely	90	225	91	12 (8.1-18)
Perrier et al, ³⁰⁻³² 1996, 1997, 1999	985	27	Low	≤20	368	9	0.21 (0.15-0.29)
			Moderate	21-79	523	33	1.1 (1.0-1.3)
			High	≥80	94	66	4.5 (3.0-6.7)
Sanson et al, ³³ 2000	413	31	Low	0-19	58	19	0.53 (0.28-0.99)
			Moderate	20-80	278	29	0.92 (0.79-1.1)
			High	>80	77	46	1.9 (1.3-2.8)
Musset et al, ³⁴ 2002 (ESSEP)	1041	34	Low	0-19	231	12	0.26 (0.18-0.38)
			Moderate	20-79	525	26	0.67 (0.58-0.78)
			High	80-100	285	68	4.0 (3.3-5.0)

Abbreviations: CI, confidence interval; ESSEP, Evaluation du Scanner Spirale dans l'Embolie Pulmonaire; LR, likelihood ratio; PIOPED, Prospective Investigation of Pulmonary Embolism Diagnosis.

^aSummary data (LR [95% CI]) for empirical pretest probability assessments are the following: low, 0.25 (0.14-0.45); moderate, 0.92 (0.71-1.2); and high, 4.7 (2.3-9.7). These summary data exclude results from the studies by Perrier et al³⁰⁻³² because the pretest probability was used to manage subgroups of patients.

Sanson et al³³ conducted a study in 6 Dutch teaching hospitals. The clinical gestalt was quantified into the pretest probability for pulmonary embolism, and patients underwent ventilation/perfusion lung scanning followed by angiography if the lung scan finding was nondiagnostic. The estimate of the pretest probability was performed by the attending physician on a visual analog scale; however, the results of chest radiographs, ECGs, and arterial blood gas analysis were not always available when the pretest probability was documented. In this study, assessment of pretest probability was less predictive than other studies of the clinical gestalt.

The Evaluation du Scanner Spirale dans l'Embolie Pulmonaire study group³⁴ assessed the accuracy of contrast spiral computed tomography (CT) of the chest for pulmonary embolism in 1041 patients. Using simple prespecified guidelines and empirical assessment based on patient medical history, physical examination findings, and results of routine investigations, clinicians stratified patients into low-, moderate-, or high-pretest-probability groups. The presence or absence of pulmonary embolism largely was based on the combined results of spiral CT and routine bilateral compression ultrasonography of the legs. If the clinical suspicion was high and the test results were negative, or if test results were inconclusive, further assessment with lung scanning and pulmonary angiography was performed. The study demonstrated reasonable discriminative ability among the 3 pretest groups.

When interpreted together, the studies show that, when experienced clinicians use clinical gestalt, the prevalence of pulmonary embolism increases with increasing pretest probability.

The PIOPED and PISA-PED studies demonstrate the influence that clinical gestalt has on the interpretation of results of subsequent tests. In the PISA-PED study, a positive scan result for pulmonary embolism (single or multiple perfusion defects with or without matching chest radiograph abnormalities), together with a possible or likely clinical pretest probability, was associated with pulmonary embolism in 92% and 99% of patients, respectively.³⁴ On the other hand (similar to the PIOPED study results), when patients had an unlikely (low) clinical pretest probability but a positive finding on perfusion scan, pulmonary embolism was diagnosed in only 50% to 60% of individuals.

The findings in the study by Sanson et al³³ suggest that the clinical gestalt is not particularly discriminating. However, the study still showed increasing prevalence of pulmonary embolism according to pretest probability.

Clinical Prediction Rules

The PISA-PED study group analyzed clinical data from their accuracy study (Table 43-2)²³ to derive a structured clinical rule.³⁵ Clinical variables were divided into 3 categories: (1) signs and symptoms; (2) results of routine tests (chest radiograph, ECG, and arterial blood gas analysis); and (3) evidence of an obvious alternative diagnosis.

Wells et al¹⁴ initially developed a 40-variable clinical rule and subsequently refined the rule after a limited pilot study. This rule (extended Wells) was used in a large multicenter study in which 1239 patients were enrolled and assigned a clinical probability of pulmonary embolism after taking a patient medical history, performing a physical examination,

and assessing chest radiography, arterial blood gas analysis, and ECG findings. A checklist of specific symptoms and signs was compiled to help assign the pretest probability. Patients were assessed for type of symptoms (“typical,” “atypical,” or “suggestive” of severe pulmonary embolism), the presence or absence of risk factors, and the presence or absence of an alternative diagnosis as likely as or more likely than pulmonary embolism to account for the patient’s symptoms.

The corresponding prevalence and LRs for pulmonary embolism in each of the 3 pretest probability categories are listed in Table 43-3.^{14,35-38} The utility of pretest probability assessment in combination with lung scanning again was highlighted. Only 8 of 27 (30%) patients with a low pretest probability and a high-probability lung scan result had angiographically proven pulmonary embolism.¹⁴

Clinical data collected on the 1239 patients by Wells et al³⁹ also were used to derive a simplified clinical rule. With a step-wise logistic regression model, 7 key variables were identified and selected for inclusion in the final rule. Cut points were identified to classify patients as low (<2), moderate (2-6), or high (>6) probability for pulmonary embolism (Table 43-4).³⁹ With this simplified rule, only 3% (LR, 0.17; 95% CI, 0.11-0.27) of patients with a low pretest probability had pulmonary embolism vs 63% (LR, 8.6; 95% CI, 5.7-13) of those with a high pretest probability.

Wicki et al³⁶ pooled clinical data obtained from the patient medical history and physical examination, together with results of the chest radiograph, ECG, and arterial blood gas analysis collected during the 3 studies, involving 986 consecutive patients. A 7-variable rule was derived by logistic

regression and statistically cross-validated (Table 43-5). A score based on a weighted sum of variables present, was used to estimate the pretest probability of pulmonary embolism. Patients with scores of less than 5 had low pretest probability of pulmonary embolism, of 5 to 8 had moderate pretest probability, and of greater than 8 had high pretest probability. The prevalence of pulmonary embolism correlated well with pretest probability.

A large emergency department-based study involving 7 US centers systematically assessed 934 patients with suspected pulmonary embolism and derived a 6-variable model from this database (Figure 43-1).³⁷ This model used 2 screening variables to assess all patients’ age and shock index (heart rate divided by systolic blood pressure). Patients younger than 50 years and with a shock index less than 1 are deemed “non-high risk”; the remaining patients are then further assessed with 4 variables. The model classified 79% of patients as non-high risk patients in whom the prevalence of pulmonary embolism was 13%, whereas the prevalence in the high-risk group (21% of patients) was 42%. Two medical students subsequently were employed to assess 117 patients presenting to one of the participating centers, and they demonstrated a high degree of interobserver agreement (weighted κ , 0.83).³⁷

The PISA-PED investigators have reanalyzed data from their initial study and included data on a further 350 patients; the latter were assessed and treated as in the first study.³⁸ Using appropriate statistical techniques, they derived and cross-validated a 15-variable model (Table 43-6). Unlike other structured models, the authors calculated and displayed the actual pretest probability for individual patients rather than the ordinal descriptors of low, moderate, and

Table 43-3 Accuracy of Clinical Prediction Rules for Assessing Pretest Probability of Pulmonary Embolism in Derivative Studies^a

Source, y	No. of Patients	Prevalence of Pulmonary Embolism, %	Prospective Validation	Pretest Probability Category	Pretest Probability, %	LR (95% CI)
Wells et al, ¹⁴ 1998 (Extended)	1239	17.5	Yes	Low	3	0.17 (0.12-0.25)
				Moderate	28	1.8 (1.5-2.1)
				High	78	17 (11-27)
Minati et al, ³⁵ 1999 (PISA-PED)	750	41	Yes	Unlikely	6	0.05 (0.03-0.10)
				Possible	46	0.99 (0.75-1.3)
				Very likely	97	47 (23-98)
				High	63	8.6 (5.7-13)
Wicki et al, ³⁶ 2001 (Geneva rule)	986	27	Yes	Low	10	0.31 (0.24-0.40)
				Moderate	38	1.7 (1.5-1.9)
				High	81	11 (6.1-21)
Kline et al, ³⁷ 2002	934	19.4	No	Nonhigh	13.3	0.64 (0.56-0.73)
				High	42.1	3.0 (2.4-3.8)
Minati et al, ³⁸ 2003 (PISA-PED II)	1100	40	No	Low	4	0.07 (0.04-0.11)
				Moderate	26	0.72 (0.6-0.87)
				High	98	66 (31-137)

Abbreviations: CI, confidence interval; LR, likelihood ratio; PISA-PED, Prospective Investigative Study of Acute Pulmonary Embolism Diagnosis.

^aSummary of pretest probability (LR [95% CI]) of structured clinical rules is as follows: low, 0.12 (0.05-0.31); moderate, 1.1 (0.76-1.6); and high, 23 (7.6-69). This summary excludes data from Kline et al,³⁷ because that study categorized patients only into low and high categories, and from Wells et al¹⁴ because the pretest probability was used to guide management, which likely resulted in case-finding bias.

Table 43-4 The Simplified Wells Scoring System^a

Findings	Score ^b
Clinical signs/symptoms of deep vein thrombosis (minimum of leg swelling and pain with palpation of the deep veins of the leg)	3.0
No alternate diagnosis likely or more likely than pulmonary emboli	3.0
Heart rate > 100/min	1.5
Immobilization or surgery in last 4 wk	1.5
History of deep vein thrombosis or pulmonary emboli	1.5
Hemoptysis	1.0
Cancer actively treated within last 6 months	1.0

^aAdapted from Wells et al³⁹ with permission.^bCategory scores are as follows: low, <2; moderate, 2-6; and high, >6. The patient's clinical score is calculated by the summing of the scores (weight) of the predictor variables that are present.**Table 43-5** The Clinical Prediction Rule by Wicki et al³⁶ (Geneva Rule)^a

Variable	Point Score ^b
Age, y	
60-79	1
≥80	2
Previous pulmonary emboli or deep vein thrombosis	2
Recent surgery	3
Pulse rate > 100/min	1
Paco ₂ , kPa ^c	
<4.8	2
4.8-5.19	1
Pao ₂ , kPa ^c	
<6.5	4
6.5-7.99	3
8-9.49	2
9.5-10.99	1
Chest radiograph appearance	
Platelike atelectasis	1
Elevated hemidiaphragm	1

^aAdapted from Wicki et al.³⁶^bThe pretest probability categories (clinical probability score range, prevalence of disease [95% confidence interval], and percentage of patients in the pretest probability category) are as follows: low (0-4, 10% [8%-13%, 49%]; intermediate (5-8, 38% [34%-43%, 38%); and high (9-16, 81% [69%-90%, 6%), respectively.^ckPa/0.133 = mm Hg. Thus, Paco₂ < 4.8 kPa becomes Paco₂ < 36 mm Hg.

high probability. Nonetheless, the probability of pulmonary embolism in the low, moderate, moderately high, and very high pretest strata shows clear discrimination among the groups (for ease of comparison, we have combined the moderate and moderately high groups).

Validation of Derived Clinical Prediction Rules

Two hundred fifty patients with suspected pulmonary embolism were assessed prospectively by the PISA-PED group.³⁵ In this study, 90% of patients were categorized correctly as hav-

ing or not having pulmonary embolism, which compared favorably with an 88% diagnostic accuracy in the initial study.

The extended Wells model has been tested prospectively by Sanson et al³³ and by Kruip et al.⁴⁰ The pretest probability in the study by Sanson et al³³ was determined retrospectively by a second physician, who used clinical information collected by the assessing physician; both physicians remained blind to the results of diagnostic testing for pulmonary embolism. Unfortunately, about 50% (212 of the 414 patients) of study patients enrolled were assessed. The Sanson et al³³ study assignments of low, moderate, and high pretest probabilities corresponded to rates of pulmonary embolism of 28%, 39%, and 46%, respectively. These results showed less discrimination among the subgroups than other studies. Kruip et al⁴⁰ combined the pretest probability assessment of patients with the results of D-dimer analysis and withheld objective testing and anticoagulant therapy in those patients categorized with a low pretest probability and a negative D-dimer result (normal level). All other patients were tested with the combination of compression ultrasonography of the legs followed by pulmonary angiography, if the ultrasonography results were negative. The model showed considerable discriminative ability when used by Kruip et al,⁴⁰ with the prevalence of pulmonary embolism ranging from 4% in the low pretest probability group to 28% and 63% in the moderate- and high-pretest-probability groups, respectively. For the subgroup of patients with a low pretest probability and a negative D-dimer result, the 3-month rate of venous thromboembolism was 0% (95% CI, 0%-6%) (Table 43-7).

The simplified Wells model also was tested by 3 groups.^{33,41,42} As with the extended Wells model, Sanson et al³³ used a second physician to assign patients retrospectively a pretest score based on the clinical data collected by the attending physician (Table 43-7). Although the attending physician was required to specify the presence of any alternate diagnosis that was more likely than pulmonary embolism, a second physician inferred this from reviewing the medical notes when the judgment was missing. The lack of an alternate diagnosis is a critical limitation of the study, given the relative importance of this factor in the model. Sanson et al³³ reported that the simplified Wells model was less discriminating in this study than in the original Wells et al¹⁴ study. Patients with a low pretest probability had a 28% prevalence of pulmonary embolism compared with 3% in the study by Wells et al,³⁹ and only 38% of patients with a high pretest probability had pulmonary embolism compared with 63% in the study by Wells et al.³⁹

At variance with these data is the subsequent prospective validation of the simplified clinical prediction rule by Wells et al⁴¹ in 4 Canadian centers and Chagnon et al⁴² in 3 centers in France and Switzerland. The Canadian study included patients assessed by one of the 43 emergency department physicians; patients with a low pretest probability and a negative D-dimer test result had no further testing per-

formed but were followed up for 3 months. The model reliably categorized patients into low-, moderate-, and high-pretest-probability subgroups, with the prevalence of disease being 1.3%, 16%, and 41%, respectively.⁴¹

In the study by Chagnon et al,⁴² emergency department residents collected and recorded clinical data on 277 consecutive patients with suspected pulmonary embolism to create a score. Although the final score was calculated retrospectively, all the variables were documented clearly. Subsequent treatment of patients was determined by the results of D-dimer testing. Patients with a positive D-dimer result were further investigated with a combination of ultrasonographic testing of the legs, lung scanning, and pulmonary angiography.³² Consistent with the prospective validation by Wells et al,⁴¹ the emergency department residents were able to stratify patients into low-, moderate-, and high-pretest-probability categories, with ascending prevalences of pulmonary embolism.

The clinical model derived by Wicki et al³⁶ has been validated prospectively by Chagnon et al.⁴² Emergency department residents collected all the relevant data on consecutive patients with suspected pulmonary embolism and assigned each patient a pretest probability according to the Wicki model. The results of the assessment of patients using the Wicki model showed that patients identified as low, moderate, or high pretest probability for pulmonary embolism showed ascending prevalences of pulmonary embolism.

Precision of the Examination and Components of the Clinical Prediction Rules

To be useful, the pretest probability for pulmonary embolism needs to be reproducible. Put simply, when the same patient is assessed, 2 physicians' clinical gestalt should yield similar estimates of the pretest probability. None of the individual studies documented interobserver variability for the clinical gestalt.

Wells et al¹⁴ documented observer variability for the pretest probability using the extended model ($\kappa = 0.86$). Kline et al³⁷ employed 2 medical students to test the observer variability of their rule and demonstrated excellent observer agreement (weighted κ , 0.83). Chagnon et al⁴² did not document concordance between 2 observers for either of the 2 models they tested, but they documented modest agreement between the Wells simplified model and the Wicki model (weighted κ , 0.43) and found that in only 2 of 277 cases was there extreme disagreement in the pretest probability assessment.

D-dimer Assay

D-dimer, a specific fibrin degradation product, is generated by the action of plasmin on cross-linked fibrin.^{19,43-47} D-dimer assay is sensitive for the presence of venous thrombosis and can be used to help exclude deep vein thrombosis and pulmonary embolism. Although several assays are available, to be useful, a D-dimer assay must be highly sensitive for pulmonary embolism so that patients with this disease are not missed. In addition, for the assay to be useful, the specificity should be high enough so that the number of false-positive

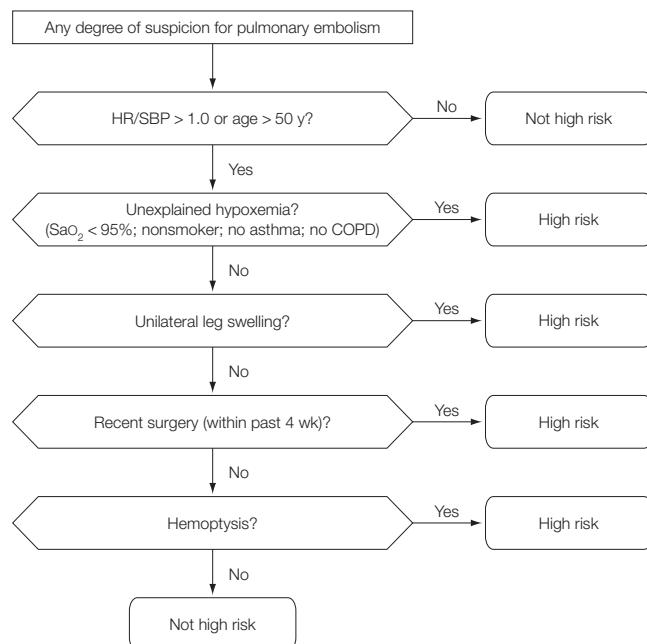


Figure 43-1 Decision Rule for Pulmonary Embolism

This model uses 2 screening variables to assess all patients' age and shock index (HR divided by SBP). Abbreviations: COPD, chronic obstructive pulmonary disease; HR, heart rate; SBP, systolic blood pressure. Adapted from Kline et al,³⁷ with permission from the American College of Emergency Physicians.

Table 43-6 Structured Clinical Model Derived by the PISA-PED Group^a

Factor	Regression Coefficient
Male sex	0.81
Age, y	
63-72	0.59
≥73	0.92
Preexisting disease	
Cardiovascular	-0.56
Respiratory	-0.97
Thrombophlebitis (ever)	0.69
Symptoms	
Dyspnea (sudden onset)	1.29
Chest pain	0.64
Hemoptysis	0.89
Temperature > 38°C	-1.17
Electrocardiogram signs of acute right ventricular overload	1.53
Chest radiograph findings	
Oligemia	3.86
Amputation of hilar artery	3.92
Consolidation (infarction)	3.55
Consolidation (no infarction)	-1.23
Pulmonary edema	-2.83

Abbreviation: PISA-PED, Prospective Investigative Study of Acute Pulmonary Embolism Diagnosis.

^aAdapted from Miniati et al,³⁸ with permission from Excerpta Medica.

Table 43-7 Accuracy of Clinical Prediction Rules for Pulmonary Embolism When Tested Prospectively

Source, y	No. of Patients	Prevalence of Pulmonary Embolism, %	Rule Prospectively Tested	Pretest Probability Category	Posttest Probability, %	LR (95% CI)
Sanson et al, ³³ 2000	237	38	Extended Wells ¹⁴	Low	28	0.66 (0.4-1.1)
				Moderate	39	1.1 (0.86-0.13)
				High	46	1.4 (0.81-2.5)
Sanson et al, ³³ 2000	414	29	Simplified Wells ³⁹	Low	28	0.93 (0.69-1.3)
				Moderate	30	1.0 (0.88-1.2)
				High	38	1.4 (0.35-5.9)
Wells et al, ⁴¹ 2001	930	9.5	Simplified Wells ³⁹	Low	1.3	0.13 (0.06-0.26)
				Moderate	16	1.9 (1.6-2.3)
				High	41	5.9 (3.7-9.3)
Kruip et al, ⁴⁰ 2002	234	22	Extended Wells ¹⁴	Low	4	0.15 (0.07-0.33)
				Moderate	28	1.5 (1.01-2.2)
				High	63	5.85 (3.51-9.74)
Chagnon et al, ⁴² 2002	277	26	Simplified Wells ³⁹	Low	12	0.39 (0.26-0.58)
				Moderate	40	2.0 (1.5-2.6)
				High	91	29 (3.8-223)
Chagnon et al, ⁴² 2002	277	26	Wicki (Geneva rule) ³⁶	Low	13	0.44 (0.30-0.65)
				Moderate	38	1.8 (1.4-2.3)
				High	67	5.8 (1.8-19)

Abbreviations: CI, confidence interval; LR, likelihood ratio.

results is sufficiently low. Newer assays can be performed rapidly, making them suitable for use in individual patients.⁴³⁻⁴⁷ The D-dimer assay is complementary to the clinical pretest probability because pulmonary embolism can be reliably excluded in patients with a negative D-dimer result and a low pretest probability.⁴¹ The accuracy indices of 3 currently available D-dimer assay types are summarized in Table 43-8.^{43,45,46}

Unfortunately, D-dimer assays vary in their sensitivities and specificities, so the posttest probability for a given patient with suspected pulmonary embolism will vary according to which D-dimer assay is used. Before clinicians use a particular D-dimer assay to revise their pretest probability, they should be aware of the differences and interpret the results of the assay accordingly.^{44,47}

Table 43-8 Estimated Accuracy Indices of 3 D-dimer Assays

D-dimer Assay	% (95% CI)		LR (95% CI)	
	Sensitivity	Specificity	Positive	Negative
Organon	96 (90-99)	45 (40-49)	1.7 (1.5-1.9)	0.09 (0.04-0.11)
Teknika latex immunoassay ⁴⁵				
Vidas Rapid ELISA assay ⁴⁶	90 (81-96)	45.1 (39-51)	1.6 (1.4-1.8)	0.22 (0.11-0.44)
SimpliRED D-dimer assay ⁴³	84.8 (79-89)	68.4 (65-71)	2.7 (2.4-3.0)	0.22 (0.16-0.3)

Abbreviations: CI, confidence interval; ELISA, enzyme-linked immunosorbent assay; LR, likelihood ratio.

CLINICAL SCENARIOS—RESOLUTIONS

CASE 1 This young woman has no risk factors or signs of pulmonary embolism (no tachycardia, features of deep vein thrombosis, or hemoptysis). No clear alternate diagnosis is present that is at least as likely as or more likely than pulmonary embolism. According to the Wells simplified clinical prediction rule, her score would be 3, a moderate pretest probability for pulmonary embolism (approximately 20%). Her whole-blood red blood cell agglutination D-dimer assay result is negative (negative LR, 0.22).⁴³ Therefore, the probability of pulmonary embolism after the results of the D-dimer assay are obtained is about 5%. The finding from a perfusion scan is normal (LR for pulmonary embolism with a normal lung scan, 0.1).⁴⁸ Therefore, her posttest probability after the above combination of tests is 0.5%, and pulmonary embolism can be ruled out.

CASE 2 This elderly patient has a high pretest probability for pulmonary embolism (approximately 65%) with the simplified Wells rule because of the combination of immobilization, tachycardia, previous deep vein thrombosis/pulmonary embolism, and the absence of an alternate diagnosis as likely as or more likely than pulmonary embolism. This combination of findings results in a score of 7, which falls into the category of a high pretest probability. In combination with a negative whole-blood red blood cell agglutination D-dimer assay result (LR, 0.22),⁴³ the revised pretest probability is approximately 30%. A ventilation/perfusion scan is reported as intermediate probability (LR, 1.2)⁴⁸; therefore, his posttest

probability of pulmonary embolism is about 33% and pulmonary embolism has not been ruled out. Further testing with compression ultrasonography and, if the finding is normal, pulmonary angiography should be considered.

THE BOTTOM LINE

Clinical assessment alone is insufficient to diagnose or rule out pulmonary embolism, although experienced clinicians can use clinical gestalt to assign a pretest probability of pulmonary embolism with reasonable accuracy. Clinical prediction rules appear to have similar accuracy to that of the clinical gestalt for patients in the low- and high-probability categories. We advocate the use of any one of the clinical prediction rules because they are simple and maintain their accuracy when used by less-experienced clinicians. In deciding which of the several rules to use, clinicians could justifiably make decisions on the scale that is easiest for them to use consistently. Factors that could affect the decision are availability of the rule in clinical reminder systems and the availability of the required clinical data. We are unable to say with confidence whether one structured clinical rule performs better than another. In outpatients with new onset or recent worsening of symptoms within the preceding 3 days, the combination of pretest probability assessment with the results of D-dimer testing improves diagnostic accuracy. Furthermore, there is emerging evidence that outpatients with a low pretest probability for pulmonary embolism can have anticoagulant therapy safely withheld when the results of D-dimer testing are negative.^{41,43}

Author Affiliations at the Time of Original Publication

Department of Medicine, McMaster University, Hamilton, Ontario, Canada (Drs Chunilal and Ginsberg); Department of Haematology, The Queen Elisabeth Hospital, Woodville, South Australia, Australia (Dr Chunilal); Department of Haematology, Royal Perth Hospital, Perth, Australia (Dr Eikelboom); Center for Clinical Epidemiology and Biostatistics, University of Newcastle, Newcastle, Australia (Dr Attia); Institute of Clinical Physiology, National Research Council of Italy, Pisa, Italy (Dr Miniati); Henderson Division, Hamilton Health Sciences Corporation, Hamilton, Ontario, Canada (Dr Panju); and Durham Veterans Affairs Medical Center and Department of Medicine, Duke University, Durham, North Carolina (Dr Simel).

Acknowledgments

Dr Chunilal is the recipient of the Noonan Fellowship, McMaster University. Dr Eikelboom is the recipient of a research fellowship from the Heart and Stroke Foundation of Canada. Dr Ginsberg is the recipient of a Career Investigator Award from the Heart and Stroke Foundation of Ontario and a research chair from CIHR-AstraZeneca.

We thank Jonathan Edlow, MD, David B. Matchar, MD, and Eugene Oddone, MD, for their important suggestions on the manuscript, as well as Arnaud Perrier, MD, Harry Buller,

PhD, and Marieke J. H. A. Kruip, MD, for providing clarification of data.

REFERENCES

- Anderson FA Jr, Wheeler HB, Goldberg RJ, et al. A population-based perspective of the hospital incidence and case fatality rate of deep vein thrombosis and pulmonary embolism: the Worcester DVT study. *Arch Intern Med.* 1991;151:933-938.
- Silverstein MD, Heit JA, Mohr DN, Petterson TM, O'Fallon WM, Melton LJ III. Trends in the incidence of deep vein thrombosis and pulmonary embolism: a 25-year population based study. *Arch Intern Med.* 1998;158(6):585-593.
- Barritt DW, Jordan SC. Anticoagulant drugs in the treatment of pulmonary embolism: a controlled trial. *Lancet.* 1960;1(7138):1309-1312.
- Goldhaber SZ, Visani L, De Rosa M. Acute pulmonary embolism: clinical outcomes in the International Cooperative Pulmonary Embolism Registry (ICOPER). *Lancet.* 1999;353(9162):1386-1389.
- PIOPED Investigators. Value of the ventilation/perfusion scan in acute pulmonary embolism: results of the Prospective Investigation of Pulmonary Embolism Diagnosis (PIOPED). *JAMA.* 1990;263(20):2753-2759.
- Stein PD, Wills PW, DeMets DL. History and physical examination in acute pulmonary embolism in patients without pre-existing cardiac or pulmonary disease. *Am J Cardiol.* 1981;47:218-223.
- Hoellerich VL, Wigton RS. Diagnosing pulmonary embolism using clinical findings. *Arch Intern Med.* 1986;146(9):1699-1704.
- Heit JA, Mohr DN, Silverstein MD, Petterson TM, O'Fallon WM, Melton LJ III. Risk factors for deep vein thrombosis and pulmonary embolism: a population-based case control study. *Arch Intern Med.* 2000;160(6):809-815.
- Grady D, Wenger NK, Herrington D, et al. Postmenopausal hormone replacement therapy increases the risk of venous thromboembolic disease: the Heart and Estrogen/Progestin Replacement Study. *Ann Intern Med.* 2000;132(9):689-696.
- Douketis JD, Ginsberg JS, Holbrook A, Crowther M, Duku EK, Burrows RF. A reevaluation of the risk for venous thromboembolism with the use of oral contraceptives and hormone replacement therapy. *Arch Intern Med.* 1997;157(14):1522-1530.
- Goldhaber SZ. Pulmonary embolism. *N Engl J Med.* 1998;339(2):93-104.
- Hampson NB. Pulmonary embolism: difficulties in the clinical diagnosis. *Semin Respir Infect.* 1995;10(3):123-130.
- Moser KM, Fedullo PF, LitteJohn JK, Crawford R. Frequent asymptomatic pulmonary embolism in patients with deep vein thrombosis. *JAMA.* 1994;271(3):223-225.
- Wells PS, Ginsberg JS, Anderson D, et al. Use of a clinical model for safe management of patients with suspected pulmonary embolism. *Ann Intern Med.* 1998;129(12):997-1005.
- Wells PS, Hirsh J, Anderson DR, et al. Accuracy of clinical assessment of deep vein thrombosis. *Lancet.* 1995;345(8961):1326-1333.
- Anand SS, Wells PS, Hunt DH, Brill-Edwards P, Cook D, Ginsberg JS. The rational clinical examination: does this patient have deep vein thrombosis? *JAMA.* 1998;279(14):1094-1099.
- Kearon C, Ginsberg JS, Crowther M, Brill-Edwards P, Weitz JI, Hirsh J. Management of suspected deep vein thrombosis in outpatients by using clinical assessment and D-dimer testing. *Ann Intern Med.* 2001;135(2):108-111.
- Bates SM, Kearon C, Crowther M, et al. A diagnostic strategy involving a quantitative latex D-dimer assay reliably excludes deep venous thrombosis. *Ann Intern Med.* 2003;138(10):787-794.
- Becker DM, Philbrick JT, Bachhuber T, Humphries JE. D-dimer testing and acute venous thromboembolism: shortcut to diagnosis? *Arch Intern Med.* 1996;156(9):939-945.
- Stein PD, Athanasioulis C, Alavi A, et al. Complications and validity of pulmonary angiography in acute pulmonary embolism. *Circulation.* 1992;85(2):462-468.
- Hull RD, Hirsh J, Carter CJ, et al. Pulmonary angiography, ventilation lung scanning, and venography for clinically suspected pulmonary embolism with abnormal perfusion lung scan. *Ann Intern Med.* 1983;98(6):891-899.
- Kruit WHJ, De Boer AC, Sing AK, Van Roon F. The significance of venography in the management of patients with clinically suspected pulmonary embolism. *J Intern Med.* 1991;230(4):333-339.

23. Miniati M, Pistolesi M, Marini C, et al. Value of perfusion lung scan in the diagnosis of pulmonary embolism: results of the Prospective Investigative Study of Acute Pulmonary Embolism Diagnosis (PISA-PED). *Am J Respir Crit Care Med.* 1996;154(5):1387-1393.
24. Hull RD, Raskob GE, Ginsberg JS, et al. A noninvasive strategy for the treatment of patients with suspected pulmonary embolism. *Arch Intern Med.* 1994;154(3):289-297.
25. Wasson JH, Sox HC, Neff RK, Goldman L. Special article: clinical prediction rules: application and methodological standards. *N Engl J Med.* 1985;313(13):793-799.
26. Suskin D, Super DH. *Metastat: Version 1*. Cleveland, OH: MetroHealth Medical Center, Case Western Reserve University, Department of Pediatrics; 1990.
27. Gardner SD, Winter PD, Gardner MJ. *Confidence Interval Analysis: Statistics With Confidence*. London, England: British Medical Journal; 1989.
28. Eddy DM, Hasselblad V, Shachter RD. *Meta-analysis by the Confidence Profile Method: The Statistical Synthesis of Evidence*. San Diego, CA: Academic Press; 1992.
29. Eddy DM, Hasselblad V. *Fast*Pro V1.8: Software for Meta-analysis by the Confidence Profile Method*. San Diego, CA: Academic Press; 1992.
30. Perrier A, Desmarais S, Goerhing C, et al. D-dimer testing for suspected pulmonary embolism in outpatients. *Am J Respir Crit Care Med.* 1997;156(2 pt 1):492-496.
31. Perrier A, Bounameaux H, Morabia A, et al. Diagnosis of pulmonary embolism by a decision analysis-based strategy including clinical probability, D-dimer levels, and ultrasonography: a management study. *Arch Intern Med.* 1996;156(5):531-536.
32. Perrier A, Desmarais S, Miron MJ, et al. Non-invasive diagnosis of venous thromboembolism in outpatients. *Lancet.* 1999;353(9148):190-195.
33. Sanson BJ, Lijmer JG, MacGillavry MR, Turkstra F, Prins MH, Buller HR. Comparison of a clinical probability estimate and two clinical models in patients with suspected pulmonary embolism. *Thromb Haemost.* 2000;83(2):199-203.
34. Musset D, Parent F, Meyer G, et al; Evaluation du Scanner Spirale dans l'Embolie Pulmonaire. Diagnostic strategy for suspected pulmonary embolism: a prospective multicenter outcome study. *Lancet.* 2002;360 (9349):1914-1920.
35. Miniati M, Prediletto R, Formichi B, et al. Accuracy of clinical assessment in the diagnosis of pulmonary embolism. *Am J Respir Crit Care Med.* 1999;159(3):864-871.
36. Wicki J, Perneger TV, Junod AF, Bounameaux H, Perrier A. Assessing clinical probability of pulmonary embolism in the emergency ward. *Arch Intern Med.* 2001;161(1):92-97.
37. Kline JA, Nelson RD, Jackson RE, Courtney DM. Criteria for the safe use of D-dimer testing in emergency department patients with suspected pulmonary embolism: a multicenter US study. *Ann Emerg Med.* 2002;39(2):144-152.
38. Miniati M, Monti S, Bottai M. A structured clinical model for predicting the probability of pulmonary embolism. *Am J Med.* 2003;114(3):173-179.
39. Wells PS, Anderson DR, Rodger M, et al. Derivation of a simple clinical model to categorize patients with a probability of pulmonary embolism: increasing the models utility with the SimpliRED D-dimer. *Thromb Haemost.* 2000;83(3):416-420.
40. Kruij JP, Slob MJ, Schijen JHEM, van der Heul C, Buller HR. Use of a clinical decision rule in combination with D-dimer concentration in diagnostic workup of patients with suspected pulmonary embolism: a prospective management study. *Arch Intern Med.* 2002;162(14):1631-1635.
41. Wells PS, Anderson D, Rodger M, et al. Excluding pulmonary embolism at the bedside without diagnostic imaging: management of patients with suspected pulmonary embolism presenting to the emergency department by using a simple clinical model and D-dimer. *Ann Intern Med.* 2001;135(2):98-107.
42. Chagnon I, Bounameaux H, Aujesky D, et al. Comparison of two clinical prediction rules and implicit assessment among patients with suspected pulmonary embolism. *Am J Med.* 2002;113(4):269-275.
43. Ginsberg JS, Wells PS, Kearon C, et al. Sensitivity and specificity of a rapid whole blood assay for D-dimer for the diagnosis of pulmonary embolism. *Ann Intern Med.* 1998;129(12):1006-1011.
44. van Beek EJR, van den Ende B, Berckmans RJ, et al. A comparative analysis of D-dimer assays in patients with clinically suspected pulmonary embolism. *Thromb Haemost.* 1993;70(3):408-413.
45. Bates SM, Grand'Maison A, Johnston M, Naguit I, Kovacs MJ, Ginsberg JS. A latex D-dimer reliably excludes venous thromboembolism. *Arch Intern Med.* 2001;161(3):447-453.
46. Sijens PE, van Ingen HE, van Beek EJ, Berghout A, Oudkerk M. Rapid ELISA assay for plasma D-dimer in the diagnosis of segmental and sub-segmental pulmonary embolism: a comparison with pulmonary angiography. *Thromb Haemost.* 2000;84(2):156-159.
47. de Moerloose P. D-dimer assay for the exclusion of venous thromboembolism: which test for which diagnostic strategy? *Thromb Haemost.* 2000;83(2):180-181.
48. Jaeschke R, Guyatt GH, Sackett DL. User's guide to the medical literature, III: how to use an article about a diagnostic test, B: what are the results and will they help me in caring for my patients? *JAMA.* 1994;271(9):703-707.

Prepared by Sanjeev Chunilal, MB ChB, FRACP, FRCPA,
and Jeff Ginsberg, MD, FRCPC

Reviewed by Phil Wells, MD

CLINICAL SCENARIO

A 25-year-old woman presents to the emergency department with pleuritic chest pain, having just returned home after a 12-hour plane flight. She is taking no medications, other than an oral contraceptive pill. Her clinical examination reveals coryza without tachypnea, and the remainder of the examination results are unremarkable. A pregnancy test result is negative and a chest radiograph result is normal. A D-dimer test shows a positive result.

Original Review

Chunilal S, Eikelboom J, Attia J, et al. Does this patient have pulmonary embolism? *JAMA*. 2003;290(21):2849-2858.

UPDATED LITERATURE SEARCH

We applied the same search criteria as was used in the original Rational Clinical Examination article to identify studies of the clinical pretest probability of pulmonary emboli. We ran a second search combining the terms “physical exam,” “medical history taking,” “sensitivity and specificity,” “observer variation,” diagnostic test, routine,” “decision support techniques,” and “pulmonary embolism.” Each search was limited to English-language articles published between 2002 and 2004. The first strategy yielded a total of 160 articles; the latter yielded 123 articles. Titles and abstracts were reviewed with the same criteria used for the original article. To find studies in which patients with suspected pulmonary embolism were enrolled in an unselected consecutive manner, participating physicians in the studies had to have been blinded to the results of diagnostic testing and had to estimate the pretest probability of pulmonary embolism. Validated algorithms to exclude or confirm the diagnosis of pulmonary embolism had to have been used.

New Findings

- New studies focus primarily on whether a low or moderate clinical probability estimate in combination with a normal D-dimer result rules out a pulmonary embolus. For such patients, the summary likelihood ratio (LR) for a pulmonary embolus is 0 with an upper 95% confidence interval

(CI) of 0.06. This combination of results effectively rules out a pulmonary embolus.

- The simplified Wells criteria have good reliability.

Details of the Update

For this update, no new clinical prediction rules were identified. Four management studies were identified with the above search strategy. One of these evaluated the performance of a logistic model that used only demographic features, symptoms, clinical signs, and radiograph results without a D-dimer assay. The other 3 studies evaluated outcomes after management that combined the results of a clinical prediction rule with the D-dimer (see Table 43-9).

IMPROVEMENTS IN THE DATA PRESENTED IN THE ORIGINAL PUBLICATION

In the original publication, a weighted κ was available for a limited number of structured clinical models. In a recent small study of patients with suspected pulmonary embolism, 2 clinicians assessed the patient initially with the extended Wells model, and then, from the data collected, each clinician was asked to determine the pretest probability by applying the simplified Wells model. The weighted κ value for the extended Wells model was 0.54 (95% CI, 0.28-0.80) vs 0.6 (95% CI, 0.34-0.85) for the simplified Wells model. In the same sub-study, there was less agreement between the extended clinical model and the pretest probability determined by clinical gestalt (weighted κ , 0.23; 95% CI, 0.05-0.42).⁵ These data suggest that the reproducibility of the clinical assessment with a structured clinical prediction rule is at best moderate but not dissimilar to the other components of the clinical examination.⁶

A reliability study of 153 patients⁷ (11% with pulmonary emboli, using helical computed tomography [CT]) assessed the simplified Wells study. The criteria had substantial agreement, with κ less than 0.70. The criterion of an “alternative diagnosis that is less likely than pulmonary embolism” had a κ of 0.58 (95% CI, 0.44-0.72), which is still considered moderate agreement. The weighted κ value for a low vs moderate vs high probability of pulmonary embolus, recalculated from the raw data displayed in the article, showed substantial agreement (0.62; 95% CI, 0.50-0.74). The results need confirmation in a larger sample of patients.

Table 43-9 Likelihood Ratios for the Pretest Probability of Pulmonary Embolus Derived From the Clinical Gestalt or Structured Clinical Models

Source, y	Number of Patients (Prevalence, %)	Model Tested	Pretest Probability Category	Pretest Probability, % (95% CI)	LR (95% CI)
Perrier et al, ¹ 2004	965 (23)	Geneva (with implicit override)	High	85 (75-92)	19 (10-36)
			Moderate	34 (29-39)	1.7 (1.5-2.0)
			Low	7 (5-9)	0.23 (0.17-0.31)
Minati et al, ² 2003	390 (41)	PISA-PED	High	100 (97-100)	297 (16-4746)
			Moderately High	86 (70-95)	8.6 (3.4-22)
			Intermediate	24 (16-34)	0.48 (0.31-0.74)
			Low	3 (1-7.0)	0.04 (0.02-0.11)
			81%-100%	67 (52-81)	8.5 (4.6-16)
Ten Wolde et al, ³ 2004	504 (20)	Empiric	51%-80%	29 (21-37)	1.6 (1.2-2.3)
			21%-50%	15 (11-19)	0.74 (0.57-0.93)
			0%-20%	8 (5-14)	0.37 (0.22-0.63)
			High	50 (32-68)	2.4 (1.3-4.5)
Leclercq et al, ⁴ 2003	202 (29)	Wells extended	Moderate	27 (17-39)	0.87 (0.56-1.4)
			Low	25 (17-35)	0.79 (0.56-1.1)

Abbreviations: CI, confidence interval; LR, likelihood ratio; PISA-PED, Prospective Investigative Study of Acute Pulmonary Embolism Diagnosis.

Table 43-10 A Low to Moderate Clinical Probability With a Normal D-dimer Result Makes Pulmonary Emboli Unlikely

Source, y	Findings	LR (95% CI)
Ten Wolde et al, ³ 2004	Clinical probability < 20% and normal D-dimer result	0 (0-0.32)
Perrier et al, ¹ 2004	Moderate or low probability and normal D-dimer result	0 (0-0.13)
Leclercq et al, ⁴ 2003	Moderate or low probability and normal D-dimer result	0 (0-0.36)
Summary		0 (0-0.06) ^a

Abbreviations: CI, confidence interval; LR, likelihood ratio.

^aSummary LR is statistically homogenous ($P = .94$).

Righini et al⁸ reanalyzed data from which the original Geneva rule was derived and also retrospectively calculated the pretest probability by applying the simplified Wells rule. The a priori hypothesis was that the discriminative ability of the clinical models would be lower in older patients com-

pared with younger patients. There was no clinically or statistically significant effect of age (younger than 50 years, 51-74 years, and older than 75 years) on the discriminative value of either model.⁸

Because of heterogeneity in the earlier studies reported in the original Rational Clinical Examination article, we did not provide summary estimates for the prediction rules or D-dimer results. The 3 new studies that added D-dimer to the established prediction rules focused primarily on the utility of the D-dimer to rule out pulmonary emboli. The studies yielded more consistent, homogenous results and provide the opportunity to create summary measures that are especially useful for understanding the role of a negative D-dimer result in ruling out pulmonary emboli (see Tables 43-10 and 43-11).

CHANGES IN THE REFERENCE STANDARD

Computed Tomography Angiography

Despite recent advances in the visualization of pulmonary arteries with the advent of spiral CT, there are no well-designed clinical outcome studies validating its role as a standalone test in the treatment of unselected patients with suspected pulmonary embolism. Newer advances in this technology, with the advent of multidetector modalities, may improve scan acquisition times and image quality. At present, although spiral CT continues to improve in its diagnostic accuracy for pulmonary embolism, a negative spiral CT result by itself is still not sufficient to reliably exclude pulmonary embolism.⁹

D-dimer With a Rapid Enzyme-Linked Immunosorbent Assay Technique

The results of a recent systematic review¹⁰ confirm that a negative D-dimer test result by itself may safely exclude pulmonary embolism. However, these data relate primarily to the enzyme-linked immunosorbent assay (ELISA) D-dimer testing format, which has superior sensitivity and negative LR compared with other D-dimer assays. Therefore, a negative D-dimer test result with the rapid ELISA format is as diagnostically useful as a negative lung scan result in patients with suspected pulmonary embolism who present with recent onset of symptoms.

RESULTS OF THE LITERATURE REVIEW

Minati et al² applied the Prospective Investigative Study of Acute Pulmonary Embolism Diagnosis (PISA-PED) structured model to assess 390 patients with suspected pulmonary embolism by categorizing them as having low, intermediate, moderately high, or high probability of pulmonary embolism. Pulmonary embolism was diagnosed or excluded by combining the pretest probability assessment with the results of perfusion lung scanning. Within these probability groups, the prevalence of pulmonary embolism was 3%, 24%, 86%, and 100%, respectively.² These data confirm the accuracy of this model when used by this group of clinicians. To date, no

other group has tested this clinical prediction rule. Because the structured model contains many more variables than other models, clinicians cannot apply the results directly without the use of a handheld calculator that contains the variables and their regression coefficients. Thus, the results are most useful for identifying the findings that are independently useful for diagnosis of pulmonary emboli.

The remaining 3 studies used the results of D-dimer testing combined with the clinical pretest probability assessment. In a study by Perrier et al,¹ consecutive patients presenting to the emergency department with suspected pulmonary embolism were evaluated with the “Geneva rule.”¹¹ Clinicians were allowed to override the pretest probability assessment if their clinical judgment disagreed with the prediction rule. The prediction rule and clinician override were done before any additional tests were obtained, including the D-dimer. All patients had a D-dimer test performed (rapid ELISA Vidas DD; BioMerieux, Marcy l’Etoile, France) and, if the assay result was negative, no further testing was performed, anticoagulant therapy was withheld, and patients were followed up for 3 months. Patients with a positive D-dimer test result underwent a preestablished standardized sequence of tests to exclude or confirm the diagnosis. The Geneva score pretest probability score was available for only 771 patients of the total cohort of 965 patients; for the remaining 126 patients, clinicians used “implicit judgment” to assess pretest probability. Of the 771 patients who were evaluated with the Geneva rule, clinicians used their judgment to change the pretest probability in 179 (23%). The pretest probability was increased in 126 patients and decreased in 53 patients. Overall, 7% (95% CI, 5%-9%) of patients with a low pretest probability had objectively confirmed pulmonary embolism compared with 35% (95% CI, 29%-39%) in the moderate- and 85% (95% CI, 75%-92%) in the high-pretest-probability groups. Strictly speaking, this study does not validate the accuracy of the Geneva rule. On the other hand, as the authors observe, allowing physicians to override the rule improves its acceptability to clinicians and makes clinical sense. The Geneva rule does not have a variable taking into account an alternative diagnosis, which might otherwise accommodate the “implicit override” feature. No patient with a moderate or low probability and a normal D-dimer result had a pulmonary embolus (LR, 0; 95% CI, 0-0.13).

Leclercq et al⁴ assessed 202 patients referred for clinically suspected pulmonary embolism. The clinical pretest probability for pulmonary embolism was formally documented with the extended Wells model¹²; subsequent investigations were based on the results of D-dimer testing (Tinaquant D-Dimer; Roche Diagnostics, Mannheim, Germany). Patients with a low or moderate pretest probability and a negative D-dimer test result were discharged without anticoagulant therapy and followed up for 3 months; none of these patients (0%; 95% CI, 0%-5.6%) had venous thromboembolism in follow-up. The remainder of patients underwent perfusion lung scanning, followed by bilateral compression ultrasonography of the legs if the lung scan result was nondiagnostic; when the ultrasonographic result was normal, pulmonary angiography was performed. The overall prevalence of

Table 43-11 Likelihood Ratios of an Abnormal D-dimer Result for a Pulmonary Embolus

Source, y	LR (95% CI)
Leclercq et al, ⁴ 2003	1.9 (1.6-2.2)
Perrier et al, ¹ 2004	1.7 (1.5-1.8)
Summary	1.7 (1.6-1.8) ^a

Abbreviations: CI, confidence interval; LR, likelihood ratio.

^aSummary LR is statistically homogenous ($P = .09$).

pulmonary embolism was 29%; 25% (95% CI, 17%-35%) in patients with a low pretest probability, 26% (95% CI, 17%-39%) in the moderate-pretest-probability group, and 50% (95% CI, 32%-68%) in the high-pretest-probability group. These results show less discrimination than the original study by Wells et al¹² but are consistent with another Dutch study.¹³

Finally, in a multicenter study,³ clinical gestalt or “informed intuition” was used to define the pretest probability of pulmonary embolism. This study group had previously used the Wells simple clinical prediction rule¹⁴ for assessing pretest probability and found it to be no more discriminatory than the pretest probability determined by an overall assessment of the clinical signs and symptoms, along with the results of basic investigation.¹³ A total of 631 patients were assessed by study physicians in 3 trial centers and were assigned to one of the 4 pretest probabilities (0%-20%, 21%-50%, 51%-80%, >81%); patients also had blood drawn for a D-dimer assessment after the clinical probabilities were assigned (Tinaquant D-Dimer; Roche Diagnostics). Patients with a low-pretest probability for pulmonary embolism and a negative D-dimer test result were discharged without anticoagulant therapy and were followed up for 3 months. Clinicians were able to reliably discriminate between low-, intermediate-, moderate-, and high-pretest-probability groups, with the prevalence of pulmonary embolism in each of the 4 groups being 8% (95% CI, 5%-14%), 15% (95% CI, 11%-19%), 29% (95% CI, 21%-37%), and 67% (95% CI, 52%-81%). These data compare favorably with studies in which the pretest probability was assessed with a structured clinical model. However, these were experienced clinicians who had extensive and specific training in assessing patients with pulmonary embolism. No patient with a low-probability clinical assessment (0%-20%) and a normal D-dimer result had a pulmonary embolus (LR, 0; 95% CI, 0-0.32).

One of the major criticisms of 2 of the structured models (extended¹² and simplified¹⁴ Wells) is the need to specify or weight the likelihood of an alternative diagnosis apart from pulmonary embolism, introducing a global assessment of the probability of pulmonary embolism, not unlike the gestalt pretest assessment. This variable has been the most problematic in terms of its reliability.⁵ Therefore, a third model, the Geneva rule, now encompasses this variable, in part by allowing physicians to upgrade or downgrade the pretest probability with an implicit override. Pragmatically, this makes sense and may improve the acceptance of this model among clinicians.

There has been considerable controversy with respect to the association between risk of venous thromboembolism and airline travel. Well-designed case-control studies suggest an odds ratio of 2 for the association.⁶ Studies in selected patients⁶⁻⁹ suggest the absolute risk for venous thromboembolism increases with duration of travel, the presence of thrombophilia, and use of estrogen-containing therapy. There are conflicting data on the true incidence of venous thrombosis within the traveling public, with estimates ranging from as low as 1.6 events per million passengers to as high as 10% for asymptomatic, ultrasonographically detected, calf vein thrombosis.⁵⁻⁹

EVIDENCE FROM GUIDELINES

Recent guidelines from the British Thoracic Society support assessing and formally documenting the pretest probability for pulmonary embolism, but they do not specifically advocate a structured model approach over the clinical gestalt.¹⁵ The guideline reiterates the importance of establishing the pretest probability before reviewing the results of a ventilation perfusion lung scan or the results of a D-dimer test. This process can identify low-risk patients who do not need further testing.

CLINICAL SCENARIO—RESOLUTION

A 25-year-old woman presents to the emergency department with pleuritic chest pain, having just returned home after a 12-hour plane flight. Your examination confirms the coryza and absence of tachypnea (16/min). The remainder of the examination results, including that for the legs, are unremarkable. A pregnancy test result is normal, as is a chest radiograph. Results of an arterial blood gas analysis do not show hypoxia, and an electrocardiogram result is normal.

This woman poses a challenge to the assessment of the pretest probability for pulmonary embolism, largely because of the uncertainty of the magnitude of the risk for venous thrombosis after airplane travel. According to the simplified Wells model, her pretest probability for pulmonary embolism is low (absence of tachycardia, active cancer, and signs of deep vein thrombosis; no history of venous thrombosis; and no hemoptysis). The overall clinical assessment suggests that the young woman is more likely to have viral pleurisy. A low pretest probability (3%-5%), combined with a positive D-dimer result (MDA D-Dimer; BioMerieux, Inc, Durham, North Carolina) (positive LR, 1.7¹⁶), places her posttest probability for pulmonary embolism at 8%. A ventilation perfusion lung scan result is normal (LR for pulmonary embolism with a normal lung scan result, 0.1).¹⁷ Her posttest probability of pulmonary embolism is less than 2%. If you had chosen a pretest “low” probability of as much as 15%, the posttest probability would still be low, at 2.9%, after the positive D-dimer result and normal ventilation-perfusion scan result.

See next page for the “Make the Diagnosis” section.

REFERENCES FOR THE UPDATE

- Perrier A, Roy PM, Aujesky D, et al. Diagnosing pulmonary embolism in outpatients with clinical assessment, D-dimer measurement, venous ultrasound and helical computed tomography: a multicenter management study. *Am J Med.* 2004;116(5):291-299.^a
- Minati M, Monti S, Bauleo C, et al. A diagnostic strategy for pulmonary embolism based on standardised pretest probability and perfusion lung scanning: a management study. *Eur J Nucl Med Mol Imaging.* 2003;30(11):1450-1456.^a
- Ten Wolde M, Hagen PJ, Macgillavry MR, et al. Non-invasive diagnostic work-up of patients with clinically suspected pulmonary embolism: results of a management study. *J Thromb Haemost.* 2004;2(7):1110-1117.^a
- Leclerc MGL, Lutisan JG, van Marwijk M, et al. Ruling out clinically suspected pulmonary embolism by assessment of clinical probability and D-dimer levels: a management study. *J Thromb Haemost.* 2003;89(1):97-103.^a
- Hughes RJ, Hopkins RJ, Hill S, et al. Frequency of venous thromboembolism in low to moderate risk long distance air travellers: the New Zealand Air Travellers Thrombosis Study (NZATT). *Lancet.* 2003;362(9401):239-244.
- Cook DJ. Clinical assessment of central venous pressure in the critically ill. *Am J Med Sci.* 1990;299(3):175-178.
- Wolf SJ, McCubbin TR, Feldhaus KM, Faragher JP, Adcock DM. Prospective validation of Wells criteria in the evaluation of patients with suspected pulmonary embolism. *Ann Emerg Med.* 2004;44(5):503-510.
- Righini M, le Gal G, Perrier A, Bounameaux H. Effect of age on the assessment of clinical probability of pulmonary embolism by prediction rules. *J Thromb Haemost.* 2004;2(7):1206-1208.
- Stijen MJL, De Monye W, Schiereck J, et al. Single-detector helical computed tomography as the primary diagnostic test in suspected pulmonary embolism: a multi-centre clinical management study of 510 patients. *Ann Intern Med.* 2003;138(4):307-314.
- Stein PD, Hull RD, Patel K, et al. D-dimer for the exclusion of acute venous thrombosis and pulmonary embolism: a systematic review. *Ann Intern Med.* 2004;140(8):589-602.
- Wicki J, Perneger TV, Junod AF, Bounameaux H, Perrier A. Assessing the clinical probability of pulmonary embolism in the emergency ward. *Arch Intern Med.* 2001;161(1):92-97.
- Wells PS, Ginsberg JS, Anderson DR, et al. Use of a clinical model for safe management of patients with suspected pulmonary embolism. *Ann Intern Med.* 1998;129(12):997-1005.
- Sanson BJ, Lijmer LG, Macgillavry M, Turkstra F, Prins MH, Buller HR; ANTELOPE Study Group. Comparison of a clinical pretest probability estimate and two clinical models in patients with suspected pulmonary embolism. *J Thromb Haemost.* 2000;83(2):199-203.
- Wells PS, Anderson DR, Rodger M, et al. Derivation of a simple clinical model to categorise patients with a probability of pulmonary embolism: increasing the model's utility with SimpliRED D-dimer. *J Thromb Haemost.* 2000;83(3):416-420.
- British Thoracic Society Standards of Care Committee Pulmonary Embolism Guideline Development Group. British Thoracic Society guidelines for the management of suspected acute pulmonary embolism. *Thorax.* 2003;58(6):470-483.
- Bates SM, Kearon CJ, Crowther M, et al. A diagnostic strategy involving a quantitative latex D-dimer assay reliably excludes deep vein thrombosis. *Ann Intern Med.* 2003;138(10):787-794.
- Jaeschke R, Guyatt GH, Sackett DL; Evidence Based Medicine Working Group. User's guide to the medical literature, III: how to use an article about a diagnostic test, B: what are the results and will they help me in caring for my patients? *JAMA.* 1994;271(9):703-707.
- Cushman M, Tsai AW, White RH, et al. Deep vein thrombosis and pulmonary embolism in two cohorts: the longitudinal investigation of thromboembolism etiology. *Am J Med.* 2004;117(1):19-25.
- Chunilal SD, Eikelboom JW, Attia J, et al. The rational clinical examination: does this patient have pulmonary embolism? *JAMA.* 2003;290(21):2849-2858.
- Bates SM, Ginsberg JS. Treatment of deep vein thrombosis. *N Engl J Med.* 2004;351(3):268-277.
- Tapson V, Carroll B, Davidson BL, et al. The diagnostic approach to acute venous thromboembolism: clinical practice guideline. *Am J Respir Crit Care Med.* 1999;160(3):1043-1066.

^aFor the Evidence to Support the Update for this topic, see <http://www.JAMAevidence.com>.

PULMONARY EMBOLUS—MAKE THE DIAGNOSIS

PRIOR PROBABILITY

Venous thrombosis occurs in 1 to 2 persons per 1000 person-years, with approximately one-half to one-third of these episodes from pulmonary embolism.¹⁸ In published studies, the prevalence of pulmonary embolism in patients who present with a clinical suspicion ranges from 9% to more than 30%,¹⁹ which undoubtedly relates to a combination of factors, including differences in referral patterns and health practices among countries, as well as differences in patient populations. The prior probability of a pulmonary embolus is determined from the clinical findings. Although studies vary in the prevalence of disease, a useful guideline would be to think of “low probability” as approximately less than 15% and “moderate probability” as 15% to 35%.

POPULATION FOR WHOM PULMONARY EMBOLUS SHOULD BE CONSIDERED

Patients who have had recent major surgery, major trauma, immobility, or active malignancy are some of the highest-risk groups within the general population, with relative risks varying from 5 to 200.²⁰ The most common presenting symptoms of pulmonary embolism are new or worsening dyspnea, acute chest pain, and, less frequently, cough, fainting, or hemoptysis. Tachypnea and tachycardia, the most common signs of pulmonary embolism, occur frequently with exacerbations of chronic obstructive lung disease, congestive cardiac failure, and pneumonia, which highlights the poor specificity of these signs.²¹

DETECTING THE LIKELIHOOD OF PULMONARY EMBOLUS

Use a structured model to assess the pretest probability of pulmonary emboli. The simplified Wells scoring system may be the easiest to use in clinical practice, shows good reliability, and requires no laboratory tests or radiographs (see Table 43-12).

Establishing the pretest probability before, and not after, reviewing the results of a sensitive D-dimer test will identify patients at very low risk for pulmonary emboli (see Table 43-13).

When there is discordance between clinician gestalt and a clinical prediction rule, most experts would place the patient

Table 43-12 Simplified Wells Scoring System

Findings in the Simplified Wells Scoring System	Score ^a
Clinical signs/symptoms of DVT of the leg (minimum of leg swelling and pain with palpation of the deep veins)	3.0
No alternate diagnosis that is as likely as or more likely than a pulmonary embolus	3.0
Heart rate > 100/min	1.5
Immobilization or surgery in the last 4 weeks	1.5
History of DVT or PE	1.5
Hemoptysis	1.0
Cancer actively treated in the past 6 mo	1.0

Abbreviations: DVT, deep vein thrombosis; PE, pulmonary embolism.

^aCategory scores determined by the sum of the individual scores: low, <2; moderate, 2-6; high, >6. Adapted from Chunilal et al.¹⁹

Table 43-13 The Likelihood Ratios for Pulmonary Embolus for the Combination of Clinical Probability Estimate With the D-dimer Result

Clinical Probability	D-dimer	LR (95% CI)
Any probability (2 studies)	Abnormal	1.7 (1.6-1.8)
Low (<15%) to moderate (15%-35%) (3 studies)	Normal	0 (0-0.06)

Abbreviations: CI, confidence interval; LR, likelihood ratio.

into the higher pretest probability group. Combining the pretest probability with the results of D-dimer testing reduces the need for further investigations in those patients with a low-moderate probability of pulmonary embolism and a negative D-dimer result because a number of management studies now confirm the safety of this approach.

REFERENCE STANDARD

The reference standard test for proving pulmonary emboli requires visualization of the embolus by arteriography or appropriate perfusion defects with nuclear studies. However, current approaches to the diagnosis of pulmonary emboli now recognize that patients with a low to moderate pretest probability and a negative high-sensitivity D-dimer result can be treated without anticoagulation, effectively ruling out the presence of pulmonary embolism.

This page intentionally left blank

EVIDENCE TO SUPPORT THE UPDATE:

43

Pulmonary Embolus

TITLE Ruling Out Clinically Suspected Pulmonary Embolism by Assessment of Clinical Probability and D-dimer Levels: A Management Study.

AUTHORS Leclercq MGL, Lutisan JG, van Marwijk Kooy, et al.

CITATION *Thromb Haemost*. 2003;89(1):97-103.

QUESTION What are the efficiency and safety of excluding pulmonary embolism based on a normal D-dimer combined with a low or moderate clinical probability as assessed by a clinical model?

DESIGN Prospective cohort study.

SETTING Single hospital, The Netherlands.

PATIENTS Two hundred two patients (inpatients and outpatients) with suspected pulmonary embolism were enrolled from August 1999 to April 2001.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

Consecutive patients with suspected pulmonary embolism were assessed and assigned a structured pretest probability.¹ Patients with low or moderate pretest probabilities were treated subsequently according to D-dimer results. If D-dimer result was negative, no further testing was performed, anticoagulant therapy was withheld, and these patients were followed up at 3 months. All other patients were investigated with an algorithm.

Pulmonary emboli were confirmed by a high-probability perfusion lung scan and normal chest radiograph result, positive compression ultrasonography result, or pulmonary angiogram. Pulmonary emboli were excluded by a low or moderate clinical pretest probability with a normal D-dimer test result and a negative 3-month follow-up result, a normal lung scan result, or negative pulmonary angiogram result.

Data collected included patient demographics, pretest probability assessments, and D-dimer test results, as well the results of objective tests.

MAIN OUTCOME MEASURES

The proportion of patients in whom pulmonary embolism was safely excluded according to a low or moderate pretest probability and a negative D-dimer test result was compared to the rate of pulmonary emboli among all patients.

MAIN RESULTS

Twenty-nine percent of patients (59/202) had confirmed pulmonary embolism. The likelihood ratios for the clinical probability estimate alone (Table 43-14), the D-dimer result alone (Table 43-15), and the combination of the probability estimate with the D-dimer result (Table 43-16) can be calculated. Thirty-two percent of patients had pulmonary embolism safely excluded according to a low or moderate pretest probability and a negative D-dimer test result. In this group of patients, the subsequent 3-month risk of recurrent venous thromboembolism was 0% (95% confidence interval [CI], 0%-6%). The overall prevalence of pulmonary embolism in the low-, moderate-, and high-pretest-probability strata was 25% (95% CI,

Table 43-14 Likelihood Ratio of the Clinical Probability Estimate for Pulmonary Emboli

Probability, All Patients	LR (95% CI)
High	2.4 (1.3-4.5)
Moderate	0.87 (0.56-1.4)
Low	0.79 (0.56-1.1)

Abbreviations: CI, confidence interval; LR, likelihood ratio.

Table 43-15 Likelihood Ratio of the D-dimer Result for Pulmonary Emboli

All Patients	LR (95% CI)
D-dimer positive	1.9 (1.6-2.2)
D-dimer negative	0 (0-0.29)

Abbreviations: CI, confidence interval; LR, likelihood ratio.

Table 43-16 Likelihood Ratio of the D-dimer Result Among Patients With a Moderate or Low Clinical Probability Estimate for Pulmonary Emboli

Moderate or Low Clinical Probability Patients	
D-dimer Result	LR (95% CI)
Positive	2.0 (1.7-2.4)
Negative	0 (0-0.36)

Abbreviations: CI, confidence interval; LR, likelihood ratio.

17%-35%), 26% (95% CI, 17%-39%), and 50% (95% CI, 32%-68%), respectively, showing minimal discriminative value of the pretest probability model used.

CONCLUSIONS

LEVEL OF EVIDENCE Level 2.

STRENGTHS A well-designed cohort study with objective confirmation or exclusion of pulmonary embolism, as well as appropriate follow-up. The clinical probability was determined before the D-dimer results were known.

LIMITATIONS The focus of this study was on the ability of the D-dimer to rule out pulmonary emboli. In this sense, “rule out” should be interpreted in literal terms because the goal was to determine whether treatment could be withheld for patients with a moderate or low probability and a negative D-dimer result. The sample size was too small for a definitive conclusion.

This study suggests that patients with a low or moderate pretest probability and a negative D-dimer result can be treated without anticoagulant therapy or additional objective testing. However, clinicians must decide whether to accept the conclusion in light of the CI around the LR because the upper limit of the CI may not be sufficiently low (upper limit, 0.36). Given the 29% prevalence of pulmonary emboli, a patient with a moderate to low clinical probability and a normal D-dimer result could have a probability as high as 13%.

Because there were so few patients with a high clinical probability and normal D-dimer result, clinicians should continue to evaluate these patients further for pulmonary emboli and not assume that pulmonary emboli have been ruled out.

Although it appears that the clinical probability alone works as well as the D-dimer alone, such a conclusion may lack validity if the patients were identified for enrollment according to the features in the Wells et al¹ prediction model.

Reviewed by Sanjeev Chunilal, MB ChB, FRACP, FRCPA

REFERENCE FOR THE EVIDENCE

- Wells PS, Ginsberg JS, Anderson DR, et al. Use of a clinical model for the safe management of patients with suspected pulmonary embolism. *Ann Intern Med*. 1998;129(12):997-1005.

TITLE A Diagnostic Strategy for Pulmonary Embolism Based on Standardized Pretest Probability and Perfusion Lung Scanning: A Management Study.

AUTHORS Miniati M, Simonetta M, Bauleo C, et al.

CITATION *Eur J Nucl Med Mol Imaging*. 2003;30(11): 1450-1456.

QUESTION What are the efficacy and safety of making the diagnosis of pulmonary emboli according to a combination of standardized estimates of the pretest probability and perfusion lung scan results?

DESIGN Prospective cohort study.

SETTING One Italian hospital.

PATIENTS Four hundred twenty-five patients enrolled from April 2000 to September 2001.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

Patients with suspected pulmonary embolism were enrolled prospectively in this single-center study. All were examined by one of the 12 respiratory physicians who determined the clinical pretest probability by applying a standardized clinical scoring system without knowledge of whether the patient had a pulmonary embolus.¹

Pulmonary embolism was confirmed or excluded on the basis of perfusion lung scanning or pulmonary angiography.² The perfusion scan was interpreted independently of the clinical data. Pulmonary embolism was excluded if the perfusion lung was normal or near normal or pulmonary angiography result was normal; for those with a nondiagnostic perfusion scan result, pulmonary embolism was considered excluded in patients with a low clinical pretest probability. For all other patients, pulmonary embolism was diagnosed according to a high-probability perfusion scan (single or multiple wedge-shaped perfusion defects) or pulmonary angiography.

MAIN OUTCOME MEASURES

The proportion of patients with a definitive diagnosis obtained noninvasively was described. A regression model was derived to estimate the influence of clinical findings on the likelihood of a pulmonary embolus.

MAIN RESULTS

Pulmonary embolism was diagnosed noninvasively in 132 (34%) patients and by angiography in an additional 28 (7%) and was excluded in 220 (noninvasively in 191 patients [49%] and by a negative pulmonary angiogram result in 39

Table 43-17 Factors in the Regression Model That Alter the Likelihood of Pulmonary Emboli

Increased the Likelihood of PE	Decreased the Likelihood of PE
Male	Preexisting cardiovascular disease
Age > 63 y	Preexisting pulmonary disease
Thrombophlebitis (ever)	Temperature > 38°C
Dyspnea (sudden onset)	
Chest pain	
Hemoptysis	
ECG signs of acute right ventricular overload	
Radiographic findings	Radiographic findings
Oligemia	Consolidation (no infarction)
Amputation of the hilar artery	Pulmonary edema
Consolidation (infarction)	
Probability From Model of PE (Calculated From the Regression Model) ^a	LR (95% CI)
High (>90%)	297 (16-4746)
Moderately high (50% to ≤90%)	8.6 (3.4-22)
Intermediate (10% to ≤50%)	0.48 (0.31-0.74)
Low (≤10%)	0.04 (0.02-0.11)

Abbreviations: CI, confidence interval; ECG, electrocardiogram; LR, likelihood ratio; PE, pulmonary embolism.

^aData kindly provided by Massimo Miniati, MD, PhD.

[10%]). Only 1 patient had a pulmonary embolus not detected during the initial evaluation.

The regression model had a large number of demographics, risk factors, symptoms, signs, and electrocardiographic and radiographic findings (see Table 43-17).

CONCLUSIONS

LEVEL OF EVIDENCE

Level 3.

STRENGTHS Strict adherence to the diagnostic criteria to objectively exclude or prove the diagnosis of pulmonary embolism, as well as a 1-year follow-up for those patients in whom the initial evaluation result was negative. The clinical model and the perfusion scans were interpreted independently.

LIMITATIONS This is a single-center study in which the clinical prediction rule was applied by one of 12 highly specialized observers. Therefore, the generalizability of this model to other centers and observers remains to be proven. Methodologically, there was incorporation bias in which the results of the model were used as part of the reference standard. However, these criteria were specified in advance and patients who did not meet the criteria required angiography.

Clinicians collect clinical data that, when incorporated into a prediction model, is useful in stratifying patients' likelihood of a pulmonary embolus. The prediction model requires entry of the data into either a spreadsheet or handheld calculator to derive

the probability. These results are most helpful for identifying the clinical findings that are independently useful.

Reviewed by Sanjeev Chunilal, MB ChB, FRACP, FRCPA

REFERENCES FOR THE EVIDENCE

1. Miniati M, Monti S, Bottai M. A structured clinical model for predicting the probability of pulmonary embolism. *Am J Med.* 2003;114(3):173-179.
2. Miniati M, Pistolesi M, Marini C, et al. Value of perfusion lung scan in the diagnosis of pulmonary embolism: results of the Prospective Investigative Study of Acute Pulmonary Embolism Diagnosis (PISA-PED). *Am J Respir Crit Care Med.* 1996;154(5):1387-1393.

TITLE Diagnosing Pulmonary Embolism in Outpatients With Clinical Assessment, D-dimer Measurement, Venous Ultrasound, and Helical Computer Tomography: A Multicentre Management Study.

AUTHORS Perrier A, Roy PM, Aujesky D, et al.

CITATION *Am J Med.* 2004;116(5):291-299.

QUESTION What is the efficiency of a diagnostic strategy for venous thromboembolism that combines clinical assessment, plasma D-dimer, lower limb venous ultrasonography, and helical computed tomography (CT)?

DESIGN A prospective cohort study.

SETTING Two Swiss hospitals and 1 French hospital.

PATIENTS Patients presenting to the emergency department with suspected pulmonary embolism (PE) were prospectively enrolled, using predefined criteria: acute onset of new or worsening shortness of breath or chest pain without another obvious etiology. Nine hundred sixty-five patients were enrolled from October 1, 2000, to June 30, 2002.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

Physicians assessed the clinical pretest probability for PE by collating the results of a standardized clinical scoring system.¹ The scoring system categorized patients as "low," "intermediate," or "high" probability. When physicians disagreed with the results of the prediction rule, they could "override" this by recording their implicit clinical judgment. The probability estimate was recorded with knowledge of the arterial blood gas and chest radiograph results but no other laboratory or radiographic study results.

The diagnostic standard for excluding PE consisted of a negative D-dimer result (enzyme-linked immunosorbent assay), a negative helical CT scan result, and compression ultrasonography of the legs when combined with a low or moderate clinical probability or a negative pulmonary angiography result and no subsequent recurrent venous thrombosis during a 3-month follow-up. Pulmonary embolism was diagnosed according to a

positive helical CT scan result, positive result for compression ultrasonography of the legs (proximal deep vein thrombosis), or a positive pulmonary angiogram result.

MAIN OUTCOME MEASURES

The proportion of patients in whom a definitive diagnosis of PE could be made without the need for pulmonary angiography and the risk of venous thromboembolism in patients who had anticoagulants withheld because the strategy excluded PEs. We calculated likelihood ratios from data provided in the tables (see Tables 43-18, 43-19, and 43-20).

MAIN RESULTS

Pulmonary embolism was diagnosed in 222 of 965 (23%) patients, with only 2.7% of patients requiring pulmonary angiography for a definitive diagnosis. A total of 194 (20%) patients did not have the standardized scoring system applied to assess pretest probability because of incomplete data. For these patients, physicians assigned an implicit pretest probability assessment. On the other hand, there was disagreement between the standardized pretest probability assessment and physicians' implicit judgment in 179 patients (23%), with 70% of these instances requiring upgrading of the clinical score. The likelihood ratios for the clinical score alone (Table 43-18), the D-dimer result alone (Table 43-19), and the combination of the clinical score with the D-dimer result (Table 43-20) can be calculated.

Table 43-18 Likelihood Ratios for the Probability of Pulmonary Emboli According to a Clinical Score

Probability, All Patients	LR (95% CI)
High	19 (10-36)
Moderate	1.7 (1.5-2.0)
Low	0.23 (0.17-0.31)

Abbreviations: CI, confidence interval; LR, likelihood ratio.

Table 43-19 Likelihood Ratio of the D-dimer Result for Pulmonary Emboli

All Patients	LR (95% CI)
D-dimer result positive	1.7 (1.5-1.8)
D-dimer result negative	0 (0-0.08)

Abbreviations: CI, confidence interval; LR, likelihood ratio.

Table 43-20 Likelihood Ratio of the D-dimer Result Among Patients With a Moderate or Low Clinical Probability Estimate for Pulmonary Emboli

Moderate or Low Clinical Probability	LR (95% CI)
D-dimer result positive	1.6 (1.5-1.7)
D-dimer result negative	0 (0-0.13)

Abbreviation: CI, confidence interval; LR, likelihood ratio.

CONCLUSIONS

LEVEL OF EVIDENCE Level 3.

STRENGTHS A prospective, multicenter, multinational study with a large number of patients.

LIMITATIONS Twenty percent of potentially eligible patients were excluded according to a number of predefined criteria, though the most frequent reason was a protocol violation. The large number of exclusions makes the patient population nonconsecutive, which is a potentially important limitation if the exclusions were among patients who had a normal D-dimer result. Nearly 40% of study patients did not have a standardized clinical assessment because of the absence of arterial blood gas results or because the standardized score was revised by implicit clinical judgment. According to the presenting feature of acute chest pain or dyspnea, patients with deep vein thrombi confirmed by ultrasonography were assumed to have PE without further studies.

The clinical scoring system and diagnostic algorithm used in this study are primarily applicable to outpatients with recent onset of worsening or new symptoms. The standardized scoring system, occasionally overridden by clinical judgment, was good at identifying patients most likely to have a PE. However, the focus of this study was on identifying patients without PE so that additional studies and treatment could be avoided. A normal D-dimer result appears better than the scoring system and clinical judgment. However, because the scoring system and clinical judgment were applied first to identify the eligible patients, the D-dimer should be applied in light of the clinical findings. An intermediate or low probability of PE, combined with a normal D-dimer result, was efficient at identifying patients with a low likelihood of an embolus. Given the prior probability of 22% in this study, taking the upper end of the 95% confidence interval (CI) for intermediate–low probability patients and a normal D-dimer result (upper 95% CI likelihood ratio, 0.13) yields a maximum probability of 3.5%.

Reviewed by Sanjeev Chunilal, MB ChB, FRACP, FRCPA

REFERENCE FOR THE EVIDENCE

- Wicki J, Perneger TV, Junod A, Bounameaux H, Perrier A. Assessing the clinical probability of pulmonary embolism in the emergency ward with a simple score. *Arch Intern Med.* 2001;161(1):92-97.

TITLE Non-invasive Diagnostic Work-up of Patients With Clinically Suspected Pulmonary Embolism: Results of a Management Study.

AUTHORS Ten Wolde M, Hagen PJ, Macgillavry MR, et al; Advances in New Technologies Evaluating the Localization of Pulmonary Embolism Study Group.

CITATION *J Throm Haemost*. 2004;2(7):1110-1117.

QUESTION Does a diagnostic algorithm safely reduce the need for ventilation perfusion lung scintigraphy and pulmonary angiography in patients who have a low clinical pretest probability for pulmonary embolism and a negative D-dimer test result?

DESIGN Prospective cohort study.

SETTING Three teaching hospitals in The Netherlands.

PATIENTS Six hundred thirty-one consecutive inpatients and outpatients enrolled from May 1999 to April 2001.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

Patients were all assessed and assigned a clinical pretest probability ($\leq 20\%$, 21%-50%, 51%-80%, and $>80\%$), which was determined by the responsible physician taking into account the patient's medical history, findings on physical examination, and the results of routine investigations. All the clinicians were specifically trained to assess patients with suspected pulmonary embolism. All patients had a plasma D-dimer test (rapid immunoturbidimetric assay; Tinaquant Roche Diagnostics, Mannheim, Germany) performed after the clinical probability was established.

Pulmonary embolism was considered excluded in patients with a low clinical pretest probability combined with a negative D-dimer result, a normal ventilation perfusion lung scan result, or nondiagnostic lung scan with negative serial compression testing result of the lower limbs when these patients remained venous thrombosis free at a 3-month follow-up (not receiving anticoagulants).

Patients were diagnosed as having pulmonary embolism according to a high-probability lung scan, positive pulmonary angiography result, or a positive result for compression ultrasonographic examination of the legs.

Data collected included patient demographics, patient pretest probability assessment, and D-dimer results, as well as the results of objective tests (ventilation/perfusion lung scan, compression ultrasonographic examination of the legs, and pulmonary angiography).

MAIN OUTCOME MEASURES

The primary safety outcome was the incidence of confirmed venous thrombosis in patients who had venous thrombosis initially excluded.

Table 43-21 Likelihood Ratio of the D-dimer Result Combined With the Clinical Probability Estimate for Pulmonary Emboli

	LR (95% CI)
Probability $> 20\%$ or D-dimer result abnormal	1.3 (1.2-1.3)
Probability $< 20\%$ and D-dimer result normal	0 (0-0.32)

Abbreviations: CI, confidence interval; LR, likelihood ratio.

MAIN RESULTS

Of 466 patients in whom pulmonary embolism was considered excluded at presentation, 1.3% (95% confidence interval, 0.5%-2.8%) had a subsequent venous thromboembolus. Among the low-pretest-probability group, 95 patients also had a negative D-dimer result, and none of these patients had confirmed recurrence during the subsequent 3 months. A low clinical probability and a normal D-dimer result appeared to rule out pulmonary emboli (Table 43-21).

Within the entire cohort, 20% of patients had confirmed pulmonary embolism, with the prevalence of disease increasing along with increasing clinical pretest probability. The corresponding rates of pulmonary embolism for the low-, intermediate-, moderate-, and high-pretest groups were statistically different, at 8%, 15%, 29%, and 67%, respectively, confirming that these experienced clinicians using clinical gestalt were able to accurately categorize patients.

CONCLUSIONS

LEVEL OF EVIDENCE Level 2.

STRENGTHS Prospective data collection with probability estimate established before the D-dimer.

LIMITATIONS The focus of this study was on the ability of the D-dimer to rule out pulmonary emboli. In this sense, "rule out" should be interpreted literally because the goal was to determine whether treatment could be withheld for patients with a low probability ($<20\%$) and a negative D-dimer result. The sample size was too small for a definitive conclusion. A D-dimer was obtained in 82% of patients, but the majority of those in whom it was not obtained (109/112) had a clinical probability greater than 20%.

Clinicians who are specifically trained to identify pulmonary embolism can accurately identify low- and high-risk patients for pulmonary embolism. The data confirm the importance of the pretest probability in triaging patients and using additional tests on the higher-risk patients. As in other studies, the upper confidence interval for the low-probability patients with a normal D-dimer result may not convince some physicians that a pulmonary embolus has been ruled out.

Reviewed by Sanjeev Chunilal, MB ChB, FRACP, FRCPA

This page intentionally left blank

CHAPTER 44

Does This Patient Have an Instability of the Shoulder or a Labrum Lesion?

Jolanda J. Luime, MSc

Arianne P. Verhagen, PhD

Harald S. Miedema, MD

Judith I. Kuiper, PhD

Alex Burdorf, PhD

Jan A. N. Verhaar, PhD, MD

Bart W. Koes, PhD

CLINICAL SCENARIO

A 24-year-old man with a history of shoulder complaints presents to his primary care physician. At age 16 years, his shoulder was injured during karate. He recovered and did not notice recurrence of symptoms. At age 21 years, while throwing a baseball, he developed sudden sharp left shoulder pain, with a popping noise. He sensed that the arm stretched out of range. He experienced a short period with shoulder discomfort, followed by recovery.

Recently, he has started playing tennis and experiences shoulder pain that requires cessation of play. On examination, the shoulder displays no swelling or atrophy. Internal and external rotation is somewhat painful but not limited. His neck moves normally, through the full range of motion, without pain. In considering the differential diagnosis, one might wonder whether the medical history suggests instability of the shoulder and which physical examination findings confirm the diagnosis.

WHY IS THE DIAGNOSIS IMPORTANT?

The shoulder's wide range of motion gives great freedom of action because of the shallow structure of the glenoid fossa but lends minimal bony support for the large humeral head (*Figure 44-1*). The minimal bony support creates a delicate balance between muscular and ligamentous strength.¹ Each year, 30% to 40% of adults experience shoulder discomfort, causing 1% to 5% of them to visit a general practitioner.²⁻⁸ Although about half of the primary care patients with shoulder discomfort recover within a year, a substantial number experience continued discomfort or develop recurrent pain.^{6,7,9} Instability of the glenohumeral joint, frequently combined with tears of the labrum (the cartilage rim of the glenoid), creates continued problems for some of these patients.

Instability occurs when the shoulder's stabilizing structures provide too little control as the humerus moves on the glenoid. As a result, the upper arm fails to stay properly located in the glenoid fossa during normal motion. Dislocation occurs when the humeral head has no attachment to the glenoid fossa; thus, the articular surfaces separate completely. Subluxation is a symptomatic translation of the humeral head without complete separation.^{1,10-12} The resultant symptoms and signs allow clinical classification according to the degree (dislocation or subluxation) and the direction (anterior, posterior, inferior, or multidirectional) of the observed defects.^{1,10-12} The incidence of shoulder dislocation is about 1.7% of the general population.¹³ Scientific literature shows no available data on the incidence or prevalence of subluxation.

Treatment of instability depends on the type and severity of the luxation detected during the clinical examination and on the patient's functional deficits. The primary option, in most

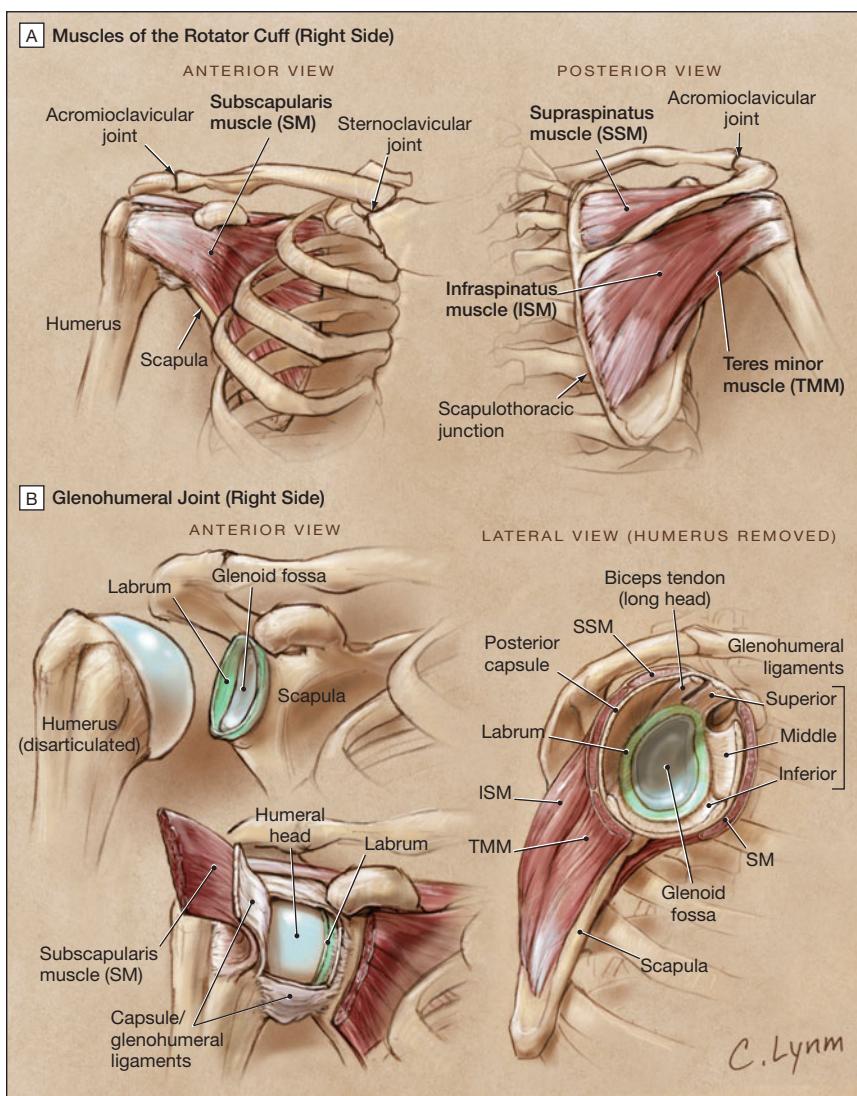


Figure 44-1 Anatomy of the Shoulder

cases, is conservative treatment^{1,10,11} of strengthening the muscles of the shoulder and increasing the coordination of the shoulder girdle. The alternative is surgery, a useful treatment if the patient has recurrent dislocation without generalized ligamentous laxity or multidirectional instability.^{1,10,11}

Labral lesions are associated with instability, although they can occur without instability because of injuries or degeneration of the shoulder joint.^{14–16} Labral lesions are classified according to their anatomic location and type of tear.¹⁴ A frequently described labral tear is the superior labrum anterior-posterior (SLAP) lesion.^{14,15} The SLAP lesion is a tear located at the superior part of the labrum that runs from the anterior to the posterior part, with or without lesions at the attachment of the long head of the biceps muscle. Surgical repairs of labral tears require an open or arthroscopic procedure.^{14,15}

Anatomy of the Shoulder

The shoulder is suited for mobility. The motions of the upper arm are the result of simultaneous motions in the glenohumeral joint, the acromioclavicular joint, the sternoclavicular joint, and the scapulothoracic junction.¹⁷ Shoulder instability and labral lesions affect the functioning of the glenohumeral joint.

The glenohumeral joint is the articulation between the large humeral head and the small glenoid fossa of the scapula (Figure 44-1). The fossa is extended by the glenoid labrum (a cartilage rim) that increases the depth and surface area of the articulation.^{1,14} The labrum cushions the apposition of the humeral head on the glenoid fossa, similar to the function of the menisci in the knee. A loose capsule surrounds the joint, strengthened by 3 thickenings called the anterior glenohumeral ligaments.¹

Seventeen muscles create the movement of the shoulder.¹⁷ The movement is a complex and subtle interaction between the 4 articulations and contributing muscles. Although knowledge of the biomechanics of the shoulder is growing, knowledge about the relationship with clinical diagnosis is still limited. An important finding related to instability is the functioning of the 4 muscles of the rotator cuff (infraspinatus, supraspinatus, teres minor, and subscapularis). These muscles play the most important roles in stabilizing the

glenohumeral joint, even when the arm is in a neutral or relaxed position.¹⁷

Mechanism of Injuries Resulting in Instability or Labral Tears

Instability has 3 causes. A generally known cause of anterior luxation includes a sudden traumatic fall with an outstretched arm (seen frequently in skiers) or blocked throwing movement of the arm. Usually, this luxation will be reduced in the field or the hospital emergency department. More typically, primary care physicians observe a second type of shoulder instability, created without obvious trauma and attributed to chronic gradual stretching during overhead activities in work or sport.¹⁰ Finally, hyperlaxity of the glenohumeral capsule, a less common cause of instability and often without any trauma,^{1,10-12} is caused by congenital excessive joint laxity that allows the shoulder to slip in different directions (multidirectional instability). Some patients with hyperlaxity of the glenohumeral capsule can dislocate their shoulder voluntarily.

The mechanisms that create labral tears without dislocation are unclear.¹⁶ The shoulder capsule and ligaments are attached to the labrum; thus, strong forces on these structures are also potentially harmful to the labrum. The occurrence of labral tears has been predominantly studied in patients with throwing injuries.¹⁸ In this group, tears are associated with strong forces of strain on the anterior capsule, ligaments, and labrum generated during the throwing motion. Labral tears are distinct from rotator cuff tears. A labral tear involves a tear of cartilage, whereas a rotator cuff tear occurs in one of the tendons of the rotator cuff muscles. Instability of the joint or labral tears can occur with rotator cuff injuries. However, rotator cuff injuries do not normally create dislocations or labral tears. Their symptoms might be different, although it is not clear from the current evidence.

CLINICAL PRESENTATION

The diagnosis of an acute shoulder dislocation is easy to establish. It is a painful condition and the patient will hold the arm in a fixed position (Figure 44-2).^{1,10-12} However, patients with shoulder instability without dislocation present in a more subtle way. Some patients may complain about a “dead arm” feeling.^{1,10} Symptoms of pain and functional disability seem to be nonspecific for the presence of instability.^{1,19} Instability of the shoulder should be considered when patients have shoulder discomfort without clear restriction of motion. A history of dislocation increases the likelihood of recurrent instability. Instability occurs more commonly in young people, although traumatic dislocation also occurs in older patients.^{1,13}

The clinical examination of the shoulder for instability is performed to evoke recurrence of the symptoms (provocation tests) or to determine laxity of the glenohumeral joint (Table 44-1).^{1,10} In a provocation test, the humeral head is placed in a position of imminent subluxation or dislocation, which makes the patient recognize the pain-provoking movement and react



Figure 44-2 Radiograph of Shoulder Luxation

with anticipated fear or pain (an apprehension test) (Figure 44-3A, C). Laxity tests of the shoulder evaluate the amount of translation of the humeral head on the glenoid in different positions of the humerus in the anterior, posterior, and inferior directions. As opposed to apprehension tests, these laxity tests are not intended to evoke discomfort.

To assess the amount of translation, rehabilitation specialists and orthopedic surgeons use a classification system such as the Hawkins grading scheme. Grade 0 denotes little to no movement, grade 1 denotes the humeral head moves onto the glenoid rim, grade 2 indicates the humeral head can be dislocated but spontaneously relocates, and grade 3 indicates the humeral head does not relocate when the pressure is removed.^{1,20} In the Hawkins scheme, grades 1 to 3 are considered positive outcomes on a laxity test.

When laxity is present in more than one direction, the diagnosis of multidirectional instability is considered and the patient should be examined for generalized ligamentous laxity (laxity in more joints of the body).^{1,10-12} There are no uniformly accepted clinical criteria for generalized ligamentous laxity. One might suspect this type of laxity when finding positive laxity tests in both shoulders. Other examples of hyperlaxity include the ability to hyperextend the elbows and a positive thumb-to-forearm test, whereby the patient can pull his or her thumb back to the point of touching the forearm. Typically, such patients will know that they can demonstrate their loose joints.

Patients with labral tears present with a variety of symptoms.¹⁶ Snyder¹⁴ suggested that the most common clinical symptoms are pain with overhead activities, deep shoulder pain, or painful catching, popping, or clicking. Stetson and Templin²¹ suggested that these symptoms were not specific for labral tears because they mimic the presence of impingement disorders, rotator cuff tears, or other shoulder problems. Although an obvious clinical presentation for labral tears cannot be described, clinicians should consider the diagnosis when the shoulder pain is related to a traumatic

Table 44-1 Clinical Tests for Instability and Laxity

Diagnostic Test	Provocation	Patient Positioning	Arm Positioning	Technique	Outcome
Provocation/Relief Tests for Instability					
Relocation ^a	Pain and apprehension	Supine	Abducted to 90 degrees and externally rotated to 90 degrees	Humeral head pressed posteriorly while arm is externally rotated	Relieves pain and apprehension
Anterior release ^a	Pain and apprehension	Supine	Abducted to 90 degrees and externally rotated to 90 degrees	Same as relocation test; then posterior pressure is suddenly released	Pain or apprehension
Apprehension ^a	Pain and apprehension	Sitting or standing	90-Degree abduction and full external rotation	Arm is externally rotated while pressure is applied anteriorly to humeral head	Pain or apprehension
Clunk	Clunk or grinding	Supine	Full abduction	Arm is rotated in full external rotation, caput humeri is pushed slightly in anterior direction	Clunk or grinding
Laxity Tests for Instability					
Load and shift anterior or posterior laxity	Anterior or posterior	Sitting, standing, or supine	Neutral position	Humeral head is fixed by clinician's hand; clinician tries to shift humeral head in anterior (or posterior) direction	Does not evoke discomfort; degree of humeral head translation on the glenoid in different positions of the humerus is evaluated using the Hawkins grading scheme ^b
Sulcus sign	Inferior laxity	Sitting or standing	Neutral position	Arm is pulled vertically downward	Positive when sulcus becomes visible between acromion and humeral head
Provocation/Relief Tests for Labral Tears					
Biceps load I	Pain	Supine	Arm is abducted 90 degrees, elbow is flexed 90 degrees	Clinician applies flexion pressure as patient resists	Positive if pain occurs
Biceps load II ^c	Pain	Supine	120-Degree abduction	Clinician applies lateral force as patient resists	Positive if pain occurs
Mimori ^c	Pain and apprehension	Sitting or standing	Arm is abducted 90 degrees, elbow is flexed 90 degrees, forearm is supine	Forearm is brought from maximum supination to maximum pronation	Positive if pain occurs
Zaslav ^c	Compares strength in internal rotation to that of external rotation, excluding impingement from labral tears	Sitting or standing	Arm is in 90-degree abduction and 80-degree external rotation, elbow is flexed 90 degrees	Patient resists external rotation force applied by the clinician, followed by applied internal rotation force	Positive (labral tear present) when the patient has good strength against external rotation and apparent weakness against internal rotation
Active compression (O'Brien)	Pain and relief	Sitting or standing	Arm is in 90-degree forward flexion, 10- to 15-degree abduction, and full internal rotation	Clinician stands in front of patient and arm is pushed down as patient resists; repeated with arm in external rotation	Positive if pain elicited with first maneuver is reduced or eliminated in the second
Compression rotation	Pain or clicking	Supine	Arm at 90-degree abduction, elbow in 90-degree flexion	Axial load placed on shoulder while rotated and circumducted (note McMurray knee test)	Positive if pain or clicking occurs
SLAP-prehension	Pain or clicking	Sitting or standing	Arm at 90-degree forward flexion	Arm is rotated internally in 90-degree flexion of the humerus	Positive if pain or clicking occurs
Speed	Pain in the anterior shoulder	Sitting or standing	90-Degree elevation	Downward force applied to forearm, full supination of forearm, and elbow is fully extended	Positive if pain occurs
Tenderness of bicipital groove	Pain	Sitting	Neutral	Palpating the bicipital groove	Positive if pain occurs
Yergason	Pain in the biceps tendon	Sitting with elbow at 90 degrees	Neutral	Patient supinates forearm against clinician's resistance, who simultaneously palpates biceps tendon	Positive if pain occurs

Abbreviation: SLAP, superior labrum anterior posterior.

^aTests shown in Figure 44-3.

^bHawkins grading scheme: grade 0 denotes little to no movement; grade 1 denotes the humeral head moves onto the glenoid rim; grade 2 indicates the humeral head can be dislocated but spontaneously relocates; grade 3 indicates the humeral head does not relocate when the pressure is removed. In the Hawkins scheme, grades 1 to 3 are seen as positive outcomes on a laxity test.

^cTests shown in Figure 44-4.

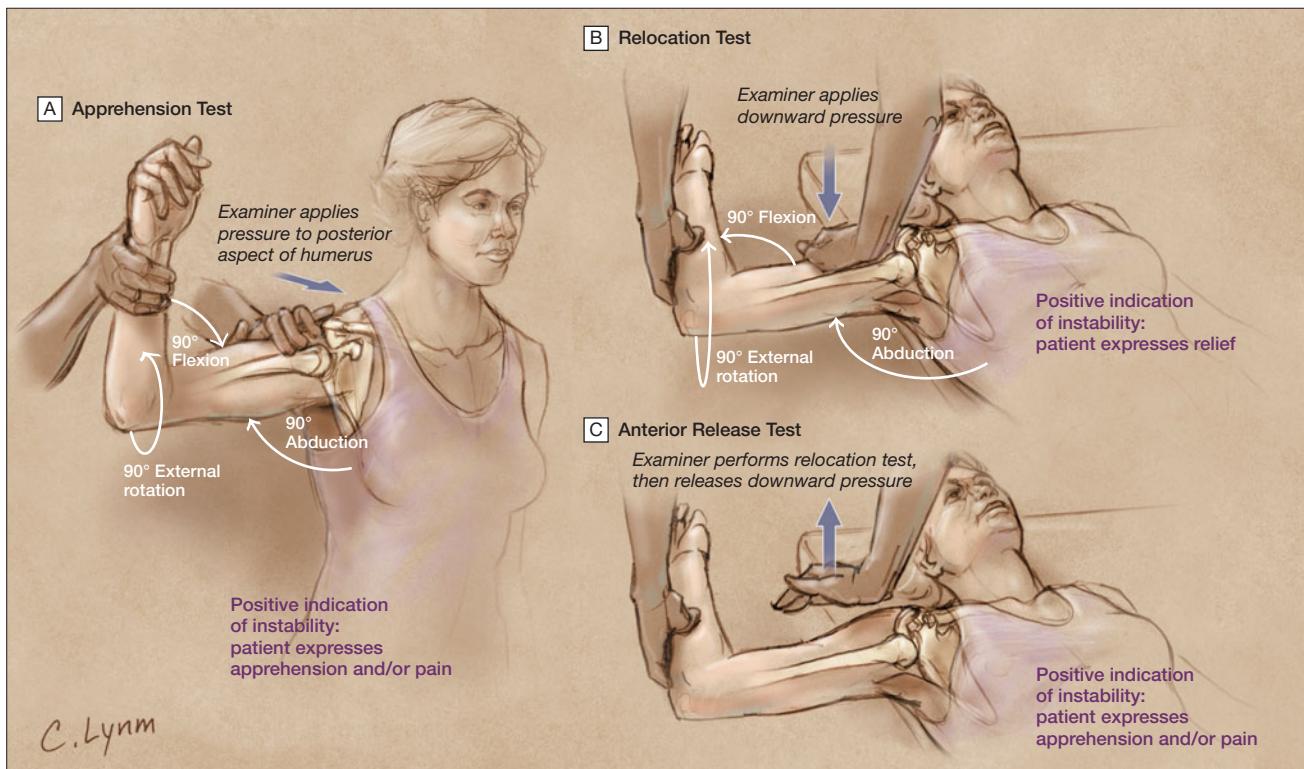


Figure 44-3 Clinical Tests to Evaluate Anterior Instability of the Shoulder

A, Apprehension test, although of limited clinical value because of low specificity, is included as part of a sequence of tests for shoulder instability. It is conducted with the patient sitting or standing, with the arm placed in 90-degree abduction and 90-degree external rotation, and the elbow flexed 90 degrees. Pressure is applied to the posterior aspect of the humerus. B, Relocation test, performed to relieve symptoms (pain and apprehension) of instability, is conducted with the patient supine and the arm abducted to 90 degrees and externally rotated to 90 degrees. Downward (posterior) pressure is applied to the humeral head. C, The anterior release test is conducted in a similar manner as the relocation test, then the examiner's hand is removed suddenly, releasing pressure on the humeral head.

injury that involves substantial forces on the glenohumeral joint (eg, falling while skiing).

Clinical tests for detecting labral tears evoke symptoms by compressing the humerus into the glenoid in an attempt to catch the labral fragment between the bony structures (compression rotation test).²² Other eponymous tests to evoke symptoms by rotating the humerus passively or actively, such as the pain provocation test of Mimori et al,¹⁸ are shown in Figure 44-4. Alternative physical examination maneuvers reproduce shoulder symptoms by asking the patient to resist the force of the clinician while the arm is held in a fixed position, such as the biceps load II test²³ shown in Figure 44-4.

Signs and symptoms for shoulder instability must be recorded accurately to add appropriate diagnostic information. We reviewed the literature on the accuracy of diagnostic studies for shoulder instability.

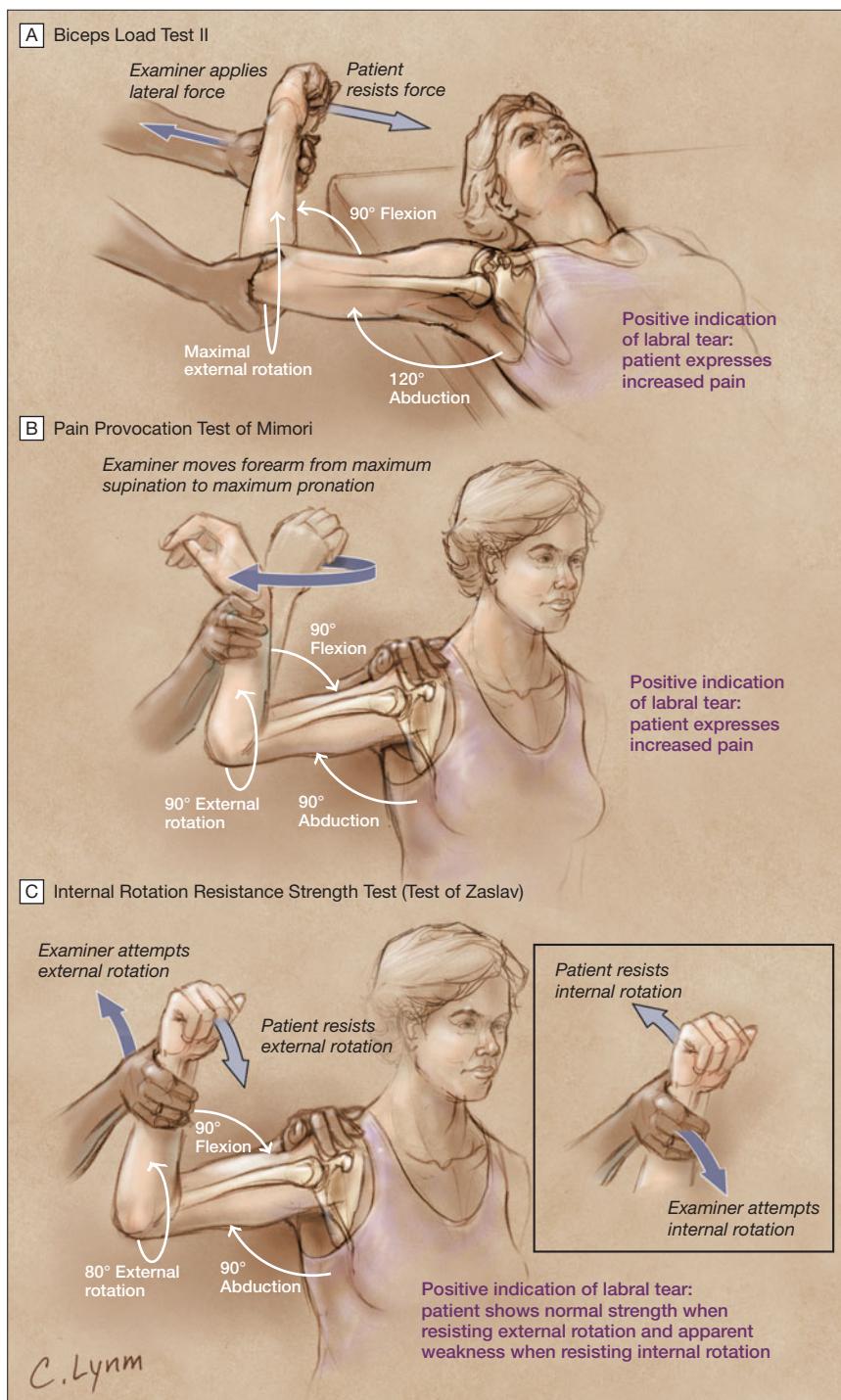
METHODS

This review is based on the guidelines for systematic reviews of studies evaluating the accuracy of diagnostic tests²⁴ identified through the MEDLINE (1966-2003), EMBASE (1980-2001), and CINAHL (1982-2001) databases. To retrieve all relevant publications related to diagnosing shoulder complaints in adults, the term “exp

shoulder” was searched. In addition, text word searches were completed for “glenohumeral,” “scapula,” “clavicula,” “acromion,” “rotator cuff,” “supraspinatus,” “supra-spinatus,” “infraspinatus,” “infra-spinatus,” “serratus anterior,” and “subscapularis.” Diagnostic studies were retrieved by exploding the phrase “sensitivity and specificity,” with additional text word searches of “specificity,” “false negative,” “screening,” and “accuracy” based on the search strategy of Deville et al.²⁵ Bibliographies of known primary and review articles were examined. One reviewer (J.J.L.) screened abstracts of the retrieved citations on clinical tests, sensitivity and specificity figures, and shoulder pain. Relevant articles were researched and their reference lists were screened to find additional studies.

Studies were screened by 2 reviewers (J.J.L., B.W.K.) and had to meet the following inclusion criteria: (1) description of clinical tests for instability or intra-articular pathology (IAP) of the shoulder, (2) use of a reference (gold) standard, (3) detailing of sensitivity and specificity, and (4) publication in English, Dutch, or German. Studies were excluded if the diagnoses included fibromyalgia or systemic disorders such as rheumatoid arthritis, fractures, tumors, or strokes.

We selected studies that compared a clinical test to surgical or arthroscopic findings, rather than noninvasive imaging tests (eg, magnetic resonance imaging, ultrasonography,

**Figure 44-4 Clinical Tests for Labral Tears**

A, Biceps load test II is performed with the patient supine, the arm is placed in 120-degree abduction (90-degree abduction in biceps load test I), and the elbow is placed in 90-degree flexion. The patient is asked to resist the lateral force applied by the examiner. B, In the pain provocation test of Mimori, the arm is placed in 90-degree abduction, the elbow in 90-degree flexion, and the forearm in maximum supination. To provoke symptoms, the examiner moves the forearm into maximum pronation. C, Internal rotation resistance strength test (test of Zaslav) is conducted with the patient standing or sitting, with the humerus in 90-degree abduction and 80-degree external rotation. The patient is asked to resist an external rotation force applied by the examiner and then to resist an applied internal rotation force.

or computed tomography). Although these imaging tests may be useful in confirming the presence of instability or a labral tear, they have a sensitivity of only 60% to 90%, depending on the type of injury²⁶ and in comparison with surgery or arthroscopy. Approximately 10% to 20% of patients with a normal reading on shoulder magnetic resonance imaging or ultrasonography²⁶ may still have shoulder instability or labral tears. Thus, these noninvasive tests might ultimately prove useful as a pragmatic reference standard for some physicians, although the presence of verifica-

tion bias (no surgery or arthroscopy implemented when the noninvasive study result is normal) and possible low sensitivity create uncertainty when the utility of the clinical examination is reviewed.

For each study, details were extracted on study population (setting, sampling, age, sex, and diagnosis), clinical tests, reference tests, and outcome (sensitivity and specificity). When raw data were available, the likelihood ratios (LRs) were calculated for individual findings, thereby describing the increase in odds that the patient had shoulder instability

when a symptom or sign was present or the opposite effect when a sign or symptom was absent.

The methodologic quality of the studies was evaluated by 2 reviewers (A.P.V., J.J.L.) with the Quality Assessment of Diagnostic Accuracy Studies checklist.²⁷ This list includes 14 questions about the spectrum of patients studied, selection criteria, test verification, test description, blinding, uninterpretable results, and study withdrawals. These questions could be scored as positive if the item was fulfilled, negative if the item was not fulfilled, or unclear if the item was not described. The limitations of each study were described. The studies were not allocated into arbitrary categories of low, medium, or high quality.

RESULTS

Our search strategy used a broad spectrum of terms for the shoulder, yielding about 21000 articles. Combined with the search strategy of Deville et al²⁵ on diagnosis, this resulted in 1449 abstracts from the 3 databases. About 130 abstracts contained information on shoulder disorders and diagnostic outcome measurements. However, most of the articles evaluated sonography vs surgery, magnetic resonance imaging vs surgery, or one type of magnetic resonance imaging vs another type.

Formal reviews were conducted for 35 articles that evaluated clinical tests. Seventeen of these studies^{16,18,19,21-23,28-38} met the selection criteria for inclusion in this review (Table 44-2). Eighteen studies were excluded: 11 because no information on instability or IAP was presented,³⁹⁻⁴⁹ 4 because data were missing on sensitivity and specificity or clinical tests,⁵⁰⁻⁵³ and 3 because they were published in French.⁵⁴⁻⁵⁶ Of the 17 studies that were selected, 5 enrolled patients when the clinician suspected shoulder instability^{19,33,35,37,38} and 12 enrolled patients when the clinician suspected labral tears or other IAP.^{16,18,21-23,28-32,34,36} All the studies were conducted in orthopedic clinics. Each study evaluated a varying number of clinical tests but lacked data on patient medical history. Surgery was used as a reference test in 6 studies,^{19,29,30,33,35,37} and arthroscopy in 11.^{16,18,21-23,28,31,32,34,36,38} The apprehension test,^{19,38,39} relocation test,^{19,38} active compression test,^{21,29} anterior slide test,^{22,34} and the test of Speed^{30,38} were evaluated in more than 1 study. Two studies reported the clinical examination of the shoulder under anesthesia with the same protocol.^{33,37} These studies were not pooled because of lack of clinical homogeneity in study populations. Although most studies had the same inclusion criterion for participant selection (having surgery or arthroscopy for shoulder complaints), the selection standards for undergoing surgery or arthroscopy were unclear. Hence, the constitution of the population might have differed. In addition, different end points of the diagnoses made it impossible to evaluate the influence of the diagnostic threshold for sensitivity and specificity.

Accuracy of Signs and Symptoms Related to Instability and Labral Tears

No diagnostic studies assessed the value of history taking in diagnosing instability. Four provocation tests for instability are presented in Table 44-3. The relocation test³⁸ and the anterior release test³⁵ have the best properties for increasing the likelihood of instability (relocation test³⁸: positive LR, 6.5; 95% confidence interval [CI], 3.0-14; and negative LR, 0.18; 95% CI, 0.07-0.45; anterior release test³⁵: positive LR, 8.3; 95% CI, 3.6-19; and negative LR, 0.09; 95% CI, 0.03-0.27). The relocation test does not work as well in determining more subtle degrees of anterior instability as opposed to more obvious cases of instability, although we were unable to evaluate the CI around the LRs for detecting less significant instability.¹⁹ The apprehension test and the clunk test were both of limited value because of low specificity and low sensitivity, respectively.

Establishment of instability was not confirmed or ruled out with the sulcus sign³⁸ or the load and shift anterior posterior laxity tests.³⁸ The likelihood of instability increased when laxity tests were performed under anesthesia (positive LR, 13; 95% CI, 3.9-43)³³; however, these tests cannot be performed in the general medical practice because of the use of anesthesia.

The possibility of detecting labral tears by arthroscopy has renewed interest in clinical tests for detecting affected patients. Thirteen studies^{16,18,21-23,28-32,34-36} have evaluated 14 clinical signs, and 8 of these^{18,21-23,28,29,32,34} allowed calculation of positive and negative LRs (Table 44-4). The anterior slide test,^{22,34} the crank test,^{16,21,28} and the active compression test^{16,21,22,29} were promising when their designers evaluated them. However, the accuracy and LRs found by other researchers were far less hopeful. Therefore, optimism should be reserved for test results that have not been duplicated in subsequent studies. The biceps load I test³² (positive LR, 29; 95% CI, 7.3-115), the biceps load II test²³ (positive LR, 26; 95% CI, 8.6-80), the pain provocation test of Mimori et al¹⁸ (positive LR, 7.2; 95% CI, 1.6-32), and the internal rotation resistance strength test³¹ (positive LR, 25; 95% CI, 8.1-76) need confirmation before they become widely adopted. Conflicting evidence was found for the test of Speed.^{16,30}

Limitation of the Literature

The results of the presented studies pose some limitations and should be interpreted with caution (Table 44-2). The diagnostic studies were all executed in specialized care; therefore, the optimal spectrum of disease was defined as patients visiting an orthopedics clinic with shoulder pain. However, in 15 studies^{16,19,21-23,28,30-38} patients were selected from waiting lists for shoulder surgery or shoulder arthroscopy. In these studies, spectrum bias cannot be excluded. Besides, this selection criterion resulted in a highly selected group of patients with severe shoulder disorders, which is also noticeable in the high prevalence values (15%-100%) of instability and labral lesions. A high prevalence among study subjects reduces the opportunity to detect both false-positive and true-negative results, which will overestimate the sensitivity and underestimate the specificity when the test is applied to patient populations with a lower prevalence of disease. It is likely that clinical findings in daily medical practice have lower sensitivity but higher specificity than suggested in the available literature.

Other limitations of the existing literature include modest sample sizes and methodologic problems. Twelve of the 17 studies did not describe the procedure for selecting patients.^{18,19,21-23,29,30,32,33,35,36,38} The time between index and reference test was unknown in most studies.^{18,23,28-38} The details of the reference

Table 44-2 Study Characteristics

Source, y	Selection Criteria	Total No. of Participants (% of Women)	Mean Age, y	Index Test	Limitations ^a
Reference Test Arthroscopy; Retrospective Design					
Berg and Ciullo, ³⁶ 1998	Identified SLAP lesions during arthroscopy	66 (NA)	NA	SLAP-prehension test	b, d, f, g
Reference Test Arthroscopy; Prospective Design					
Guanche and Jones, ¹⁶ 2003	First arthroscopy for shoulder pain, complete range of motion under anesthesia	61 (19)	38	Active compression test; anterior apprehension test; crank test; relocation test; test of Speed; test of Yergason; tenderness in bicipital groove	e
Kibler, ³⁴ 1995	Isolated glenoid labral tear, partial-thickness rotator cuff pathology, Bankart lesion, capsular deficiency, or 25-degree internal rotation deficit	226 (33)	NA	Anterior slide test	a, b, c, d, f, g, h
Kim et al, ³² 1999	Arthroscopy for unilateral recurrent anterior shoulder dislocation (based on physical examination, plain radiograph, and MRI) with a Bankart lesion	75 (15)	25	Biceps load test I	a, b, e, f
Exclusion: multidirectional instability					
Kim et al, ²³ 2001	Arthroscopy for shoulder problems	127 (30)	31	Biceps load test II	a, b, e
Exclusion: dislocation; stiff shoulder					
Liu et al, ²⁸ 1996	Shoulder surgery after failure of conservative treatment	62 (22)	28	Crank test	b, d, e
Exclusion: traumatic dislocation; weakness of subscapularis					
McFarland et al, ²² 2002	Diagnostic arthroscopy for shoulder pain	426 (NA) ^b	NA	Compression rotation test; anterior slide test; active compression test	a
Mimori et al, ¹⁸ 1999	Shoulder pain during throwing motions	32 (6)	21	Crank test; anterior apprehension test in external and internal rotation	a, b, c, f
Exclusion: instability; indications of rotator cuff tears on MRI or arthrography					
Stetson and Templin, ²¹ 2002	Diagnostic arthroscopy after failure of conservative treatment	65 (31)	46	Crank test; active compression test	a, b, f, h
T'Jonck et al, ³⁸ 2001	Shoulder arthroscopy due to disabling shoulder pain	71 (45)	NA	Active compression test; apprehension test; clunk test; lift-off test; load-and-shift test; posterior stress test; release test; relocation test; resistance test external rotation; test of Speed; sulcus sign	a
Exclusion: >65 y; previous surgery of shoulder; interaction with complaints in elbow or neck					
Zaslav, ³¹ 2001	Shoulder surgery after failure of conservative treatment; positive Neer overhead sign	110 (41) ^c	44	Internal rotation resistance strength test	b
Reference Test Surgery; Prospective Design					
Bennett, ³⁰ 1998	Surgery for shoulder pain	45 (31)	NA	Test of speed	a, b
Cofield et al, ³³ 1993	Surgery after referral for suspected recurrent instability	55 (27)	29	Laxity tests under anesthesia in anterior, posterior, inferior, anterior-inferior and posterior-inferior direction	a, b, e
Gross and Distefano, ³⁵ 1997	Subluxation or gross dislocation on examination under anesthesia; abnormal excursion during arthroscopic examination; Hill-Sachs lesion or Bankart lesion	82 (38) ^d	37	Anterior release test	a, b, e, f
O'Brien et al, ²⁹ 1998	Shoulder pain	268 (NA)	NA	Active compression test	a, b, c, d, e, f, g, h
Oliashirazi et al, ³⁷ 1999	Shoulder surgery for unilateral traumatic recurrent anterior instability	30 (17)	23	Laxity tests under anesthesia in anterior, posterior, inferior, anterior-inferior and posterior-inferior direction	a, e, f
Exclusion: treatable/observable rotator cuff lesions; multidirectional instability					
Speer et al, ¹⁹ 1994	Shoulder surgery; subtle anterior instability	100 (NA)	NA	Relocation test apprehension test	a, e
Exclusion: treatable/observable rotator cuff lesions; multidirectional instability					

Abbreviations: MRI, magnetic resonance imaging; NA, not available; SLAP, superior labrum anterior posterior.

^aLimitations pertaining to all listed studies: spectrum bias possible, patient on the list for surgery or arthroscopy, and blinding unclear; the reference test might have been interpreted with knowledge of the index test or vice versa. Key to limitations: (a) Selection criteria for waiting list entry not described. (b) Disease progression bias possible; time between index and reference test not described. (c) Partial verification bias; part of the sample did not receive the reference test. (d) Incorporation bias; results of index test are used to establish the final diagnosis. (e) The execution of the reference test was not described, causing problems with study replication. (f) Unclear whether same clinical data (radiography, MRI, or other diagnostic tools) would be available in daily practice. (g) Unclear whether uninterpretable or intermediate test results were reported. (h) Unclear whether all patients entering study were accounted for (withdrawals). Limitations of the studies were determined with the Quality Assessment of Diagnostic Accuracy Studies standardized checklist.²⁷

^bAn additional 178 patients retrospectively excluded for various reasons.

^cFive patients removed for cohort according to physical findings.

^dAn additional 18 patients retrospectively excluded for dual diagnoses.

Table 44-3 Diagnostic Accuracy of Physical Examination for Instability of the Shoulder

Index Test and Source	Diagnosis	No. of Shoulders	Sensitivity ^a	Specificity ^a	LR+ (95% CI)	LR- (95% CI)
Provocation Tests						
Apprehension test						
T'Jonck et al, ³⁸ 2001	Instability	72	0.88 (23/26)	0.50 (23/46)	1.8 (1.3-2.5)	0.23 (0.08-0.69)
Speer et al, ¹⁹ 1994	Subtle anterior instability					
Pain		100	0.54	0.44
Apprehension		100	0.68	1.0
Relocation test						
T'Jonck et al, ³⁸ 2001	Instability	72	0.85 (22/26)	0.87 (40/46)	6.5 (3.0-14)	0.18 (0.07-0.45)
Speer et al, ¹⁹ 1994	Subtle anterior instability					
Pain		100	0.30	0.58
Apprehension		100	0.57	1.0
Clunk test						
T'Jonck et al, ³⁸ 2001	Instability	72	0.35 (9/26)	0.98 (45/46)	16 (2.1-119)	0.67 (0.5-0.89)
Anterior release test						
T'Jonck et al, ³⁸ 2001	Instability	72	0.85	0.87
Gross and Distefano, ³⁵ 1997	Occult instability	100	0.92 (34/37)	0.89 (40/45)	8.3 (3.6-19)	0.09 (0.03-0.27)
Laxity Tests						
Load and shift posterior test						
T'Jonck et al, ³⁸ 2001	Instability	72	0 (0/26)	1.0 (46/46)	1.7 (0-83)	0.99 (0.93-1.1)
Sulcus sign						
T'Jonck et al, ³⁸ 2001	Instability	72	0.31 (8/26)	0.89 (41/46)	2.8 (1.0-7.7)	0.78 (0.59-1.0)
Load and shift anterior test						
T'Jonck et al, ³⁸ 2001	Instability	72	0.54 (14/26)	0.78 (36/46)	2.5 (1.3-4.8)	0.59 (0.38-0.92)
Examination under anesthesia						
Cofield et al, ³³ 1993	Instability	55	1.0 (25/25)	0.93 (28/30) ^c	13 (3.9-43)	0.02 (0-0.31)
Oliashirazi et al, ³⁷ 1999	Anterior instability	60	0.83 (25/30)	1.0 (30/30)	51 (3.2-80)	0.18 (0.08-0.38)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

^aIf data of the 2 × 2 table were presented in the study, the sensitivity and specificity calculations are shown in parentheses.

^bEllipses indicate data not available.

^cThe healthy contralateral shoulders of the subjects (n = 30) were used as control. Hence, the specificity value and likelihood ratios have been presumably overestimated.

test were missing or unclear in 9 studies.^{16,19,23,28,29,32,33,35,37} Furthermore, in 16 studies it was unclear whether the examiner of the reference test was blinded for the index test^{16,18,19,21-23,28-36,38}; in 1 study it was evident that the examiner was not blinded.³⁷ These methodologic problems complicate reproduction of study results and may have biased the outcome.

CLINICAL SCENARIO—RESOLUTION

Primary care physicians may consider the diagnosis of instability with or without a labral tear for this 24-year-old. The history of trauma at a young age and recurrent shoulder problems associated with a symptom that might have represented an acute dislocation (pop with an excessive stretch) mean that the attending physician may consider clinical tests to assess for instability and labral tears, but diagnostic accuracy would still be uncertain. Because the patient might opt for surgery to prevent recurrent dislocation, the primary care physician might consult an orthopedist to confirm the diagnosis and optimal management strategies for this patient's case.

THE BOTTOM LINE

The available evidence suggests that the relocation test and the anterior release test are best for establishing diagnosis of instability. For labral tears, the biceps loads I and II tests, the pain provocation test of Mimori, and the internal rotation resistance strength test have the best diagnostic performance characteristics (Figure 44-4). However, these results are based on single studies done in groups of selected patients who were evaluated by specialists. Despite the high prevalence of shoulder disorders in the general population, we are uncertain whether the diagnostic value of these tests or combinations thereof will be similar when used in primary care. Nonetheless, an understanding of the tests used in a specialist practice gives primary care physicians the opportunity to focus on physical examination maneuvers that might improve diagnostic skills. Although we recommend that clinicians take a careful history of the mechanism of shoulder injury, the role of the patient's medical history in diagnosing the presence of instability or labral tears has not been studied. A comparison of relevant historical characteristics of

Table 44-4 Diagnostic Accuracy of Physical Examination for Labral Tears

Index Test	Diagnosis	No. of Shoulders	Sensitivity ^a	Specificity ^{a,b}	LR+ (95% CI) ^b	LR- (95% CI) ^b
Anterior Apprehension						
Guanche and Jones, ¹⁶ 2003	Labral tears (including SLAP)	60	0.40	0.87
Guanche and Jones, ¹⁶ 2003	SLAP lesions	60	0.30	0.63
Active Compression (O'Brien Test)						
Stetson and Templin, ²¹ 2002	Labral tears	65	0.54 (14/26)	0.31 (12/39)	0.8 (0.5-1.2)	1.5 (0.8-2.8)
O'Brien et al, ²⁹ 1998	Labral tears	206	1.0 (53/53)	0.98 (150/153)	21 (10-42)	0.01 (0-0.16)
O'Brien et al, ²⁹ 1998	Acromial joint pathology	212	1.0 (55/55)	0.96 (150/157)	44 (16-123)	0.01 (0-0.16)
McFarland et al, ²² 2002	SLAP lesions	409 ^c	0.47 (18/38)	0.55 (203/371)	1.0 (0.7-1.4)	0.96 (0.70-1.3)
Guanche and Jones, ¹⁶ 2003	SLAP lesions	60	0.54	0.47
Guanche and Jones, ¹⁶ 2003	Labral tears (including SLAP)	60	0.63	0.73
Anterior Slide						
Kibler, ³⁴ 1995	Superior glenoid labral tear	226	0.78 (69/88)	0.92 ^d (125/138)	8.3 (4.9-14)	0.24 (0.16-0.36)
McFarland et al, ²² 2002	SLAP lesions	419 ^c	0.07 (3/38)	0.83 (62/381)	0.5 (0.2-1.5)	0.99 (1.1-1.2)
Biceps Load I						
Kim et al, ³² 1999	SLAP lesions	74	0.83 (10/12)	0.98 (62/63)	29 (7.3-115)	0.09 (0.01-0.58)
Biceps Load II						
Kim et al, ²³ 2001	SLAP lesions	127	0.90 (35/38)	0.96 (85/89)	26 (8.6-80)	0.11 (0.04-0.28)
Compression Rotation						
McFarland et al, ²² 2002	SLAP lesions	303 ^c	0.24 (7/29)	0.76 (207/274)	1.0 (0.5-2.0)	1.0 (0.81-2.1)
Crank						
Liu et al, ²⁸ 1996	Labral tears	62	0.91 (29/32)	0.93 (28/30)	14 (3.5-52)	0.10 (0.03-0.29)
Stetson and Templin, ²¹ 2002	Labral tears	65	0.46 (12/26)	0.56 (22/39)	1.1 (0.6-1.9)	0.95 (0.61-1.5)
Guanche and Jones, ¹⁶ 2003	Labral tears (including SLAP)	60	0.40	0.73
Guanche and Jones, ¹⁶ 2003	SLAP lesions	60	0.39	0.67
Internal Rotation Resistance Strength						
Zaslav, ³¹ 2001	Internal articular derangement	110	0.88 (23/26)	0.96 (81/84)	25 (8.1-76)	0.12 (0.04-0.35)
Pain Provocation Test of Mimori						
Mimori et al, ¹⁸ 1999	Superior labral tears	32	1.0 (22/22)	0.90 (9/10)	7.2 (1.6-32)	0.03 (0-0.47)
Relocation						
Guanche and Jones, ¹⁶ 2003	Labral tears (including SLAP)	60	0.44	0.87
Guanche and Jones, ¹⁶ 2003	SLAP lesions	60	0.36	0.63
SLAP-Prehension						
Berg and Ciullo, ³⁶ 1998	SLAP lesions	66	0.82 (54/66)
Tenderness of Bicipital Groove						
Guanche and Jones, ¹⁶ 2003	Labral tears (including SLAP)	60	0.44	0.40
Guanche and Jones, ¹⁶ 2003	SLAP lesions	60	0.48	0.52
Test of Speed						
Bennett, ³⁰ 1998	Biceps pathology (including labral tears)	46	0.90 (9/10)	0.14 (5/36)	1.1 (0.8-1.3)	0.72 (0.10-5.5)
Guanche and Jones, ¹⁶ 2003	Labral tears (including SLAP)	60	0.18	0.87
Guanche and Jones, ¹⁶ 2003	SLAP lesions	60	0.09	0.74
Test of Yergason						
Guanche and Jones, ¹⁶ 2003	Labral tears (including SLAP)	60	0.09	0.93
Guanche and Jones, ¹⁶ 2003	SLAP lesions	60	0.12	0.96

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio; SLAP, superior labrum anterior posterior.

^aIf data of the 2 × 2 table were presented in the study, the sensitivity and specificity calculations are shown in parentheses.

^bEllipses indicate data not available.

^cThe authors stated in their article that patient numbers for each test were not equal because tests were published at different times (namely, the compression rotation test, 1990; the anterior slide test, 1995; and the active compression test, 1998).

^dThe healthy contralateral shoulders of the subjects were used as control. Hence, the specificity value and LRs have been presumably overestimated.

patients with shoulder complaints, physical examination findings, and noninvasive images (eg, magnetic resonance imaging), along with arthroscopy or surgical results, would greatly enhance the knowledge base of primary care physicians who are first to evaluate shoulder conditions.

Author Affiliations at the Time of the Original Publication

The Netherlands Expert Centre for Work-related Musculoskeletal Disorders (Drs Miedema and Kuiper and Ms Luime), Department of General Practice (Drs Verhagen and Koes and Ms Luime), Department of Public Health (Dr Burdorf), and Department of Orthopedics (Dr Verhaar), Erasmus Medical Center, Rotterdam, The Netherlands.

Acknowledgment

We thank David L. Simel, MD, MHS, for his critical comments on the manuscript.

REFERENCES

1. Matsen FA, Thomas SC, Rockwood CA. Anterior glenohumeral instability. In: Rockwood CA, Matsen FA, ed. *The Shoulder*. Volume 1. Philadelphia, PA: WB Saunders Co; 1990:526-622.
2. Pope DP, Croft PR, Pritchard CM, Silman AJ. Prevalence of shoulder pain in the community: the influence of case definition. *Ann Rheum Dis*. 1997;56(5):308-312.
3. Urwin M, Symmons D, Allison T, et al. Estimating the burden of musculoskeletal disorders in the community. *Ann Rheum Dis*. 1998;57(11):649-655.
4. Makela M, Heliovaara M, Sainio P, et al. Shoulder joint impairment among Finns aged 30 years or over. *Rheumatology*. 1999;38(7):656-662.
5. Natvig B, Nassiøy I. Musculoskeletal complaints in a population: occurrence. *Tidsskr Nor Laegeforen*. 1994;114(3):323-327.
6. Croft P, Pope D, Silman A; Primary Care Rheumatology Society Shoulder Study Group. The clinical course of shoulder pain. *BMJ*. 1996;313(7057):601-602.
7. Van der Windt DA, Koes BW, de Jong BA. Shoulder disorders in general practice. *Ann Rheum Dis*. 1995;54(12):959-964.
8. Picavet HS, Schouten JS. Musculoskeletal pain in the Netherlands. *Pain*. 2003;102(1-2):167-178.
9. Macfarlane GJ, Hunt IM, Silman AJ. Predictors of chronic shoulder pain: a population based prospective study. *J Rheumatol*. 1998;25(8):1612-1615.
10. Pollock RG, Flatow EL. Classification and evaluation. In: Bigliani LU, ed. *The Unstable Shoulder*. Rosemont, IL: American Academy of Orthopaedic Surgeons; 1996:25-36.
11. Warner JJ. Overview: avoiding pitfalls and managing complications and failures of instability surgery. In: Warner JJ, Iannotti JP, Gerber C, eds. *Complex and Revision Problems in Shoulder Surgery*. Philadelphia, PA: Lippincott-Raven Publishers; 1997:3-8.
12. Gerber C. Observations on the classification of instability. In: Warner JJ, Iannotti JP, Gerber C, eds. *Complex and Revision Problems in Shoulder Surgery*. Philadelphia, PA: Lippincott-Raven Publishers; 1997:9-18.
13. Hovelius L. Incidence of shoulder dislocation in Sweden. *Clin Orthop*. 1982;(166):127-131.
14. Snyder SJ. Labral lesions (non-instability) and SLAP lesions. In: Snyder SJ. *Shoulder Arthroscopy*. New York, NY: McGraw-Hill; 1994:115-131.
15. Snyder SJ, Karzel RP, Del Pizzo W, Ferkel RD, Friedman MJ. SLAP lesions of the shoulder. *Arthroscopy*. 1990;6(4):274-279.
16. Guanche CA, Jones DC. Clinical testing for tears of the glenoid labrum. *Arthroscopy*. 2003;19(5):517-523.
17. Van der Helm FC. A finite element musculoskeletal model of the shoulder mechanism. *J Biomech*. 1994;27(5):551-569.
18. Mimori K, Muneta T, Nakagawa T, Shinomiya K. A new pain provocation test for superior labral tears of the shoulder. *Am J Sports Med*. 1999;27(2):137-142.
19. Speer KP, Hannafin JA, Altchek DW, Warren RF. An evaluation of the shoulder relocation test. *Am J Sports Med*. 1994;22(2):177-183.
20. McFarland EG, Torpey BM, Curl LA. Evaluation of shoulder laxity. *Sports Med*. 1996;22(4):264-272.
21. Stetson WB, Templin K. The Crank test, the O'Brien test, and routine magnetic resonance imaging scans in the diagnosis of labral tears. *Am J Sports Med*. 2002;30(6):806-809.
22. McFarland EG, Kim TK, Savino RM. Clinical assessment of three common tests for superior labral anterior-posterior lesions. *Am J Sports Med*. 2002;30(6):810-815.
23. Kim SH, Ha KI, Ahn JH, Kim SH, Choi HJ. Biceps load test II. *Arthroscopy*. 2001;17(2):160-164.
24. Deville WL, Buntinx F, Bouter LM, et al. Conducting systematic reviews of diagnostic studies: didactic guidelines. *BMC Med Res Methodol*. 2002;2:9.
25. Deville WL, Bezemer PD, Bouter LM. Publications on diagnostic test evaluation in family medicine journals. *J Clin Epidemiol*. 2000;53(1):65-69.
26. Dinnis J, Loveman E, McIntyre L, Waugh N. The effectiveness of diagnostic tests for the assessment of shoulder pain due to soft tissue disorders: a systematic review. *Health Technol Assess*. 2003;7(29):1-166.
27. Whiting P, Rutjes AW, Reitsma JB, Bossuyt PM, Kleijnen J. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Med Res Methodol*. 2003;3:25.
28. Liu SH, Henry MH, Nuccion SL. A prospective evaluation of a new physical examination in predicting glenoid labral tears. *Am J Sports Med*. 1996;24(6):721-725.
29. O'Brien SJ, Pagnani MJ, Fealy S, McGlynn SR, Wilson JB. The active compression test. *Am J Sports Med*. 1998;26(5):610-613.
30. Bennett WF. Specificity of the Speed's test. *Arthroscopy*. 1998;14(8):789-796.
31. Zaslav KR. Internal rotation resistance strength test: a new diagnostic test to differentiate intra-articular pathology from outlet (Neer) impingement syndrome in the shoulder. *J Shoulder Elbow Surg*. 2001;10(1):23-27.
32. Kim SH, Ha KI, Han KY. Biceps load test. *Am J Sports Med*. 1999;27(3):300-303.
33. Cofield RH, Nessler JP, Weinstabl R. Diagnosis of shoulder instability by examination under anesthesia. *Clin Orthop*. 1993;(291):45-53.
34. Kibler WB. Specificity and sensitivity of the anterior slide test in throwing athletes with superior glenoid labral tears. *Arthroscopy*. 1995;11(3):296-300.
35. Gross ML, Distefano MC. Anterior release test. *Clin Orthop*. 1997;(339):105-108.
36. Berg EE, Ciullo JV. A clinical test for superior glenoid labral or "SLAP" lesions. *Clin J Sport Med*. 1998;8(2):121-123.
37. Oliashirazi A, Mansat P, Cofield RH, Rowland CM. Examination under anesthesia for evaluation of anterior shoulder instability. *Am J Sports Med*. 1999;27(4):464-468.
38. T'Jonck L, Staes F, Smet L, Lysens R. The relationship between clinical shoulder tests and the findings in arthroscopic examination. *Geneeskunde Sport*. 2001;34:15-24.
39. Lyons AR, Tomlinson JE. Clinical diagnosis of tears of the rotator cuff. *J Bone Joint Surg Br*. 1992;74(3):414-415.
40. Leroux JL, Thomas E, Bonnel F, Blotman F. Diagnostic value of clinical tests for shoulder impingement syndrome. *Rev Rheum Engl Ed*. 1995;62(6):423-428.
41. Ure BM, Tiling T, Kirchner R, Rixen D. Zuverlässigkeit der klinischen Untersuchung der Schulter im Vergleich zur Arthroskopie: eine prospektive Studie. *Unfallchirurg*. 1993;96(7):382-386.
42. Hertel R, Ballmer FT, Lambert SM, Gerber C. Lag signs in the diagnosis of rotator cuff rupture. *J Shoulder Elbow Surg*. 1996;5(4):307-313.
43. Litaker D, Pioro M, El Bilbeisi H, Brems J. Returning to the bedside. *J Am Geriatr Soc*. 2000;48(12):1633-1637.
44. MacDonald PB, Clark P, Sutherland K. An analysis of the diagnostic accuracy of the Hawkins and Neer subacromial impingement signs. *J Shoulder Elbow Surg*. 2000;9(4):299-301.
45. Walch G, Boulahia A, Calderone S, Robinson AH. The "dropping" and "hornblower's" signs in evaluation of rotator-cuff tears. *J Bone Joint Surg Br*. 1998;80(4):624-628.
46. Calis M, Akgun K, Birtane M, Karacan I, Calis H, Tuzun F. Diagnostic values of clinical diagnostic tests in subacromial impingement syndrome. *Ann Rheum Dis*. 2000;59(1):44-47.
47. Read JW, Perko M. Shoulder ultrasound. *J Shoulder Elbow Surg*. 1998;7(3):264-271.

48. Kaneko K, De Mouy EH, Brunet ME. Massive rotator cuff tears. *Clin Imaging*. 1995;19(1):8-11.
49. Rahme H, Solem-Bertoft E, Westerberg CE, et al. The subacromial impingement syndrome. *Scand J Rehabil Med*. 1998;30(4):253-262.
50. McFarland EG, Neira CA, Gutierrez MI, Cosgarea AJ, Magee M. Clinical significance of the arthroscopic drive-through sign in shoulder surgery. *Arthroscopy*. 2001;17(1):38-43.
51. Adolfsson L, Lysholm J. Arthroscopy and stability testing for anterior shoulder instability. *Arthroscopy*. 1989;5(4):315-320.
52. Nelson MC, Leather GP, Nirschl RP, Pettrone FA, Freedman MT. Evaluation of the painful shoulder. *J Bone Joint Surg Am*. 1991;73(5):707-716.
53. Westerberg CE, Solem-Bertoft E, Lundh I. The reliability of three active motor tests used in painful shoulder disorders. *Scand J Rehabil Med*. 1996;28(2):63-70.
54. Nove-Josserand L, Levigne C, Noel E, Walch G. Isolated lesions of the subscapularis muscle: a propos of 21 cases [in French]. *Rev Chir Orthop Reparatrice Appar Mot*. 1994;80(7):595-601.
55. Gazielly DF, Gleyze P, Montagnon C, Bruyere G, Prallet B. Functional and anatomical results after surgical treatment of ruptures of the rotator cuff [in French]. *Rev Chir Orthop Reparatrice Appar Mot*. 1995;81(1):8-16.
56. Lerat JL, Chotel F, Besse JL, Moyen B, Brunet Guedj E. Dynamic anterior jerk of the shoulder [in French]. *Rev Chir Orthop Reparatrice Appar Mot*. 1994;80(6):461-467.

Prepared by Catherine P. Kaminetzky, MD, MPH

Reviewed by David L. Simel, MD, MHS,
and Jolanda J. Luime, PhD

CLINICAL SCENARIO

A 24-year-old man with shoulder pain had a shoulder injury when he was 16 years old. For the last 3 years, he experienced sudden right shoulder discomfort and felt a pop every time he tried to throw a baseball with excessive force. However, the discomfort always resolved on its own. He has started to play tennis, and shoulder pain is affecting his performance.

Inspection and palpation of the shoulder reveals no abnormalities. He has no neck discomfort or limitation in neck range of motion. Although he has full range of external and internal rotation of the shoulders, the right shoulder causes some discomfort throughout the arc of motion. You decide to assess for instability of the shoulder.

UPDATED SUMMARY ON SHOULDER INSTABILITY AND LABRAL TEARS

Original Review

Luime JJ, Verhagen AP, Miedema HS, et al. Does this patient have an instability of the shoulder or a labrum lesion? *JAMA*. 2004;292(16):1989-1999.

UPDATED LITERATURE SEARCH

Our literature search replicated that of the original article, confined to 2004 to April 2006. We identified 89 potential articles and reviewed the abstracts to find articles that included consecutive, prospectively identified patients whose shoulder problems were suspicious for instability or a labral tear and who were assessed by arthroscopy or surgery. No new studies describe the sensitivity and specificity of findings for instability or labral tear symptoms and signs. One study described the precision of various maneuvers for anterior instability.

NEW FINDINGS

- The recommended tests for shoulder instability, the relocation and anterior release tests, may also be the most reproducible.

The reliability improves when apprehension during the maneuver, rather than pain, is used to judge the results as positive vs negative.

Details of the Update

Four members of an orthopedic shoulder clinic team prospectively examined patients referred with shoulder symptoms and a medical history suggestive of instability.¹ Each patient had to be able to endure examinations by each member of the team, resulting in 13 of 25 potentially eligible patients undergoing the complete examinations. The final diagnoses were not reported, but the intraclass correlations were reported for 2 laxity tests (load and shift) and 4 provocation tests (apprehension, relocation, augmentation, and release tests). For the laxity tests, the results were reported on an ordinal scale, and for the provocation tests the results were considered “positive” or “negative” according to a response of patient apprehension or pain. The load and shift tests had good reproducibility for motions in the anterior and inferior direction but not the posterior direction. For provocation tests, the assessment of an apprehensive response to each maneuver was more reproducible than the assessment of a response of pain. Among the 4 tests, the relocation test to assess apprehension (intraclass correlation, 0.71) and the release test to assess apprehension (intraclass correlation, 0.63) were the most reproducible.

A study by Holtby and Razmjou² evaluated a large number of patients referred for shoulder problems ($n = 152$), of whom 50 patients had their disease status confirmed by arthroscopy.² The 2 tests of interest were Speed test and Yergason test, both initially described as tests for bicipital tendonitis. The verification bias and the categorization of disease (any biceps tendon lesion or a superior labral anterior posterior lesion) prohibited assessment of isolated labral tears, but the positive likelihood ratio (LR+) and negative likelihood ratio (LR-) for Speed and Yergason tests had confidence intervals (CIs) that crossed 1. If the data had been corrected for verification bias, the likelihood ratio for a positive Yergason test result (LR+, 2.0; 95% CI, 0.86-4.7) might have appeared more promising.

A systematic review of the incidence and prevalence of shoulder discomfort in the general population provides a context for assessing the likelihood that a patient will have shoulder instability or a labral lesion.³ The annual incidence of shoulder discomfort is 0.9% to 2.5%. However, shoulder discomfort does not

immediately resolve, so prevalence rates are much higher. At any given time, shoulder discomfort is present in 6.9% to 26% of the general population.

IMPROVEMENTS IN THE DATA PRESENTED IN THE ORIGINAL PUBLICATION

None.

CHANGES IN THE REFERENCE STANDARD

None.

RESULTS OF LITERATURE REVIEW

Precision of Tests for Instability or Labral Tears

The sulcus sign and load and shift laxity tests have similar reproducibility (Table 44-5). Assessing a patient's apprehension to maneuvers has greater reliability than assessing his or her pain (Table 44-6).

Table 44-5 Laxity Maneuvers

Tests	Intraclass Correlation Coefficient
Sulcus sign	0.60
Load and shift (at 0-, 20-, and 90-degree arm positions)	
Anterior direction	0.53-0.72
Posterior direction	0.42-0.68
Inferior direction	0.65-0.79

Table 44-6 Provocation Maneuvers

Response to Maneuvers	Intraclass Correlation Coefficient
Apprehensive Response	
Apprehension test	0.47
Relocation	0.71
Augmentation	0.48
Release	0.63
Pain Response	
Apprehension test	0.31
Relocation	0.31
Augmentation	0.09
Release	0.31
Pain or Apprehensive Response	
Apprehension test	0.44
Relocation	0.44
Augmentation	0.33
Release	0.45

EVIDENCE FROM GUIDELINES

No governmental guidelines address the evaluation of patients for shoulder instability.

CLINICAL SCENARIO—RESOLUTION

Shoulder instability, with or without a labral tear, is a diagnostic consideration for this patient with a history of a shoulder injury. The popping sensation is suggestive of instability, but the physical examination maneuvers are more important. The apprehension maneuver should be performed, followed by the relocation test and anterior release tests. The assessment of an apprehensive response to the relocation and anterior release tests is the most reliable provocation test. A positive response increases the likelihood of instability approximately 6 to 8 times, whereas negative responses decrease the likelihood by approximately 0.1 to 0.20 times. Labral tears are assessed through the biceps load tests I and II. These tests differ only by the position of the arm (abduction at 90 degrees for biceps load I and at 120 degrees for biceps load II). An increase in pain on the biceps load tests increases the likelihood of a labral tear by 26 to 29 times, whereas the lack of increased pain decreases the likelihood 0.09 to 0.11 times.

REFERENCES FOR THE UPDATE

1. Tzannes A, Paxinos A, Callanan M, Murrell GAC. An assessment of the interexaminer reliability of tests for shoulder instability. *J Shoulder Elbow Surg.* 2004;13(1):18-23.
2. Holtby R, Razmjou H. Accuracy of the Speed's and Yergason's tests in detecting biceps pathology and SLAP lesions: comparison with arthroscopic findings. *Arthroscopy.* 2004;20(3):231-236.
3. Luime JJ, Koes BW, Hendriksen IJ, et al. Prevalence and incidence of shoulder pain in the general population: a systematic review. *Scand J Rheumatol.* 2004;33(2):73-81.

SHOULDER INSTABILITY—MAKE THE DIAGNOSIS

PRIOR PROBABILITY

There are no adequate data for assessing the prevalence of these conditions among patients with shoulder discomfort because the existing data come only from patients undergoing surgery or arthroscopy. The incidence of shoulder discomfort is 0.9% to 2.5%. However, because shoulder pain can be chronic, the prevalence at a single point in time is 6.9% to 26%.

POPULATION FOR WHOM SHOULDER INSTABILITY OR LABRAL TEARS SHOULD BE CONSIDERED

Patients with shoulder pain should be screened for shoulder instability and labral tears. The annual incidence of shoulder dislocation in the general population may be as high as 1.7%. There are no data for the incidence or prevalence of labral tears.

DETECTING THE LIKELIHOOD OF SHOULDER INSTABILITY OR A LABRAL TEAR

The anterior release and relocation tests have the best measurement properties for shoulder instability (Table 44-7 and Figure 44-3). The assessment of apprehension will be more reliable than the assessment of pain for these maneuvers. The biceps load tests should be performed to assess for labral tears (Table 44-7 and Figure 44-3).

Table 44-7 Likelihood Ratios for Tests of Shoulder Instability or a Labral Tear^a

Test	LR+ (95% CI)	LR- (95% CI)
Shoulder Instability		
Anterior release test	8.3 (3.6-19)	0.09 (0.03-0.27)
Relocation test	6.5 (3.0-14)	0.18 (0.07-0.45)
Labral Tear		
Biceps load I	29 (7.3-115)	0.09 (0.01-0.58)
Biceps load II	26 (8.6-80)	0.11 (0.04-0.28)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

^aThese data come from examinations done by orthopedists and not generalist physicians.

REFERENCE STANDARD TESTS

Arthroscopy or surgery.

This page intentionally left blank

CHAPTER 45

Does This Patient Have Sinusitis?

John W. Williams, Jr, MD, MHS

David L. Simel, MD, MHS

CLINICAL SCENARIO

A patient presents to your office with a “bad cold.” Her symptoms began 5 days ago, when a runny nose, a scratchy throat, generalized malaise, and a nonproductive cough developed. Her symptoms are gradually improving with an over-the-counter cough medicine, but during the past 24 hours a “sinus headache” has developed. The patient is concerned that she may have “sinus.” It is the middle of cold and flu season, and this is the fifth patient you have treated today who has upper respiratory tract symptoms.

WHY IS THIS AN IMPORTANT QUESTION TO ANSWER WITH A CLINICAL EXAMINATION?

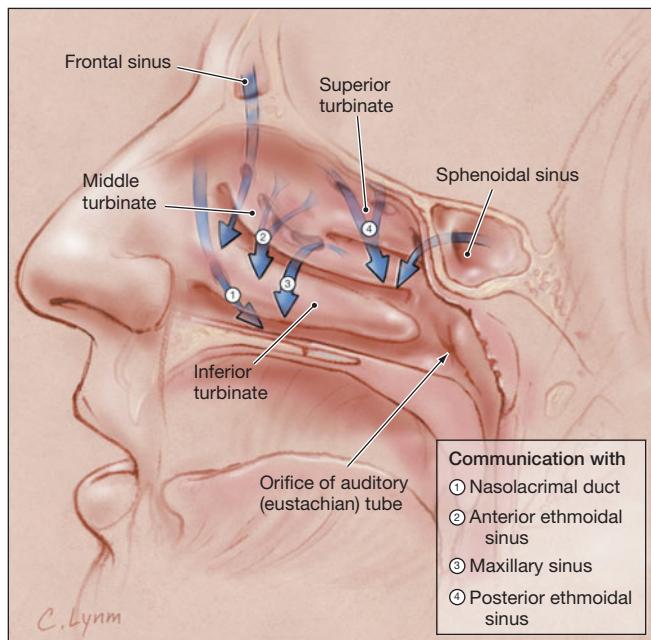
The patient’s story is familiar to primary care clinicians. Among the most frequent diagnoses made by primary care practitioners are nasal problems such as allergic and infectious rhinitis, vasomotor rhinitis, and bacterial sinusitis.¹ Given the constant assault of allergens, environmental pollutants, respiratory viruses, and rapid temperature changes, it is not surprising that nasal complaints are so common. However, not all “sinus” is sinusitis. Sinusitis can be defined simply as inflammation of one or more paranasal sinuses but usually refers to infection of the sinuses. In recent years, many new medications have become available that allow effective medical treatment of sinus problems so that it is important to diagnose nasal complaints accurately to deliver appropriate treatment.² When this can be accomplished by the clinical examination, it obviates the need for more expensive testing such as radiography.

The list of differential diagnoses for patients with nasal congestion or discharge is long (Table 45-1), but a handful of conditions encompass the majority of cases.³ These conditions can be divided into those causing inflammation of the nose (rhinitis) and those causing inflammation of the sinuses (sinusitis). Rhinitis is most frequently due to viral infection, allergens (seasonal or perennial), or vasomotor instability (eg, caused by extreme temperature change or excessive use of vasoconstrictive medications). When these conditions are severe, the sinus ostia may become blocked and the sinuses infected secondarily. However, the implications of diagnosing rhinitis are different from diagnosing sinusitis. Rhinitis may respond to antihistamines, nasal decongestants, nasal steroids, or cromolyn sodium, but randomized trials have shown that sinusitis requires antibiotics for rapid resolution.^{4,5} Sinusitis also occurs as an occult illness that may be associated with asthmatic exacerbations or chronic headache. This overview will focus on the medical history and physical examination findings that distinguish bacterial sinusitis from rhinitis and other conditions.

Table 45-1 Differential Diagnosis of Nasal Congestion/Rhinorrhea

Allergic
Seasonal allergic rhinitis (pollens) ^a
Perennial allergic rhinitis (dusts, molds) ^a
Vasomotor
Idiopathic (vasomotor rhinitis) ^a
Abuse of nose drops (rhinitis medicamentosa) ^a
Drugs (reserpine, guanethidine, prazosin, cocaine abuse)
Psychological stimulation (anger, sexual arousal)
Mechanical
Polyps
Tumor
Deviated septum
Crusting (as in atrophic rhinitis)
Hypertrophied turbinates (chronic vasomotor rhinitis)
Foreign body
Central nervous system fluid leak
Chronic inflammatory
Sarcoidosis
Wegener granulomatosis
Midline granuloma
Infectious
Acute viral infection ^a
Acute or chronic bacterial infection of paranasal sinuses ^a
Atrophic rhinitis (secondary infection)
Hormonal
Pregnancy
Hypothyroidism

^aMost common causes of nasal symptoms.

**Figure 45-1** Sagittal View of Paranasal Sinuses

SINUSITIS REQUIRES ANTIBIOTICS FOR RAPID CURE

Reference Standard for Diagnosing Sinusitis

The reference (or gold) standard for diagnosing infectious sinusitis is sinus aspiration and culture. Its use is particularly appropriate for guiding antibiotic choice in patients with complicated or refractory sinusitis. However, in general practice, sinus radiographs are readily obtained and can be considered a pragmatic reference standard. A 4-view sinus series is highly concordant with a single Waters view,^{6,7} and when it reveals sinus opacity, an air-fluid level, or 6 mm or more of mucosal thickening, a 4-view sinus series is 72% to 96% as accurate for maxillary sinusitis as aspiration and culture respectively.^{8,9} The chief limitations of sinus radiographs are poor visualization of the ethmoid air cells and difficulty distinguishing between infection, tumor, and polyp in the completely opacified sinus. Other potentially useful diagnostic tests are ultrasonography and computed tomography. Ultrasonography is nonionizing but correlates only moderately well with sinus radiographs or sinus aspiration.¹⁰⁻¹² Computed tomography of the sinuses is superior to sinus radiography for visualizing the ethmoid air cells, for evaluating opacified sinuses or mucoceles, and for differentiating the bony changes of chronic inflammation from osteomyelitis.¹³ Sinus computed tomography may become the diagnostic test of choice but is not as readily available as radiographs and has not been evaluated against sinus puncture. This caveat is important because computed tomography may be highly sensitive, yet lack specificity.¹⁴

Normal Anatomy and Pathophysiology of Sinusitis

The nose humidifies, warms, and filters inspired air as it passes through the nasal vestibule and over the nasal turbinates.¹⁵ The nasal turbinates promote turbulent air flow that causes particulate matter to fall on the nasal mucosa, where it is swept by ciliated pseudostratified columnar cells to the nasopharynx. Respiratory epithelium also lines the paranasal sinuses and creates drainage into the nasal cavity via the superior meatus (sphenoid and posterior ethmoid) and middle meatus (maxillary and anterior ethmoids) (Figure 45-1).¹⁶ Properly functioning ciliated cells are critical because maxillary sinus drainage is uphill (Figure 45-2). Patients predisposed to infectious sinusitis may have mucosal edema (eg, allergic rhinitis, viral rhinitis), mechanical obstruction of the meatus (eg, polyps, deviated nasal septum), or impaired ciliary activity (eg, Kartagener syndrome).^{3,17} Under these conditions, viruses and bacteria proliferate in the poorly draining sinus and provoke acute sinusitis.

How to Elicit the Relevant Symptoms and Signs

Although patients may give a simple description, such as "sinus trouble," the examiner should seek a more complete medical history. Symptoms that may increase the likelihood of sinusitis include fever, malaise, cough, nasal congestion, maxillary toothache, purulent nasal discharge, little improvement with nasal decongestants, and headache or facial pain exacerbated by bending forward.

Examination of the nostrils can be performed with a short, wide speculum mounted on a handheld otoscope. The speculum should be directed posterolaterally, avoiding the sensitive nasal septum. The nasal mucosa should be inspected for color, edema, character of nasal secretions, polyps, and structure of the nasal septum (Figure 45-3). Purulent secretion from the middle meatus is reported to be highly predictive of maxillary sinusitis but may be difficult to see unless the examiner shrinks the nasal mucosa with a topical vasoconstrictive agent (eg, oxymetazoline hydrochloride) and uses a nasal speculum to enhance visualization.¹⁸ Septal deviation or nasal polyps are important findings because they may contribute to nasal obstruction and promote recurrent sinusitis.

Palpation for sinus tenderness should be performed over the maxillary and frontal sinuses (Figure 45-4). In addition, checking for tenderness by tapping the maxillary teeth with a tongue blade may be valuable because 5% to 10% of maxillary sinusitis is a result of dental root infection.¹⁹ The ethmoid and sphenoid sinuses cannot be adequately evaluated during the routine physical examination.

Transillumination of the maxillary sinuses may be performed by 2 methods. The best-studied method is performed by placing a Welch-Allyn-Finnoff transilluminator (Welch-Allyn Inc, Skaneateles Falls, New York) over the infraorbital rim, shielding the light source from the observer's eyes, and judging light transmission between sides through the hard palate (Figure 45-5). The examination must be performed in a completely darkened room after allowing the observer's vision to adapt fully to darkness. Obviously, the patient's dentures should be removed. Most experts report the transillumination results as opaque (no light transmission), dull (reduced light transmission), or normal (light transmission typical of a normal subject). An alternative method is to place a light source in the patient's mouth and have the patient make a tight seal around the transilluminator; the observer judges light transmitted through the maxillary sinuses. This technique has the advantage of being able to simultaneously compare sides but requires sterilization of the instrument between patient examinations.

The frontal sinuses can be examined by placing a light source below the supraorbital rim, but interpretation is difficult because the frontal sinuses naturally develop asymmetrically. This normal variation may falsely suggest sinusitis but is resolved by routine radiography.

Precision of Symptoms and Signs

A total of 111 patients with nasal complaints were examined by a general internist and a second examiner who was a physician assistant, internal medicine resident, or attending internist.²⁰ Agreement was high between examiners for 11 of the 15 historical items, including headache (κ , 0.78); subjective fever, chills, or sweats (κ , 0.71); cough (κ , 0.68); colored nasal discharge (κ , 0.68); facial pain (κ , 0.65); and maxillary toothache (κ , 0.60). (Sackett²¹ gives a further explanation of the κ statistic and the other special terms and ideas used in this overview.) On physical examination, agreement was high

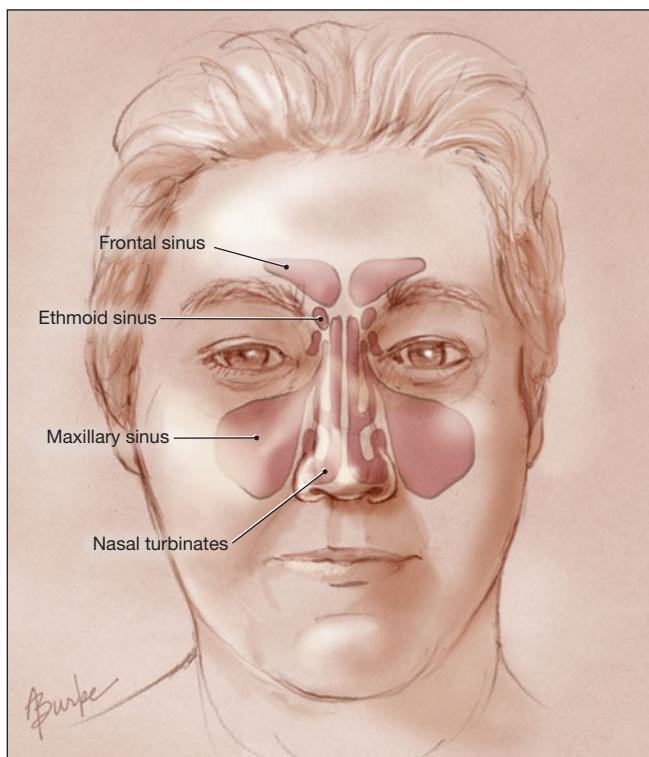


Figure 45-2 Coronal View of Paranasal Sinuses

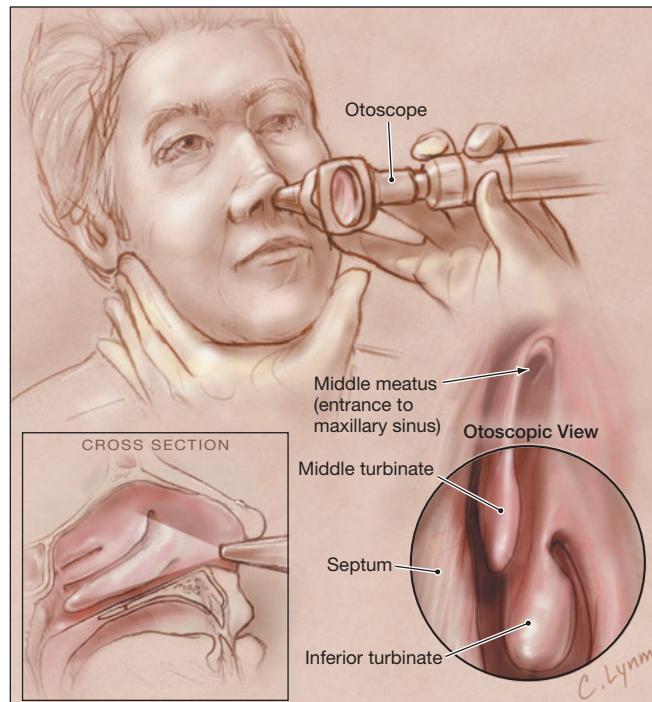


Figure 45-3 Examination of the Nose Through an Otoscope With a Disposable Speculum

The middle meatus is usually not visible behind the turbinates.

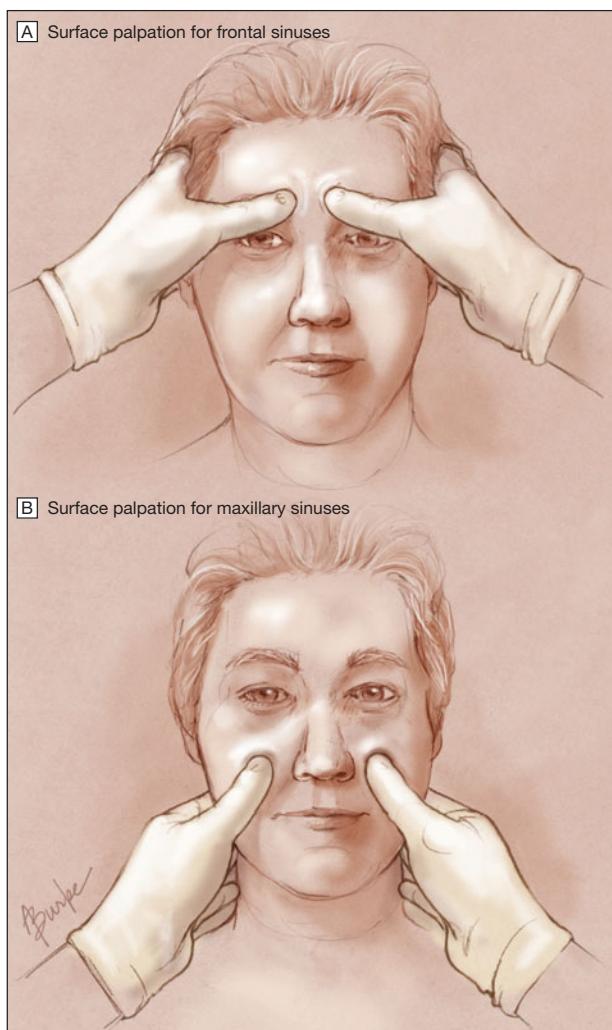


Figure 45-4 Surface Landmarks for Palpation of Frontal Sinuses (Left) and Maxillary Sinuses (Right)

Some experts recommend palpating the frontal sinuses by placing the fingers on the orbital roof below the eyebrow.

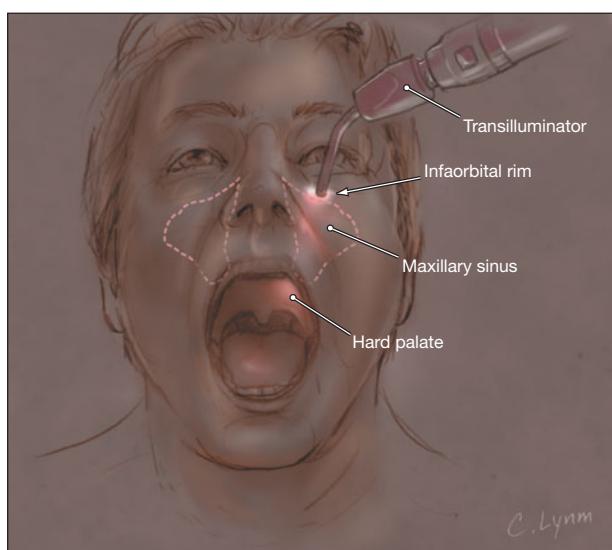


Figure 45-5 Transillumination of the Maxillary Sinus

The light source should be shielded from the examiner's vision with the free hand.

only for sinus tenderness (κ , 0.59) and was fair for maxillary sinus transillumination (simple agreement, 61%; κ , 0.22). In the only other study of observer variability for transillumination, otolaryngologists also had modest agreement between examiners for the maxillary sinuses (simple agreement, 62%), but agreement was good for the frontal sinuses (simple agreement, 95%).²²

Observer agreement is high for most patient symptoms, but for the physical examination agreement is high only for sinus tenderness.

Accuracy of Symptoms and Signs of Sinusitis

There have been few attempts to systematically evaluate the accuracy of the clinical examination for sinusitis. Three studies assessed the discriminative ability of sinusitis symptoms and signs in adults. One evaluated 69 historical items among 164 consecutive patients with sinusitis suspected by the patient or otolaryngologist.²³ These symptoms were compared to a reference standard of 4-view radiography (Caldwell, Waters, lateral, and submental vertex projections). Six symptoms (preceding upper respiratory infection, any nasal discharge or purulent nasal discharge, painful mastication, malaise, cough, and hyposmia) were significantly ($P < .01$) more common in patients with abnormal radiographs, but no single finding was highly accurate.

We compared symptoms to radiograph in 247 consecutive male patients who had rhinorrhea or facial pain unrelated to trauma or who suspected they might have sinusitis.²⁰ Colored nasal discharge, cough, and sneezing were the most sensitive symptoms (72%, 70%, and 70%, respectively) but were not specific (52%, 44%, and 34%, respectively). One symptom, maxillary toothache, was highly specific (93%), but only 11% of patients reported this symptom. Historical items thought to make sinusitis less likely, such as sore throat (sensitivity, 52%; specificity, 56%), itchy eyes (sensitivity, 52%; specificity, 43%), and constitutional symptoms (sensitivity, 56%; specificity, 47%), were not useful.

A third study compared symptoms to ultrasonographic findings in 400 general practice patients selected for study because their physician intended to test or treat for sinusitis.²⁴ Results from this study should be interpreted with caution because the reference standard (ultrasonography) was not interpreted independent of the clinical findings and is less accurate than radiography.^{11,12} In the study by van Duijn et al,²⁴ preceding common cold (sensitivity, 85%; specificity, 28%), pain at bending forward (sensitivity, 65%; specificity, 59%), and purulent rhinorrhea (sensitivity, 62%; specificity, 67%) were the most useful findings. Toothache was found to be highly specific (specificity, 83%).

Studies in children are limited to sensitivities for a few clinical findings. Clear or purulent discharge (sensitivity, 76%-84%) and cough (sensitivity, 48%-80%) are the most sensitive findings (Table 45-2), but the discriminating power of these findings is not known.²⁵⁻²⁸

The most studied but least understood physical examination maneuver is paranasal sinus transillumination.^{5,8,20,22,25,27,29-32} Since the technique was first described in 1889 by Voltolini,³³

its value as a diagnostic test has been hotly debated. Several authors have described transillumination as “highly predictive of disease,” whereas another author has described the use of transillumination as an act of criminal negligence.³⁴ Most studies of transillumination have methodologic limitations, and 2 of the more complete studies had differing results.^{20,30}

Our own study compared the results of transillumination to paranasal sinus radiographs in 247 consecutive patients with nasal symptoms who were treated in general medicine clinics at a Veterans Affairs medical center.²⁰ Transillumination, using a Welch-Allyn-Finnoff transilluminator or Mini MagLite (Mag Instrument Inc, Ontario, California) placed over the infraorbital rim, did little to change the posttest probability of sinusitis. It generated a likelihood ratio (LR) of only 1.6 if either maxillary sinus was dull or opaque and 0.5 if both maxillary sinuses transilluminated normally. Clearly, as a single finding, transillumination could not be relied on to rule in or rule out sinusitis.

The second study included 113 patients with nasal symptoms and abnormal sinus radiographs and found different results.³⁰ In the subset of these patients who were examined by an otolaryngologist (using the same transillumination technique as our study), transillumination was highly useful when the sinus was either completely opaque (LR, ∞) or completely normal (LR, 0.04) but less useful when the finding was dull transillumination (LR, 0.41). In contrast to the previous study, opaque transillumination ruled in sinusitis and normal transillumination ruled out sinusitis.

Why did these 2 studies yield such disparate results? First, the study populations were different (a primary care walk-in clinic vs an otolaryngology clinic) and may have created different degrees of expectation bias. Second, the examiners’ training was different; otolaryngologists may be better transilluminators than general internists. These 2 studies suggest that transillumination may be more useful for diagnosing sinusitis when performed by otolaryngologists.

Because the paranasal sinuses develop at different rates among children, transillumination may be less reliable than in adult patients. Three studies have examined the value of transillumination in children. In one, the examination could not be performed in 24% of the children because of poor patient cooperation.⁵ For the remaining children, there was agreement between transillumination and radiographic findings in 53% and disagreement in 27%, and transillumination was nondiagnostic in 20%.⁵ The other 2 studies reported sensitivities of only 76% (19/25) in one²⁷ and 48% (23/48) in the other, which was performed in children with opaque maxillary sinuses on radiographs who were undergoing sinus drainage for chronic purulent sinusitis.³² The sensitivity of transillumination should have been maximal in this latter patient group with severe disease but nevertheless performed poorly.

Information is limited for other commonly assessed physical examination components. In adults, sinus tenderness was found to have poor sensitivity and specificity (48% to 50% and 62% to 65%, respectively),^{20,24} but other findings (temperature, nasal mucosal color, and percussion tenderness of the maxillary teeth) have not been well studied. In children,

Table 45-2 Sensitivities (%) for Signs and Symptoms of Acute Sinusitis in Children

Sign or Symptom	Source			
	Swischuk et al ²⁵ (n = 63)	Wald et al ²⁶ (n = 30)	McClean ²⁷ (n = 25)	Kogutt and Swischuk ²⁸ (n = 96)
Nasal discharge	76	77	84	77
Cough	60	80	60	48
Headache	48	33	... ^a	...
Fever	46 ^b	63 ^b	12 ^c	21 ^c
Facial pain or swelling	...	30	8 ^d	...
Fetor oris	...	50

^aEllipses indicate information not available.

^bFever not defined.

^cTemperature > 38.3°C (101°F).

^dPain or swelling detected by examination.

tympanic membrane changes from otitis media (sensitivity, 68%) is the most common physical examination finding associated with sinusitis, whereas a documented temperature higher than 38.3°C (101°F) (sensitivity, 12% to 21%) is uncommon.^{27,28}

Accuracy of Combinations of Symptoms and Signs

Despite the poor accuracy of the individual symptoms and signs, these findings used in combination can be diagnostic for sinusitis. We used logistic regression modeling to identify signs and symptoms that best predict sinusitis. This statistical procedure selects findings that independently contribute toward making the diagnosis of sinusitis. Three symptoms (maxillary toothache, poor response to nasal decongestants, and history of colored nasal discharge) and 2 signs (purulent nasal secretion and abnormal transillumination) were the best predictors of sinusitis (Table 45-3).²⁰ When none of these findings were present, sinusitis could be ruled out (LR, 0.1), and when 4 or more were present, the LR was 6.4 (Table 45-4). One study compared 11 clinical findings elicited by experienced otolaryngologists with radiograph and maxillary sinus aspiration in 155 patients presenting to an emergency department with

Table 45-3 Independent Predictors of Sinusitis^a

Symptom or Sign	LR+ (95% CI)	LR- (95% CI)
Maxillary toothache	2.5 (1.2-5.0)	0.9 (0.8-1.0)
Purulent secretion	2.1 (1.5-3.0)	0.7 (0.5-0.8)
Poor response to decongestants	2.1 (1.4-3.1)	0.7 (0.6-0.9)
Abnormal transillumination	1.6 (1.3-2.0)	0.5 (0.4-0.7)
History of colored nasal discharge	1.5 (1.2-1.9)	0.5 (0.4-0.8)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

^aData from Williams et al.²⁰

Table 45-4 Likelihood Ratios by Number of Signs and Symptoms Present^a

No. of Symptoms and Signs	Sinusitis Present	Sinusitis Absent	LR
≥4	16	4	6.4
3	29	18	2.6
2	27	39	1.1
1	14	48	0.5
0	2	32	0.1
Total	88	141	... ^b

Abbreviation: LR, likelihood ratio.

^aSymptoms and signs comprise maxillary toothache, purulent nasal secretion, poor response to decongestants, transillumination (normal bilaterally vs any abnormality), and colored nasal discharge by medical history. Data from Williams et al.²⁰

^bEllipses indicate not applicable.

suspected sinusitis.³⁵ With similar statistical techniques, a history of purulent rhinorrhea or unilateral sinus pain and the presence of pus in the nasal cavity on examination were highly predictive of sinusitis. Maxillary toothache, response to decongestants, and transillumination were not studied.

Physicians appear able to integrate individual signs and symptoms into an overall assessment that accurately diagnoses sinusitis. In our study, an overall impression that sinusitis was “definitely or most likely present” generated an LR of 4.7, and an overall impression that sinusitis was “unlikely or definitely absent” generated a rather low LR of 0.4. When the impression was intermediate, the LR was 1.4.^{20,36} These findings are in agreement with a study that investigated otolaryngologists’ ability to diagnose purulent sinusitis in patients with chronic symptoms. In the study by Berg et al,³⁷ the overall clinical evaluation was compared with sinus aspiration, with the following results: definitely sinusitis, LR = 19; probably sinusitis, LR = 4; probably not sinusitis, LR = 0.14; definitely not sinusitis, LR = 0.19. The general internist’s overall assessment of the likelihood of sinusitis performs well compared with radiograph or sinus aspiration.

To summarize, primary care practitioners frequently evaluate patients with nasal symptoms, and in many instances, sinusitis can be confidently ruled in or ruled out according to the clinical examination. Further studies are needed to examine clinical findings that have not been studied (such as headache when leaning forward) and to test whether the 5 clinical findings found to be useful for adult men can be exported to other patient populations.

THE BOTTOM LINE

1. Sinusitis is insidious in children. Concurrent otitis media is common.
2. Considered in combination, maxillary toothache, poor response to nasal decongestants, abnormal transillumination, and colored nasal discharge by medical history or examination are the most useful clinical findings in primary care populations. When all 5 features are present,

the odds of sinusitis increase sharply (LR, 6.4), and when none are present, sinusitis is ruled out.

3. Transillumination requires a completely darkened room, adequate time for dark adaptation, and practice.
4. The overall medical history and physical examination in symptomatic adult patients is accurate.

Author Affiliations at the Time of the Original Publication

Division of General Internal Medicine, University of Texas Health Sciences Center at San Antonio, and the Ambulatory Care Service, Audie L. Murphy Memorial Veterans’ Hospital, San Antonio, Texas (Dr Williams); Ambulatory Care Service and the Center for Health Services Research in Primary Care, Durham Veterans Affairs Medical Center, and the Division of General Internal Medicine, Duke University Medical Center, Durham, North Carolina (Dr Simel).

Acknowledgments

This work was supported in part by a grant from the A. W. Mellon Foundation and Veterans Affairs Health Services Research and Development grant 89-065.A.

The authors wish to thank Donald Hitch, MD, for reviewing an early draft of this manuscript.

REFERENCES

1. National Center for Health Statistics. *National Ambulatory Medical Care Survey of Visits to General and Family Practitioners, January–December 1975*. Advance Data from Vital and Health Statistics, No. 15. DHEW Pub. No. (HRA) 78-1250. Hyattsville, MD: Public Service; 1977.
2. MacKay I, ed. *Rhinitis: Mechanisms and Management*. Park Ridge, NJ: Parthenon Publishing; 1989:169.
3. Simon HB. Approach to the patient with sinusitis. In: Goroll AH, May LA, Mulley AG, eds. *Primary Care Medicine: Office Evaluation of the Adult Patient*. Philadelphia, PA: JB Lippincott; 1987:883-885.
4. Axelson A, Chidekel N, Grebelius N, Jensen C. Treatment of acute maxillary sinusitis: a comparison of four different methods. *Acta Otolaryngol (Stockh)*. 1970;70(1):71-76.
5. Wald ER, Chiponis D, Ledesma-Medina J. Comparative effectiveness of amoxicillin and amoxicillin-clavulanate potassium in acute paranasal sinus infections in children: a double-blind, placebo-controlled trial. *Pediatrics*. 1986;77(6):795-800.
6. Williams JW Jr, Roberts L, Distell B, Simel DL. Diagnosing sinusitis by x-ray: comparing a single Waters view to 4-view paranasal sinus radiographs. *J Gen Intern Med*. 1992;7(5):481-485.
7. Hayward MW, Lyons K, Ennis WP, Rees J. Radiography of the paranasal sinuses: one or three views? *Clin Radiol*. 1990;41(3):163-164.
8. Evans FO Jr, Sydnor JB, Moore WE, et al. Sinusitis of the maxillary antrum. *N Engl J Med*. 1975;293(15):735-739.
9. Hamory BH, Sande MA, Sydnor A Jr, Seale DL, Gwaltney JM Jr. Etiology and antimicrobial therapy of acute maxillary sinusitis. *J Infect Dis*. 1979;139(2):197-202.
10. Berg O, Carenfelt C. Etiological diagnosis in sinusitis: ultrasonography as clinical complement. *Laryngoscope*. 1985;95(7 pt 1):851-853.
11. Shapiro GG, Furukawa CT, Pierson WE, Gilbertson E, Bierman CW. Blinded comparison of maxillary sinus radiography and ultrasound for diagnosis of sinusitis. *J Allergy Clin Immunol*. 1986;77(1 pt 1):59-64.
12. Rohr AS, Spector SL, Siegel SC, Katz RM, Rachelefsky GS. Correlation between A-mode ultrasound and radiography in the diagnosis of maxillary sinusitis. *J Allergy Clin Immunol*. 1986;78(1 pt 1):58-61.
13. Unger JM, Shaffer K, Duncavage JA. Computed tomography in nasal and paranasal sinus disease. *Laryngoscope*. 1984;94(10):1319-1324.
14. Havas TE, Motbey JA, Gullane PJ. Prevalence of incidental abnormalities on computed tomographic scans of the paranasal sinuses. *Arch Otolaryngol Head Neck Surg*. 1988;114(8):856-859.

15. Guyton AC. *Textbook of Medical Physiology*. Philadelphia, PA: WB Saunders Co; 1986:477.
16. Hollinshead WH, Rosse C. *Textbook of Anatomy*. Philadelphia, PA: Harper & Row; 1985:982.
17. Carson JL, Collier AM, Hu SS. Acquired ciliary defects in nasal epithelium of children with acute viral upper respiratory infections. *N Engl J Med*. 1985;312(8):463-468.
18. Burtoff S. Evaluation of diagnostic methods used in cases of maxillary sinusitis, with a comparative study of recent therapeutic agents employed locally. *Arch Otolaryngol Head Neck Surg*. 1947;45:516-542.
19. Loyal V, Jones J, Noyek A. Management of odontogenic maxillary sinus disease. *Otolaryngol Clin North Am*. 1976;9:213-222.
20. Williams JW Jr, Simel DL, Roberts L, Samsa G. Clinical evaluation for sinusitis: making the diagnosis by history and physical examination. *Ann Intern Med*. 1992;117(9):705-710.
21. Sackett DL. A primer on the precision and accuracy of the clinical examination. *JAMA*. 1992;267(19):2638-2644.
22. Rachelefsky GS, Spector SL. Sinusitis and asthma. *J Asthma*. 1990;27(1):1-3.
23. Axelsson A, Runze U. Symptoms and signs of acute maxillary sinusitis. *ORL J Otorhinolaryngol Relat Spec*. 1976;38(5):298-308.
24. van Duijn NP, Brouwer HJ, Lamberts H. Use of symptoms and signs to diagnose maxillary sinusitis in general practice: comparison with ultrasonography. *BMJ*. 1992;305(6855):684-687.
25. Swischuk LE, Hayden CK Jr, Dillard RA. Sinusitis in children. *Radiographics*. 1982;2:241-252.
26. Wald ER, Milmoe GJ, Bowen A, Ledesma-Medina J, Salamon N, Bluestone CD. Acute maxillary sinusitis in children. *N Engl J Med*. 1981; 304(13):749-754.
27. McClean DC. Sinusitis in children: lessons from twenty-five patients. *Clin Pediatr*. 1970;9(6):342-345.
28. Kogutt MS, Swischuk LE. Diagnosis of sinusitis in infants and children. *Pediatrics*. 1973;52(1):121-124.
29. McNeill RA. Comparison of the findings on transillumination, x-ray and lavage of the maxillary sinus. *J Laryngol Otol*. 1963;77:1009-1013.
30. Gwaltney JM Jr, Sydnor A Jr, Sande MA. Etiology and antimicrobial treatment of acute sinusitis. *Ann Otol Rhinol Laryngol Suppl*. 1981;90(3 pt 3):68-71.
31. Ballantyne JC, Rowe AR. Some points in the pathology, diagnosis and treatment of chronic maxillary sinusitis. *J Laryngol Otol*. 1949;63(6):337-341.
32. Otten FWA, Grote JJ. The diagnostic value of transillumination for maxillary sinusitis in children. *Int J Pediatr Otorhinolaryngol*. 1989;18(1):9-11.
33. Voltolini FER. Die ersten Operationen in der Kehlkopfshöhle vom Munde aus, bei der Durchleuchtung des Kehlkopfes von aussen. *Dtsch Med Wochenschr*. 1889;15:340-343.
34. Smith MEN. Sinus transillumination: an archaic survival. *J Laryngol Otol*. 1961;75:93-94.
35. Berg O, Carenfelt C. Analysis of symptoms and clinical signs in the maxillary sinus empyema. *Acta Otolaryngol (Stockh)*. 1988;105(3-4): 343-349.
36. Simel DL, Feussner JR, DeLong ER, Matchar DB. Intermediate, indeterminate, and uninterpretable diagnostic test results. *Med Decis Making*. 1987;7(2):107-114.
37. Berg O, Bergstedt H, Carenfelt C, Lind MG, Perols O. Discrimination of purulent from nonpurulent maxillary sinusitis: clinical and radiographic diagnosis. *Ann Otol Rhinol Laryngol*. 1981;90(3 pt 1):272-275.

This page intentionally left blank

Prepared by David L. Simel, MD, MHS,
and John W. Williams Jr, MD, MHS
Reviewed by Rowena Dolor, MD, MHS

CLINICAL SCENARIO

A 36-year-old woman reports that she has “sinus headaches” about once every 2 to 3 months. On many days, she thinks she is about to get a sinus headache, but the symptoms resolve. On the day of her visit, she reports pressure in the sinuses, a headache, and nasal congestion that occurred when she woke up. There is no fever, cough, or nasal discharge. She requests an antibiotic. Your examination does not reveal pus in the nares or nasal polyps, though you find she does have some discomfort when you apply pressure to the sinuses. Before you turn off the light to transilluminate the sinuses, what additional lines of inquiry could be explored?

UPDATED SUMMARY ON SINUSITIS

Original Review

Williams JW Jr, Simel DL. Does this patient have sinusitis? diagnosing acute sinusitis by history and physical examination. *JAMA*. 1993;270(10):1242-1246.

UPDATED LITERATURE SEARCH

Our literature search used the parent search strategy for The Rational Clinical Examination series, combined with the subject “exp sinusitis,” published in English from 1992 to June 2004. The results yielded 191 titles, for which we reviewed the titles and abstracts; 22 were selected for additional review. These articles were reviewed to identify studies that assessed the sensitivity and specificity of medical history or physical examination features for sinusitis. We required that the studies be conducted with outpatients, involve prospectively collected data, and use radiologic imaging, endoscopy, or sinus puncture as a criterion standard for acute sinusitis. We excluded studies that had major design biases such as a sample confined to patients with a clinical diagnosis of sinusitis. No new original studies were identified. Two meta-analyses were identified, so the update focuses on those systematic reviews rather than individual studies.

NEW FINDINGS

- Among patients with a suspicion of sinusitis in general medical practice, the prevalence of disease from sinus aspirates is about 50%.
- The radiograph serves as a pragmatic reference standard for primary care practice, correctly diagnosing about 4 of 5 patients.

Details of the Update

Two meta-analyses of essentially the same original studies led their respective authors to distinctly different interpretations about the outcomes, though both reported that the radiographs appeared better than ultrasonography. Engels et al¹ took a pragmatic approach to the reference standard for sinusitis and compared radiographs with sinus puncture and clinical examination with both sinus puncture and radiographs. In addition, they also evaluated varying thresholds for sinus radiograph positivity (opacity, air-fluid level, or mucosal thickening) and risk score for the clinical examination. Not surprisingly, the radiograph had a slightly better summary receiver operating characteristic curve area than the clinical examination (0.83 vs 0.74, respectively), with the authors concluding that evaluating combinations of individual findings as reported in the original Rational Clinical Examination article may perform better than the overall clinical impressions. A reappraisal of the studies reported by Engels et al¹ shows a summary positive likelihood ratio (LR) for radiographs of 4.2 (95% confidence interval [CI], 2.6-6.7) and negative LR of 0.25 (95% CI, 0.17-0.37). From a pragmatic standpoint, using the radiograph as a reference standard will result in the correct classification of 4 of 5 patients compared with sinus puncture. In the 2 original articles comparing radiographs with sinus puncture for general medical patients with a suspicion of sinusitis, the prevalence of sinusitis was 49% and 51%.^{2,3}

Varonen et al⁴ evaluated essentially the same studies, although they counted one study as 2 separate studies, which produced slightly different results. However, the authors did not evaluate varying thresholds of positivity for the radiographs and did not include risk scores for the clinical examination because the scores have not been compared with puncture.² In addition, the authors could not evaluate varying levels of the overall clinical examination (eg, high, intermediate, or low

probability) because they have not been compared with sinus puncture. Despite results that appear clinically similar, the authors concluded that the clinical examination was not reliable and that ultrasonography or radiography should be used “if a correct diagnosis is considered important.”

IMPROVEMENTS IN THE DATA PRESENTED IN THE ORIGINAL PUBLICATION

The original data for LRs, based on the number of signs and symptoms, were given without their CIs. The most important findings were maxillary toothache, purulent nasal secretion, poor response to decongestant, abnormal transillumination result, and patient report of colored nasal discharge. We recalculated the LRs for greater than or equal to 4, 3, 2, 1, and 0 findings present. Patients with greater than or equal to 4 findings have an LR of 6.4 (95% CI, 2.2-19), whereas those with 0 findings have an LR of 0.1 (95% CI, 0.02-0.41).

CHANGES IN THE REFERENCE STANDARD

For clinical research, the reference standard is sinus puncture. However, for clinical care, the radiograph will correctly classify 4 of 5 patients and may serve as a pragmatic standard for evaluating the clinical examination.

RESULTS OF LITERATURE REVIEW

Univariate Findings for Sinusitis

Radiographs perform well compared to the reference standard of sinus puncture ([Table 45-5](#)).

Table 45-5 Likelihood Ratios of Radiographs Compared to Sinus Puncture

Finding	LR+ (95% CI)	LR- (95% CI)
Radiographs vs sinus puncture ^a (n = 6 studies)	4.2 (2.6-6.7)	0.26 (0.17-0.37)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

^aCalculated from the data in the studies as summarized by Engels et al.¹

EVIDENCE FROM GUIDELINES

A panel of experts met to discuss the definition of sinusitis for clinical research and clinical care.⁵ A useful clinical concept was

preference for the word *rhinosinusitis* over *sinusitis* because sinusitis is usually associated with nasal inflammation and rhinitis. The experts did not perform a structured systematic literature review but used consensus-building strategies to derive recommendations. For diagnosing acute bacterial rhinosinusitis, the panel’s expert opinion was based on a combination of 3 “major” findings and 9 “minor” symptoms. The panel accepted a previously proposed case definition for acute rhinosinusitis (that has not been validated), requiring the presence of 2 or more major symptoms (purulent anterior nasal drainage, purulent posterior nasal drainage, or cough) or 1 major and at least 2 minor symptoms (headache, facial pain, periorbital edema, earache, halitosis, tooth pain, sore throat, increased wheeze, fever). The objective documentation of acute bacterial rhinosinusitis requires either visualization of purulent drainage by the clinician or radiographic evidence.

CLINICAL SCENARIO—RESOLUTION

This patient presents with a common set of symptoms. Patients frequently self-diagnose sinusitis, and many will self-medicate or present to their primary care provider with a request for antibiotics. The prevalence of acute bacterial sinusitis among patients who the physician suspects may have the disease is about 50%. However, unless this patient proves to have abnormal maxillary sinus transillumination results, she has none of the symptoms commonly associated with radiographic-proven sinusitis. The probability of sinusitis with none of the 5 findings is about 9%.

The keys to additional lines of inquiry are recognizing that acute bacterial sinusitis is not something that comes and goes within a given day but is more persistent. Migraine headaches frequently begin on awakening and are associated with nasal stuffiness, leading patients to “misdiagnose” themselves. The absence of frank nasal discharge from the medical history and your examination, along with the abrupt onset of symptoms associated with the headache, supports an alternative diagnosis such as vascular headaches. The decision to obtain a sinus radiograph depends on whether you would treat with decongestants, antibiotics, or steroid inhalers for a positive result. If she has an abnormal radiographic result (LR, 4.2), the probability of acute bacterial sinusitis is about 30%, given the absence of clinical findings. You might value a normal radiographic result if it would help in persuading the patient that she does not likely have acute bacterial sinusitis. The probability of acute bacterial sinusitis is less than 3% for a patient with none of the clinical findings and a normal radiographic result.

SINUSITIS—MAKE THE DIAGNOSIS

PRIOR PROBABILITY

Among general medical patients with suspected sinusitis, the prevalence of disease as determined by sinus puncture and culture is 50%.

POPULATION FOR WHOM SINUSITIS SHOULD BE CONSIDERED

Sinusitis may be thought of as “rhinosinusitis” to emphasize the role of nasal symptoms but requires additional clinical research to determine whether the change in terminology requires a change in management approaches. Sinusitis should be considered in patients with nasal stuffiness, nasal discharge, or maxillary facial pain. Many patients will present with a self-suspicion of sinusitis.

DETECTING THE LIKELIHOOD OF SINUSITIS IN ADULTS

The presence of 4 or more findings (maxillary toothache, purulent nasal secretion, poor response to decongestant, abnormal transillumination request, patient report of colored nasal discharge) makes sinusitis much more likely, whereas the absence of any of the findings makes sinusitis unlikely ([Table 45-6](#)).

Table 45-6 Likelihood Ratios for Radiographs and the Clinical Findings for Sinusitis

	LR+ (95% CI)	LR- (95% CI)
Radiographs vs sinus puncture (6 studies) ^a	4.2 (2.6-6.7)	0.26 (0.17-0.37)
Clinical findings compared with sinus radiographs (1 study) ^a		
≥4	6.4 (2.2-19)	
3	2.6 (1.5-4.4)	
2	1.1 (0.73-1.7)	
1	0.47 (0.27-0.80)	
0	0.1 (0.02-0.4)	

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

^aMaxillary toothache, purulent nasal secretion, poor response to decongestant, abnormal transillumination result, patient report of colored nasal discharge.

REFERENCE STANDARD TESTS

Sinus puncture with culture serves as the reference standard for research. Clinicians will prefer to use sinus radiographs, although some patients (approximately 20%) will be misclassified. A recent panel of experts accepts an abnormal radiographic result as evidence of acute bacterial rhinosinusitis for patients with appropriate symptoms.⁵

REFERENCES FOR THE UPDATE

- Engels EA, Terrin N, Barza M, Lau J. Meta-analysis of diagnostic tests for acute sinusitis. *J Clin Epidemiol*. 2000;53(8):852-862.^a
- van Buchem FL, Knottnerus JA, Schrijnemakers VJ, Peeters MF. Primary-care-based randomised placebo-controlled trial of antibiotic treatment in acute maxillary sinusitis. *Lancet*. 1977;349(9053):683-687.
- Laine K, Maätä T, Varonen H, Mäkelä M. Diagnosing acute maxillary sinusitis in primary care: a comparison of ultrasound, clinical examination and radiography. *Rhinology*. 1998;36(1):2-6.
- Varonen H, Mäkelä M, Savolainen S, Läärä E, Hilden J. Comparison of ultrasound, radiography, and clinical examination in the diagnosis of acute maxillary sinusitis: a systematic review. *J Clin Epidemiol*. 2000;53(9):940-948.^a
- Meltzer EO, Hamior DL, Hadley JA, et al. Rhinosinusitis: establishing definitions for clinical research and primary care. *J Allergy Clin Immunol*. 2004;114(6 suppl):S155-S212.

^aFor the Evidence to Support the Update for this topic, see <http://www.JAMAEvidence.com>.

This page intentionally left blank

EVIDENCE TO SUPPORT THE UPDATE:

45

Sinusitis

TITLE Meta-analysis of Diagnostic Tests for Acute Sinusitis.

AUTHORS Engels EA, Terrin N, Barza M, Lau J.

CITATION *J Clin Epidemiol.* 2000;53(8):852-862.

QUESTION With a hierarchy of accuracy based on the reference standard, how well do radiography, ultrasonography, and the clinical examination perform in identifying patients with sinusitis?

DESIGN Systemic review and meta-analysis.

DATA SOURCES Original articles were identified through MEDLINE, along with a review of the reference lists and review articles.

STUDY SELECTION AND ASSESSMENT English-language articles from 1996 to 1998 that met prespecified criteria. Studies had to be among patients with symptoms consistent with sinusitis, and all patients had to undergo evaluation so that verification bias was avoided.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

The relevant tests were radiography, ultrasonography, and the clinical examination. Each test was compared with sinus puncture, when studies were available. Following this “ideal” reference standard study, the authors included studies in which ultrasonography or the clinical examination was compared to radiography as a pragmatic reference standard. No adequate studies of computed tomography or magnetic resonance imaging were identified.

MAIN OUTCOME MEASURES

The studies were assessed for the country, setting, patient characteristics, adequacy of blinding, definition of tests, and number of cut points assessed. A summary receiver operating characteristic (ROC) curve was generated for each comparison, along with summary sensitivity and specificity estimates. An estimate of the likelihood ratio (LR) was estimated from

the summary sensitivity and specificity, rather than calculating from the original data.

MAIN RESULTS

The authors identified 4070 potential articles. From these articles, they found the following that met their inclusion criteria: studies comparing radiology with puncture ($n = 6$), ultrasonography with puncture ($n = 5$), clinical examination with puncture ($n = 1$), ultrasonography with radiology ($n = 3$), clinical examination with radiology ($n = 3$) (Table 45-7). All studies were done in Europe, except for an ultrasonography and a study that compared clinical examination to radiographs done in the United States. Of the 4 clinical studies, only 1 study restricted to children was not included in the original Rational Clinical Examination article.

In the 2 puncture studies of adults in a generalist clinical practice, the prevalence of sinusitis was 49% and 51%.

The diagnostic odds ratio for radiographs is 18 (95% confidence interval [CI], 12-27; $P = .09$ for heterogeneity, with

Table 45-7 Sensitivity and Specificity of Radiographs and the Clinical Examination

Test (No.)	Result	Summary Sensitivity (95% CI)	Summary Specificity (95% CI)	Summary ROC Curve Area
Radiographs vs puncture (6)	Opacity	0.41 (0.33-0.49)	0.85 (0.76-0.91)	0.83
	Fluid or opacity	0.73 (0.60-0.83)	0.80 (0.71-0.87)	
	Fluid, opacity, or mucus membrane thickening	0.90 (0.68-0.97)	0.61 (0.20-0.91)	
Clinical examination vs radiography (3) ^a				0.74

Abbreviations: CI, confidence interval; ROC, receiver operating characteristic.

^aOne study compared the overall clinical impression to sinus radiography, one evaluated a risk score for children, and one evaluated a risk score for adults. The 2 risk score studies show similar points on the ROC curve.

$I^2 = 48\%$). Sinus radiographs vs sinus puncture showed an accuracy of 81%, with reasonably narrow CIs, despite statistical heterogeneity (95% CI, 74%-87%). We calculated these results for the odds ratio and accuracy from data in the original reports (see Table 45-7).

Although the comparison of a clinical risk score with puncture had a summary ROC area of 0.91, the authors identified potential problems with internal validity. The studies comparing ultrasonography with puncture had too much variability for adequate ROC curve assessment, whereas those compared with radiography were so close together that a curve could not describe the points.

CONCLUSIONS

LEVEL OF EVIDENCE Systematic review.

STRENGTHS The inclusion criteria are well specified and inclusive.

LIMITATIONS A quality score was not assigned, and some studies were included that lacked appropriate blinding or description of the factors that defined a positive result. Tests for homogeneity were not done, nor were summary estimates of the LRs given with their CIs.

Some readers will be uncomfortable with the decision to pool studies of varying quality and that were potentially biased by the lack of blinding and case definitions. The authors used the summary sensitivity and specificity to estimate the positive and negative LRs (LR+ and LR-) of radiographs (LR+, 3.7; LR-, 0.34) for “fluid or opacity” compared with sinus puncture. The absence of fluid, opacity, or mucosal thickening decreased the LR- to 0.16, but the specificity had broad CIs. On the other hand, the CIs suggest that the LR- could be much lower and that, even with a prevalence of 50%, a completely normal radiograph result would greatly decrease the probability of sinusitis. From a pragmatic standpoint, using radiographs as the reference standard will result in the misclassification of about 20% of patients.

Sinus ultrasonography is a test primarily used in Europe. Because of the substantial variability in results, the authors infer that ultrasonography may require experience that is more extensive before clinicians can rely on it. No studies of computed tomography were of sufficient quality to meet their inclusion criteria.

The authors conclude that the clinical examination does have “moderate” ability to identify patients with sinusitis. They recommend further evaluation of risk scores for children and adults because they are less reliant on the experience of the examining clinician.

Reviewed by David L. Simel, MD, MHS

TITLE Comparison of Ultrasound, Radiography, and Clinical Examination in the Diagnosis of Acute Maxillary Sinusitis: A Systematic Review.

AUTHORS Varonen H, Mäkelä M, Savolainen S, Läärä E, Hilden J.

CITATION *J Clin Epidemiol.* 2000;53(9):940-948.

QUESTION Compared with sinus puncture or computed tomography (CT), how well do radiographs, ultrasonography, and the overall clinical examination perform in diagnosing acute maxillary sinusitis?

DESIGN Systemic review and meta-analysis.

DATA SOURCES Original articles were identified through MEDLINE (1996 to April 1999) and a Finnish database (Medic), along with a review of the reference lists, review articles, and hand searching of 4 relevant journals.

STUDY SELECTION AND ASSESSMENT Included studies compared the tests of interest with sinus puncture or CT for adults with suspected acute maxillary sinusitis and symptoms of fewer than 3 months’ duration.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

Radiograph results were considered abnormal when the patient had at least 6-mm mucosal thickening, an air-fluid level, or a complete opacity. Ultrasonographic results were considered abnormal according to previously published criteria.¹ The overall clinical impression was evaluated as positive or negative for sinusitis.

MAIN OUTCOME MEASURES

The studies were assessed for validity with standard criteria from the Cochrane Collaboration. A summary receiver operating characteristic (ROC) curve was generated for each comparison, along with summary sensitivity and specificity estimates. Fixed-effects summary likelihood ratios (LRS), without confidence intervals (CIs), were provided. The authors did not provide quantitative estimates of heterogeneity.

MAIN RESULTS

The authors identified 1054 potential articles. From these articles, they found 11 articles that met their inclusion; all were studies compared with sinus puncture, except for 1 study that used CT. The LRs were similar for radiographs, ultrasonography, and clinical examination (Table 45-8).

Despite clinically similar LRs, the authors observed that ultrasonographic findings were heterogeneous. They also report that, as the prevalence of disease decreased in studies, the sensitivity decreased.

Table 45-8 Likelihood Ratio for Sinusitis for Radiographs, Ultrasonography, and the Clinical Examination

Test (No.)	Summary Sensitivity (95% CI)	Summary Specificity (95% CI)	Summary LR+	Summary LR-
Radiographs (7)	0.87 (0.85-0.88)	0.89 (0.88-0.91)	3.4	0.26
Ultrasonography (7)	0.85 (0.84-0.87)	0.82 (0.80-0.83)	2.8	0.30
Clinical examination (2)	0.69 (0.65-0.73)	0.79 (0.75-0.82)	3.2	0.40

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

CONCLUSIONS

LEVEL OF EVIDENCE Systematic review.

STRENGTHS The inclusion criteria are well specified and inclusive.

LIMITATIONS A quality score was not assigned, and some studies were included that lacked appropriate blinding or description of the factors that defined a positive result. Tests for homogeneity were not reported. For the summary ROC curve, the authors do not allow “gradations” of positivity for the tests of interest. Thus, information may have been lost in studies that dichotomize the overall clinical impression into “positive” or “negative.” In addition, the authors used fixed-effects measures without CIs for the summary LR.

The included studies are almost identical to the meta-analysis published by Engels et al.² Not surprisingly, the results

are similar, with differences explained more by the methods used for summarizing results than the results themselves. For radiographs, the only difference in the studies included is that the data from one study were broken out in this meta-analysis into 2 separate studies with different results for radiographs (likewise, they broke out the ultrasonographic data from this single study as if they were 3 separate studies).¹ This likely created some bias in the outcomes.

These authors concluded, as did Engels et al,² that there was too much heterogeneity for ultrasonography and that radiographs may perform better.

Despite LRs that are not appreciably different from radiographs, the authors conclude that the clinical examination is not reliable and that radiographs should be used when a “correct diagnosis is required.” Given that the clinical examination was used to select the patients for radiographs and that no CIs were provided for the LR, it is difficult to conclude that the clinical examination is useless. Furthermore, analyzing the clinical examination as either positive or negative may dilute the efficiency of the clinical examination when patients with an estimated intermediate probability of disease are forced into the positive or negative categories.

REFERENCES FOR THE EVIDENCE

- Revonta M. Ultrasound in the diagnosis of maxillary and frontal sinusitis. *Acta Otolaryngol (Stockh)*. 1980;370(suppl):1-55.
- Engels EA, Terrin N, Barza M, Lau J. Meta-analysis of diagnostic tests for acute sinusitis. *J Clin Epidemiol*. 2000;53(8):852-862.

Reviewed by David L. Simel, MD, MHS

This page intentionally left blank

CHAPTER 46

Does This Patient Have Splenomegaly?

Steven A. Grover, MD, MPA, FRCPC

Alan N. Barkun, MD, FRCPC

David L. Sackett, MD, FRSC, FRCPC

CLINICAL SCENARIO

Among the patients you are seeing today are the following 3:

The first is an elderly woman who complains of easy fatigability, and her conjunctivae and nail beds are pale. You suspect that she is anemic because of gastrointestinal blood loss, but among your differential diagnoses you consider a lymphoproliferative disorder and decide to examine her for splenomegaly.

The second is a college student with failing appetite, ability to concentrate, energy, and grades. You think that he is depressed but want to rule out infectious mononucleosis and decide to examine him for splenomegaly.

The third is an otherwise healthy man with well-controlled hypertension and a normal cardiovascular examination result. As he lies on the examining table, stripped to his waist, you wonder whether you should take the time to examine him for splenomegaly.

WHY EXAMINE THE SPLEEN?

We examine the spleen to see whether it is palpable. Most palpable spleens are enlarged, and splenomegaly in an adult requires an explanation, for it may be a manifestation of disease. Despite many important causes of splenomegaly, including cancers, infections, and connective tissue diseases, many of these diagnoses are relatively uncommon such that isolated splenomegaly in an otherwise healthy adult is most often associated with nonspecific infections or no obvious cause.¹

ANATOMIC LANDMARKS AND SPLENIC SIZE

The normal spleen is a curved wedge that follows the course of the bony portion of the left 10th rib (Figure 46-1A). Its narrow posterior pole points back and to the right, toward the spine. Its outer surface is convex and lies just beneath the left side of the diaphragm, and its blunt anterior pole approaches the midaxillary line, pointing toward the left side of the colic flexure. Its inner convex surface bears a large impression from the posterior wall of the stomach, and its inferior edge bears impressions from the upper pole of the left kidney and, occasionally, the tail of the pancreas.

HOW LARGE IS THE NORMAL SPLEEN?

Autopsies after sudden traumatic death in individuals free of disorders likely to lead to splenomegaly have provided information on the usual weight of the spleen. In Philadelphia,

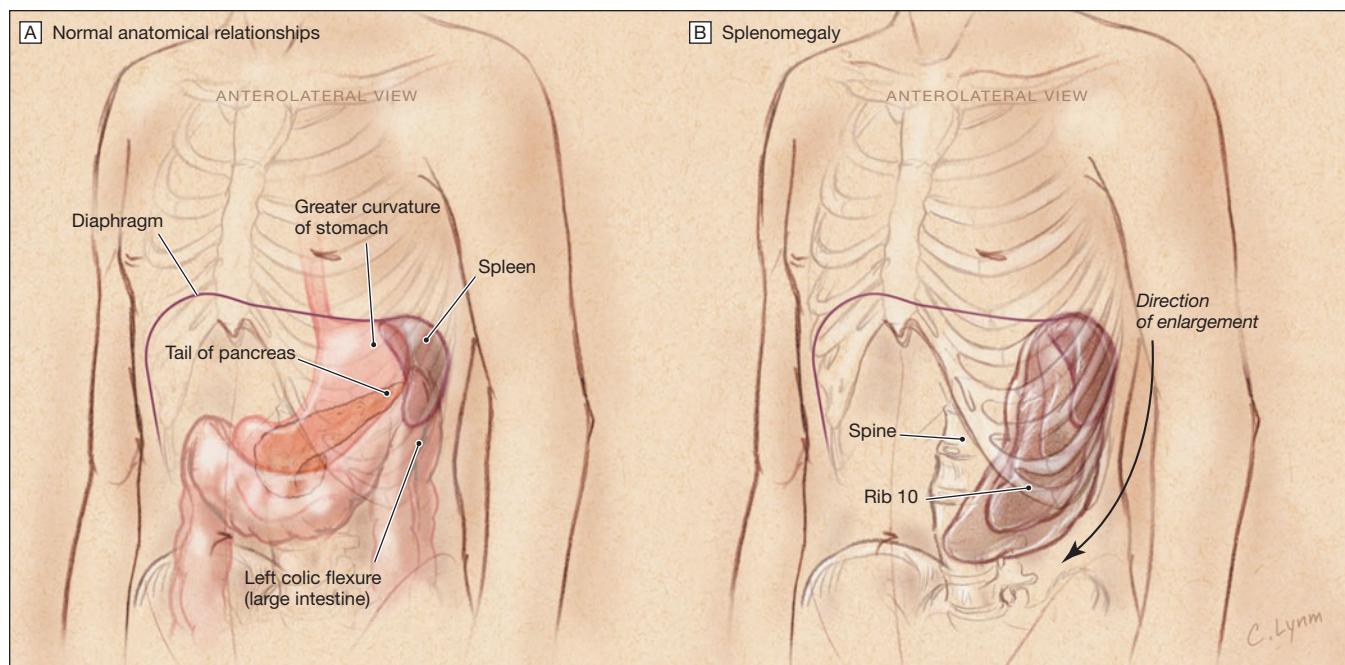


Figure 46-1 The Normal Sized Spleen Rests Hidden Under the Rib Cage

A, The normal spleen is a curved wedge that follows the course of the bony portion of the left 10th rib. Its narrow posterior pole points back and to the right, toward the spine. Its outer surface is convex and lies just beneath the left side of the diaphragm, and its blunt anterior pole approaches the midaxillary line, pointing toward the left side of the colic flexure. Its inner convex surface bears a large impression from the posterior wall of the stomach, and its inferior edge bears impressions from the upper pole of the left kidney and occasionally the tail of the pancreas. B, As the spleen enlarges, its anterior pole continues to follow the left 10th rib as the spleen descends below the rib cage and across the abdomen toward the right iliac fossa.

Pennsylvania, such spleens exhibited median weights from 90 g (among black women) to 170 g (among young white men), with intermediate values for black men (100 g), white women (115 g), and elderly white men (130 g). The pathologists who conducted these studies stated that the “best rule of thumb is to regard any spleen under 250 g as normal.”² This biologic variation in average spleen size underscores the need for a criterion standard definition of splenic enlargement that is acceptable to patients (ie, having one's spleen weighed is painful) and reproducible for clinicians.

One such standard is the radioisotopic scintiscan, presented (with the most commonly used normal values in parentheses) as maximum values for length (12 cm) and width (7 cm),³ surface area (80 cm²),⁴ or volume (250 cm³).⁵ Most recently, an ultrasonographic criterion standard has been suggested, with splenomegaly defined as a cephalo-caudad diameter of 13 cm or more.^{6,7}

THE CONSEQUENCES OF SPLENOMEGLY FOR THE CLINICAL EXAMINATION

Because the normal-sized spleen almost always lies entirely within the rib cage, it usually cannot be palpated. However, as it enlarges it displaces the stomach but cannot displace the spine, diaphragm, or kidney. Therefore, its anterior pole continues to follow the projection of the bony portion of the left 10th rib, descending below the rib cage and across the abdomen toward the right iliac fossa (Figure 46-1B).

HOW TO EXAMINE FOR SPLENOMEGLY

Inspection

Inspection of the left upper quadrant might reveal a bulging mass emerging from under the left costal margin and descending on inspiration. There are no published assessments of the accuracy of clinical inspection. Nonetheless, this sign would be expected to have low sensitivity because only massive spleens will distort the abdominal wall sufficiently to be seen. Moreover, because other large masses (a polycystic kidney or gastric or colon cancer) also can distort the abdominal wall and may descend on inspiration, this sign probably does not have perfect specificity either. In the absence of previous documentation or suspicion of massive splenomegaly, this is unlikely to be a useful sign.

Percussion

Percussion seeks to identify the loss of tympany as the enlarging spleen impinges on the adjacent air-filled lung, stomach, and colon.

Percussion is often claimed to be more sensitive than palpation for lesser degrees of splenomegaly, although evidence to support this claim (described herein) is scant.

Three percussion methods have been validated against ultrasonography or scintigraphy:

1. Percussion by Nixon Method (as Modified by Sullivan and Williams)

The patient is placed in the right lateral decubitus position. Percussion is initiated midway along the left costal margin and continued upward along a line perpendicular to the

costal margin (Figure 46-2). In a normal examination, a full stomach can result in initial percussion dullness, but as percussion continues along the perpendicular line tympany then becomes present because of the overlying lung. Splenomegaly is diagnosed when the dullness is present more than 8 cm above the costal margin.^{8,9}

2. Percussion by Castell Method

The patient is placed in the supine position. Percussion is carried out in the lowest intercostal space in the left anterior axillary line in both expiration and full inspiration (Figure 46-3). In a normal examination result, the percussion note remains resonant throughout this maneuver. Splenomegaly is diagnosed when the percussion note is dull or becomes dull on full inspiration.¹⁰

3. Percussion of Traube Space

The patient is supine, with the left arm slightly abducted for access to the entire Traube space (after its description by Ludwig Traube, who ascribed its disappearance to pleural effusion, not an enlarged spleen),¹¹ defined by the sixth rib superiorly, the midaxillary line laterally, and the left costal margin inferiorly (Figure 46-3). With the patient breathing normally, this triangle is percussed across 1 or more levels from its medial to lateral margins. Normal percussion yields a resonant or tympanitic note. Splenomegaly is diagnosed when the percussion note is dull.¹²

Palpation

Although many methods for palpation of the spleen have been reported in clinical texts and journals, only 3 have had their precision or accuracy documented in the clinical literature and will be described herein. Relaxation of the abdominal wall is a prerequisite for successful palpation and can be assisted by both the examiner (friendly, gentle, and warm hands) and the patient (flexed, supported knees).

Two-Handed Palpation With Patient in Right Lateral Decubitus

With the patient in the right lateral decubitus position, the examiner's left hand is slipped from front to back around the left lower thorax, gently lifting the left lowermost rib cage anteriorly and medially. The tips of the fingers of the examiner's right hand are pressed gently just beneath the left costal margin, and the patient is asked to take a long, deep breath as the palpation of a descending spleen is sought. If none is felt, the procedure is repeated, lowering the right hand 2 cm toward the umbilicus each cycle, until the examiner is confident that a massive spleen has not been missed. (Some authorities suggest starting palpation over the lower abdomen and moving up toward the costal margin.) The same procedure can be carried out with the patient supine.

One-Handed Palpation With Patient Supine

This method is identical to the former one, except that no counterpressure is applied by the left hand to the rib cage. With the patient supine, the tips of the fingers of the examiner's right hand are pressed gently just beneath the left costal margin, and the patient is asked to take a long, deep breath as the palpation of a descending spleen is sought. If none is felt,

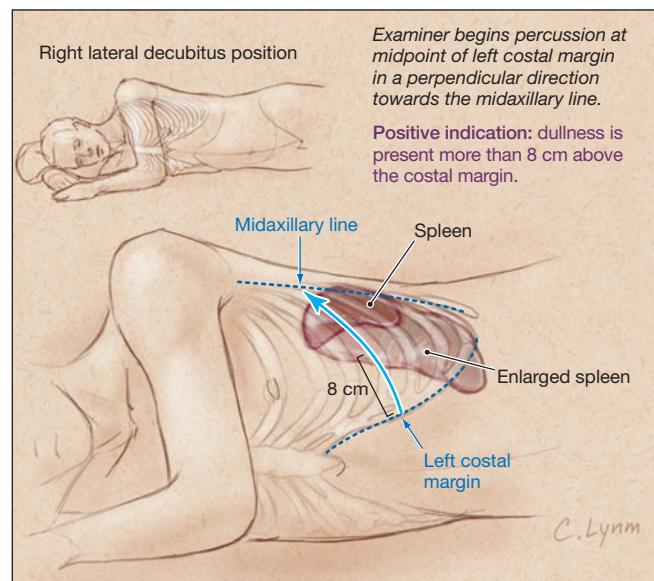


Figure 46-2 Nixon Method to Detect Splenomegaly

Nixon method of percussion requires that the patient be placed in the right lateral decubitus position. Percussion is started at the midpoint of the left costal margin and proceeds perpendicularly. Splenomegaly is diagnosed when the dullness is present more than 8 cm above the costal margin.

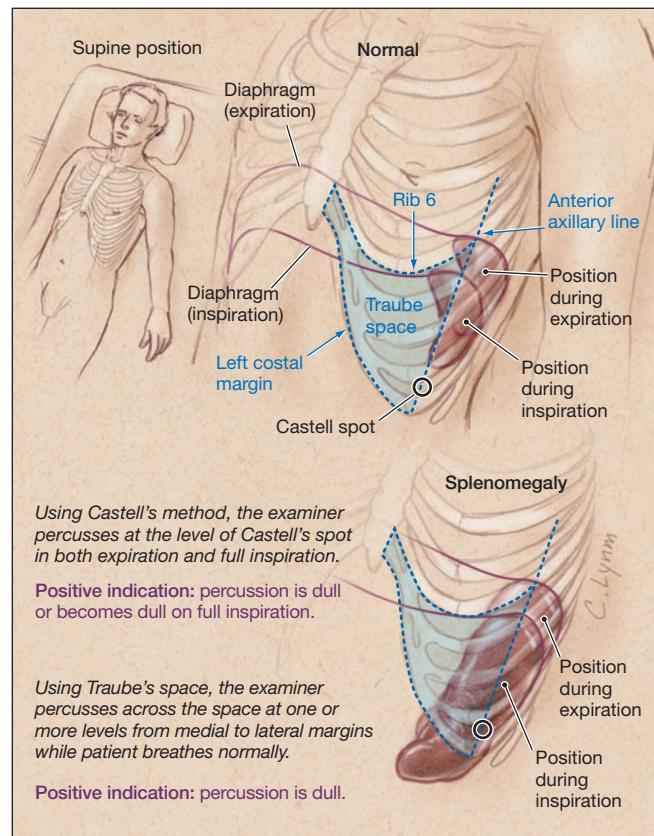


Figure 46-3 Percussion in Traube Space and at Castell Spot to Detect Splenomegaly

Traube space is defined by the sixth rib superiorly, the left anterior axillary line laterally, and the costal margin inferiorly. Castell spot is located at the junction of the lowest intercostal space and the left anterior axillary line.

the procedure is repeated, lowering the right hand 2 cm toward the umbilicus each cycle, until the examiner is confident that a massive spleen has not been missed. Some examiners like to apply counterpressure to the patient's flank with the left hand while palpating with the right.

Hooking Maneuver of Middleton With Patient Supine

The patient is asked to lie flat with his or her left fist under the left costovertebral angle. The examiner is positioned to the patient's left, facing the patient's feet. The fingers of both the examiner's hands are curled under the left costal margin, and the patient is asked to take a long, deep breath as the palpation of a descending spleen is sought.¹³

Additional Features of the Palpable Spleen

Given its origin within the rib cage, most texts state that it is never possible to palpate (get above) the upper border of the spleen, helping distinguish it from other abdominal masses that may present an upper border. If a spleen is greatly enlarged, it may be possible to feel a hilar notch along its medial border.

PRECISION OF THE SIGNS FOR SPLENOMEGALY

When groups of inpatients with and without splenomegaly had their Traube spaces percussed by 3 internists, the interexaminer agreement (κ values) ranged from 0.19 to 0.41, which is modest at best.¹² However, recent food intake reduced the accuracy of Traube space percussion in this study and probably decreased the test precision when different physicians examined the same patient at various times after meals. Among the same patients, a second study¹⁴ showed that the interexaminer agreement for palpation ranged from 0.56 to 0.70, demonstrating that reproducibility between examiners of palpation was better than percussion.

When tested among 50 patients with alcoholism, agreement among different examiners (using 2-handed palpation with the patient in the right lateral decubitus and 1-handed palpation with the patient supine) demonstrated an interclass correlation coefficient of 0.75 and was as good as that for ascites (and marginally better than that for jaundice,

Dupuytren contracture, vascular spiders, gynecomastia, palmar erythema, asterixis, or clubbing).¹⁵ Senior gastroenterologists exhibited marginally better agreement than more junior physicians (intraclass correlation coefficients of 0.81 and 0.73, respectively). When different examiners were asked to report the extent to which the spleen tip extended below a specific bony landmark (eg, the xiphisternal-sternal junction), their estimates varied on average by 6 cm.¹⁶

ACCURACY OF THE SIGNS FOR SPLENOMEGALY

Table 46-1 summarizes studies on the accuracy of percussion. Using ultrasonographic results as the criterion standard, percussion of Traube space had a sensitivity of 62% (95% confidence interval [CI], 51%-72%) and a specificity of 72% (95% CI, 65%-80%).¹² Percussion sensitivity was reduced by the presence of obesity (more false-negative results), and its specificity was decreased by recent food intake (more false-positive results). Accordingly, among leaner patients who had not eaten in the previous 2 hours, percussion sensitivity was 78% (95% CI, 62%-90%), and its specificity was 82% (95% CI, 70%-90%).

A second study⁹ examined the sensitivity and specificity, individually and in combination, of the Nixon and Castell methods of percussion (as well as 2-handed palpation in the supine and right lateral decubitus positions). In comparing the Nixon to the Castell method of percussion, the Castell method exhibited a higher sensitivity (82% vs 59%) but lower specificity (83% vs 94%) (Table 46-1).

Table 46-2 summarizes 7 studies of the accuracy of palpation. The first 2 studies^{17,18} assessed the accuracy of the routine examination for splenomegaly by abstracting the clinical examinations (performed by a large number and range of clinicians) from routine clinical charts. Both studies found low sensitivity (20%-28%) but high specificity (98%-100%). Most enlarged spleens were missed (a high rate of false-negative results, leading to low sensitivity), but few examiners reported palpating spleens that were not there (a low rate of false-positive results, leading to high specificity). When the results of these 2 studies were combined, the routine examination for splenomegaly had a sensitivity of 27% (95% CI, 19%-36%) and a specificity of 98% (95% CI, 96%-100%).

In the other 5 palpation studies^{4,5,9,14,19} (Table 46-2), the examination for splenomegaly was performed as part of the study. Because the examiners knew that they were under scrutiny, it is not surprising that both their true-positive reports and false-positive reports of splenomegaly increased; that is, the overall sensitivity of palpation was higher and the specificity lower than in the 2 previously described studies that assessed the routine examination as recorded in clinical notes.

One study⁹ compared percussion methods and palpation and demonstrated that the Castell method of percussion may be somewhat more sensitive than palpation (82% vs 71%) (Tables 46-1 and 46-2). Finally, if splenomegaly was declared when any of the 4 signs (2 for percussion and 2 for palpation) were positive, true-positive and false-positive declarations of splenomegaly increased because the increase in sensitivity to

Table 46-1 Studies of the Accuracy of Percussion

No. of Patients	Criterion Standard	Maneuver	Sensitivity, % (No.)	Specificity, % (No.)
118 ^a	Ultrasonography	Traube space percussion		
		All patients ¹²	62 (58/94)	72 (109/151)
		Nonobese patients who have not eaten recently ¹²	78 (29/37)	82 (54/66)
65	Scintigraphy	Nixon method ⁹	59 (10/17)	94 (45/48)
		Castell method ⁹	82 (14/17)	83 (40/48)

^aEach patient was examined by 1 to 3 examiners, for a total of 245 examinations.

Table 46-2 Studies of the Accuracy of Palpation

No. of Patients	Criterion Standard	Maneuver	Sensitivity, % (No.)	Specificity, % (No.)
Based on Routine Examinations Recorded in Clinical Charts				
47	Autopsy ¹⁷	Physical examination	20 (3/15)	100 (32/32)
217	Scintigraphy ¹⁸	Clinical impressions	28 (26/92)	98 (122/125)
		Overall	27 (29/107)	98 (154/157)
Based on Specific Examinations Done as Part of the Study				
99	Scintigraphy ⁴	Palpation	57 (31/54)	100 (45/45)
32	Operation ¹⁹	Palpability	59 (16/27)	100 (5/5)
100	Scintigraphy ⁵	Supine 2-handed palpation	56 (47/84)	69 (11/16)
65	Scintigraphy ⁹	Supine and right lateral decubitus palpation	71 (12/17)	90 (43/48)
118 ^a	Ultrasonography ¹⁴	Supine palpation or Middleton maneuver	56 (53/94)	93 (140/151)
		Overall	58 (159/276)	92 (244/265)

^aEach patient was examined by 1 to 3 examiners, for a total of 245 examinations.

88% (fewer large spleens missed) was accompanied by a decrease in specificity to 83% (more normal-sized spleens mistakenly called large).

The final study¹⁴ evaluated the accuracy of bedside diagnostic maneuvers, using receiver operating characteristic curve analysis. This analytic technique evaluates the discriminating ability of different tests by comparing the true-positive rate (sensitivity) and false-positive rate (1 – specificity) of each test using different definitions of a positive test result (test thresholds). The discriminating ability refers to the probability of correctly selecting the patient with splenomegaly between 2 patients: one with an enlarged spleen and one with a normal spleen. A test with a discriminating ability of zero performs no better than chance alone, whereas a perfect test has a discriminating ability of 100%.

In this study, supine palpation, right lateral decubitus palpation, and Middleton maneuver all demonstrated similar discriminating abilities (73%–79%). The discriminating ability of palpation and percussion was similar, although the test specificity of palpation appeared to be generally superior to percussion.

The most important finding of this study was that palpation was a better discriminator among patients in whom percussion result was positive. (As might be expected, these patients have the largest spleens.) When percussion dullness was present, palpation discriminated correctly 87% of the time. However, if percussion was not dull, palpation was a poor discriminator (55%) or only slightly better than chance. This confirms that percussion and palpation should be used together because percussion dullness identifies a subset of patients in whom palpation is a useful test. If percussion dullness is absent, there is no need to palpate, because palpation is a poor test among such patients.

Finally, this study also demonstrated that, given a clinical suspicion (the prior probability or disease prevalence) of splenomegaly before examining the patient of 10% to 90%, it is difficult to substantially decrease the likelihood of an enlarged spleen because the false-negative rate of bedside diagnosis was

28%, even if percussion and palpation results were negative. On the other hand, when a positive bedside examination result was defined as both percussion and palpation results being positive, the high test specificity of 97% significantly increased the likelihood of splenic enlargement to 60% or more.

IS SPLENOMEGLY RESULT EVER NORMAL?

About 3% of otherwise healthy students entering a US college were found to have unexplained palpable spleens¹ and, on incomplete follow-up, appeared to fare none the worse²⁰; similarly, 12% of otherwise healthy postpartum women at a Canadian hospital had palpable spleens.²¹

THE BOTTOM LINE

Guidelines for examining for splenic enlargement are summarized in Table 46-3.

1. Splenomegaly is uncommon but occurs in a wide variety of conditions. Given the low sensitivity of the clinical examination, it can be argued that the routine examination for splenomegaly cannot definitively rule in or rule out splenomegaly in normal, asymptomatic patients when the prevalence is less than 10% and additional imaging tests will be required. Rather, the examination for splenomegaly is most useful to rule in the diagnosis of splenomegaly among patients in whom there is a clinical suspicion of at least 10%.
2. The bedside examination of the spleen should start with percussion. If percussion is not dull, there is no need to palpate because the results of palpation will not effectively rule in or rule out splenic enlargement. If the possibility of missing splenic enlargement remains an important clinical concern, then ultrasonography or scintigraphy is indicated. In the presence of percussion dullness, palpation should follow. If both test results are positive, the diagnosis of splenomegaly is established (providing that the clinical suspicion

Table 46-3 Guidelines for Examining for Splenic Enlargement

Recommendations and Rationale	
Clinical Suspicion (Prior Probability) of Splenic Enlargement	
Less than 10%	Percussion or palpation for splenomegaly of limited usefulness Maneuvers are not sufficiently sensitive to rule out splenomegaly Given the low pretest probability of splenomegaly, test specificity of clinical examinations is not sufficiently high to rule in splenic enlargement, even if both test results are positive
10% Or more	Percussion and palpation can be used to rule in splenomegaly if both results are positive Percuss first, and if result is positive, then palpate If percussion result is negative but your clinical suspicion remains high, order ultrasonography because palpation in the presence of abdominal tympany is not specific enough to rule in splenomegaly If percussion result is positive but palpation result is negative, then ultrasonography is also needed to confidently evaluate spleen size
To confidently rule out splenomegaly, a radiologic procedure is necessary because of the limited sensitivity of bedside examination	

of splenomegaly was at least 10% before examination). If palpation result is negative, diagnostic imaging will be required to confidently rule in or rule out splenomegaly.

CLINICAL SCENARIO—RESOLUTION

Returning to the 3 patients originally described at the beginning of this article, you may be able to confidently rule in splenic enlargement in the pale elderly women complaining of fatigue if your preeexamination clinical suspicion of splenomegaly is at least 10% and if both percussion and palpation results are positive. Abdominal examination is not sufficiently sensitive to rule out splenic enlargement in the college student with symptoms of depression. Finally, you may choose to examine for splenic enlargement in the asymptomatic man with hypertension, but a negative examination result may be a false negative, and a positive examination result will require radiologic confirmation to rule in splenomegaly.

Author Affiliations at the Time of the Original Publication

Divisions of General Internal Medicine, Gastroenterology, and Clinical Epidemiology, Montreal General Hospital, and the Departments of Medicine and Epidemiology and Biostatistics, McGill University, Montreal, Quebec, Canada (Drs Grover and Barkun); and the Division of General Internal Medicine, Department of Medicine, and the Department of Clinical Epidemiology and Biostatistics of McMaster University and Henderson General Hospital, Hamilton, Ontario, Canada (Dr Sackett).

sity and Henderson General Hospital, Hamilton, Ontario, Canada (Dr Sackett).

Acknowledgments

The authors acknowledge, with thanks, helpful comments received from Roger Williams, MD, and Andreas Laupacis, MD.

Dr Grover is a research scholar and Dr Barkun is a clinical scholar supported by the Fonds de la recherche en santé du Québec.

REFERENCES

- Ebaugh FG, McIntyre OR. Palpable spleens: ten-year follow-up. *Ann Intern Med.* 1979;90(1):130-131.
- Myers J, Segal RJ. Weight of the spleen, I: range of normal in a nonhospital population. *Arch Pathol.* 1974;98(1):33-35.
- Larson SM, Tuell SH, Moores KD, Nelp WB. Dimensions of the normal adult spleen scan and prediction of spleen weight. *J Nucl Med.* 1971;12(3):123-126.
- Westin J, Lanner SO, Larsson A, Weinfeld A. Spleen size in polycythemia: a clinical and scintigraphic study. *Acta Med Scand.* 1972;191(3):263-271.
- Zhang B, Lewis SM. Use of radionuclide scanning to estimate size of spleen in vivo. *J Clin Pathol.* 1987;40(5):508-511.
- Niederau C, Sonnenberg A, Mueller JE, Erckenbrecht JF, Scholten T, Fritsch WP. Sonographic measurements of the normal liver, spleen, pancreas, and portal vein. *Radiology.* 1983;149(2):537-540.
- Koga T, Morikawa Y. Ultrasonographic determination of the splenic size and its clinical usefulness in various liver diseases. *Radiology.* 1975;115(1):157-161.
- Nixon RK Jr. The detection of splenomegaly by percussion. *N Engl J Med.* 1954;250(4):166-167.
- Sullivan S, Williams R. Reliability of clinical techniques for detecting splenic enlargement. *BMJ.* 1976;2(6043):1043-1044.
- Castell DO. The spleen percussion sign: a useful diagnostic technique. *Ann Intern Med.* 1967;67(6):1265-1267.
- Talbott JE. *A Biographical History of Medicine: Excerpts and Essays on the Men and Their Work.* New York, NY: Grune & Stratton; 1970:594-595.
- Barkun AN, Camus M, Meagher T, et al. Splenic enlargement and Traube's space: how useful is percussion? *Am J Med.* 1989;87(5):562-566.
- Shaw MT, Dvorak V. Palpation of slightly enlarged spleens. *Lancet.* 1973;1(7798):317.
- Barkun AN, Camus M, Green L, et al. The bedside assessment of splenic enlargement. *Am J Med.* 1991;91(5):512-518.
- Espinosa P, Ducot B, Pelletier G, et al. Interobserver agreement in the physical diagnosis of alcoholic liver disease. *Dig Dis Sci.* 1987;32(3):244-247.
- Blendis LM, McNeilly WJ, Sheppard L, Williams R, Laws JW. Observer variation in the clinical and radiological assessment of hepatosplenomegaly. *BMJ.* 1970;1(5698):727-730.
- Riemenschneider PA, Whalen JP. The relative accuracy of estimation of enlargement of the liver and spleen by radiologic and clinical methods. *Am J Roentgenol.* 1965;94:462-468.
- Halpern S, Coel M, Ashburn W, et al. Correlation of liver and spleen size: determinations by nuclear medicine studies and physical examination. *Arch Intern Med.* 1974;134(1):123-124.
- Ingeberg S, Stockel M, Sorensen PJ. Prediction of spleen size by routine radioisotope scintigraphy. *Acta Haematol.* 1983;69(4):243-248.
- McIntyre RR, Ebaugh FG Jr. Palpable spleens in college freshmen. *Ann Intern Med.* 1967;66(2):301-306.
- Berris B. The incidence of palpable liver and spleen in the postpartum period. *CMAJ.* 1966;95(25):1318-1319.

Prepared by Alan N. Barkun, MD, and Steven A. Grover, MD

Reviewed by Andrew Muir, MD

CLINICAL SCENARIO

A 34-year-old man has complained of fatigue and abdominal pain. He presents to the emergency department with vague abdominal pain and fever. The medical history is also that of intermittent sweats and some weight loss. Your examination reveals diffuse adenopathy. Traube space is dull to percussion. You decide to try to palpate the spleen edge but, despite spending a few minutes examining the patient while he is supine and then while he is on his side, you decide that you cannot feel the spleen. According to your findings, how confident should you be that the spleen is not enlarged?

Original Review

Grover SA, Barkun AN, Sackett DL. The rational clinical examination: does this patient have splenomegaly? *JAMA*. 1993;270(18):2218-2221.

UPDATED LITERATURE SEARCH

Our literature search used the parent search strategy for The Rational Clinical Examination series, combined with the subject “exp splenomegaly,” published in English from 1991 to 2004, and articles that referred to the original review. The results yielded 136 articles, for which we reviewed the titles and abstracts. We found 5 articles suitable for review, although 1 was a duplicate publication. Of the remaining 4 studies, 3 were selected because they had prospective evaluation of patients for splenomegaly, with both sensitivity and specificity data collected independent of an ultrasonograph used as the reference standard test. One of the studies had information on the interobserver variability of examination techniques. We also identified 1 study of the sensitivity of the examination for splenomegaly in athletes.

SUMMARY OF NEW FINDINGS

- Palpation might have greater accuracy than percussion, especially in lean patients. However, assessment of palpation performance may be biased because in a number of studies, palpation followed percussion maneuvers.

- Examiners should become proficient in 1 palpation method and 1 percussion method because the combination of both results may be better than either alone.

Details of the Update

A study from Brazil suggested that combining the results of 2 examiners’ palpation findings (presence of a palpable spleen, presence of a spleen felt more than 4 cm below the costal margin)¹ gave good results. When both physicians palpated the spleen, the likelihood ratio (LR) of splenomegaly increased (LR, 7.6; 95% confidence interval [CI], 4.5-12); when neither physician palpated the spleen, the likelihood of splenomegaly decreased (LR, 0.31; 95% CI, 0.14-0.56). The inference is that having a colleague confirm your findings for splenomegaly might be useful.

Two studies from India suggested that palpation maneuvers may have better accuracy for diagnosing splenomegaly than percussion techniques.^{2,3} However, the order of the maneuvers was not stated, and by western standards the patients were of small stature and size (as measured by body mass index). Furthermore, in one of the studies, false-negative results for Traube space percussion were significantly higher in smaller patients.³ Therefore, although palpation may perform better than percussion in lean patients, we do not know whether the same test characteristics apply to patients with larger body mass.⁴ The clinical utility appears enhanced when the results of both percussion and palpation are considered but should be confirmed in other studies in which the order of the examination is specified.³

A study performed in a convenience sample of patients with confirmed or suspected human immunodeficiency virus (HIV) suggested that 3 palpation maneuvers and 3 percussion maneuvers were relatively insensitive but had better specificity,⁵ supporting the findings of the original Rational Clinical Examination article. Although the study sample was small (27 patients), a unique feature of the evaluation was that there were 8 observers, allowing a comparison between observers and an assessment to see whether the various maneuvers performed similarly. However, they also noted significant interobserver variability that did not depend on the years of medical practice. The poor reliability was evident in the broad range of individual assessors’ sensitivity and specificity values. The sensitivity of each of the tests seemed

to improve with the length of the trial, but the overall accuracy of all the findings was low. Because the evaluation of such a large number of individual findings may have lacked independence, and because the total number of patients was so small, we did not combine these results with other studies.

The presence of splenomegaly in athletes (often caused by mononucleosis) creates a diagnostic dilemma for clinicians who must decide when the splenomegaly has resolved so that the athlete can return to sports participation. A study of 29 athletes with splenomegaly (length, 12.5–15.5 cm) documented by ultrasonography (normal length < 12 cm), showed that the clinician could detect the spleen in only 17%.⁶ Many athletes have well-developed abdominal musculature, which makes palpation for splenomegaly even more difficult.

IMPROVEMENTS IN THE DATA PRESENTED IN THE ORIGINAL PUBLICATION

A reappraisal of the original publication showed that CIs around the signs would help understanding of their potential importance.⁷ We used the original data, in addition to data from newer articles, to create random-effects summary estimates for the LRs. In addition, we used the diagnostic odds ratios to assess whether the overall accuracy for some maneuvers might be better than others. We used only data from studies that used ultrasonography as the reference standard test.

CHANGES IN THE REFERENCE STANDARD

Although radiologic studies have suggested the possible use of competing technologies, such as nuclear scan and specialized computed tomography (CT) examinations, the most widely recognized and available gold standard remains ultrasonography. All articles assessing the utility of clinical examination maneuvers in the detection of splenomegaly published in the past 13 years used ultrasonography as the reference standard for the diagnosis of splenomegaly (a length of 12 or 13 cm).

RESULTS OF LITERATURE REVIEW

Percussion using the Nixon method (Figure 46-2) or Traube's space (Figure 46-3) works best for detecting splenomegaly

Table 46-4 Likelihood Ratios of Percussion and Palpation Maneuvers for Splenomegaly

Maneuver (No. of Combined Studies)	LR+ (95% CI)	LR- (95% CI)	DOR (95% CI)
Percussion Maneuvers			
Nixon sign (1)	3.6 (1.8–7.3)	0.41 (0.26–0.64)	8.9 (3.1–25)
Percussion of Traube space (3)	2.3 (1.8–2.9)	0.48 (0.39–0.60)	4.8 (3.2–7.3)
Castell sign (1)	1.2 (0.98–1.6)	0.45 (0.19–1.1)	2.8 (0.92–8.3)
Palpation Maneuvers			
Supine, 1-handed palpation (4)	8.2 (5.8–12)	0.41 (0.30–0.57)	22 (13–38)
Middleton hooking maneuver (1)	6.5 (3.1–15)	0.16 (0.08–0.32)	40 (11–138)

Abbreviations: CI, confidence interval; DOR, diagnostic odds ratio; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

(Table 46-4). Supine one-handed palpation has been the most widely studied palpation maneuver, which increases the confidence in the results (Table 46-4).

EVIDENCE FROM GUIDELINES

No federal guidelines discuss the assessment of splenomegaly by using physical examination.

CLINICAL SCENARIO—RESOLUTION

This patient may have a viral or myeloproliferative syndrome, so you have a good reason to assess for splenomegaly. The physical examination results seem contradictory. You have percussed dullness (which increases the likelihood of splenomegaly), but you cannot palpate the splenic tip (which decreases the likelihood of splenomegaly). The percussion findings have a lower accuracy than the palpation signs (as suggested by the diagnostic odds ratios). You decide you need to know whether the patient has splenomegaly, so you must proceed to additional testing with ultrasonography or a CT scan.

SPLENOMEGLY—MAKE THE DIAGNOSIS

During the general physical examination, patients should *not* be evaluated for splenomegaly.

PRIOR PROBABILITY

The prevalence of palpable splenomegaly in an otherwise healthy student population is low, approximating 3%⁸; 12% of normal postpartum women had palpable spleens.⁹ The prevalence of splenomegaly increases significantly among other selected populations, such as HIV patients (up to 66%¹⁰), or in areas in which schistosomiasis is prevalent.¹¹

POPULATION FOR WHOM THE PHYSICAL EXAMINATION OF SPLENOMEGLY SHOULD BE SOUGHT

- Suspected or proven viral illness, lymphoproliferative disorder, or malignancy
- Cirrhosis
- Suspected portal hypertension
- Suspected or proven malaria
- Connective tissue disorders associated with splenomegaly

DETECTING SPLENOMEGLY

In cases in which splenomegaly is questioned, the clinical examination is more specific than sensitive and is best used when ruling in the diagnosis among patients for whom the suspicion is at least 10%. Moreover, the examination should

start with Traube space percussion, followed, if dull, by supine 1-handed palpation (Table 46-5). These maneuvers have received more extensive evaluation than other maneuvers, allowing us greater confidence in the findings. Middleton maneuver, in which the physician stands to the left of the patient and hooks the examining hand under the ribs, may work as well.

Palpation may be superior to percussion, especially in lean patients. When it remains important not to miss splenomegaly, imaging will be necessary because the clinical examination does not provide sufficient clinical certainty.

Table 46-5 Summary Likelihood Ratios for Palpation to Detect Splenomegaly and Percussion of Traube Space

Test (No.)	LR+ (95% CI)	LR- (95% CI)	DOR (95% CI)
Supine 1-handed palpation (4 studies)	8.2 (5.8-12)	0.41 (0.30-0.57)	22 (13-38)
Percussion of Traube space (3 studies)	2.3 (1.8-2.9)	0.48 (0.39-0.60)	4.8 (3.2-7.3)

Abbreviations: CI, confidence interval; DOR, diagnostic odds ratio; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

REFERENCE STANDARD TESTS

Ultrasonography, CT, nuclear liver-spleen imaging.

REFERENCES FOR THE UPDATE

1. Gerspacher-Lara R, Pinto-Silva RA, Serufo JC, Rayes AA, Drummond SC, Lambertucci JR. Splenic palpation for the evaluation of morbidity due to schistosomiasis mansoni. *Mem Inst Oswaldo Cruz*. 1998;93(suppl 1):245-248.^a
2. Chongtham DS, Singh MM, Kalantri SP, Pathak S. Accuracy of palpation and percussion manoeuvres in the diagnosis of splenomegaly. *Indian J Med Sci*. 1997;51(11):409-416.^a
3. Dubey S, Swaroop A, Jain R, Verma K, Garg P, Agarwal S. Percussion of Traube's space—a useful index of splenic enlargement. *J Assoc Physicians India*. 2000;48(3):326-328.^a
4. Barkun AN, Camus M, Green L, et al. The bedside assessment of splenic enlargement. *Am J Med*. 1991;91(5):512-518.
5. Tamayo SG, Rickman LS, Mathews WC, et al. Examiner dependence on physical diagnostic tests for the detection of splenomegaly: a prospective study with multiple observers. *J Gen Intern Med*. 1993;8(2):69-75.^a
6. Dommerby H, Stangerup SE, Stangerup M, et al. Hepatosplenomegaly in infectious mononucleosis, assessed by ultrasonic scanning. *J Laryngol Otol*. 1986;100(5):573-579.
7. Grover SA, Barkun AN, Sackett DL. The rational clinical examination: does this patient have splenomegaly? *JAMA*. 1993;270(18):2218-2221.
8. Larson SM, Tuell SH, Moores KD, Nelp WB. Dimensions of the normal adult spleen scan and prediction of spleen weight. *J Nucl Med*. 1971;12(3):123-126.
9. Berris B. The incidence of palpable liver and spleen in the postpartum period. *CMAJ*. 1966;95(25):1318-1319.
10. Furrer H. Prevalence and clinical significance of splenomegaly in asymptomatic human immunodeficiency virus type 1-infected adults: Swiss HIV Cohort Study. *Clin Infect Dis*. 2000;30(6):943-945.
11. Lambertucci JR, Cota GF, Pinto-Silva RA, et al. Hepatosplenic schistosomiasis in field-based studies: a combined clinical and sonographic definition. *Mem Inst Oswaldo Cruz*. 2001;96(suppl):147-150.

^aFor the Evidence to Support the Update for this topic, see <http://www.JAMAevidence.com>.

This page intentionally left blank

EVIDENCE TO SUPPORT THE UPDATE:

46

Splenomegaly

TITLE Accuracy of Palpation and Percussion Maneuvers in the Diagnosis of Splenomegaly.

AUTHORS Chongtham DS, Singh MM, Kalantri SP, Pathak S.

CITATION *Indian J Med Sci.* 1997;51(11):409-416.

QUESTION What are the sensitivity and specificity of palpation and percussion maneuvers in diagnosing splenomegaly?

DESIGN Prospective, independent comparison of non-consecutive cases.

SETTING Medical ward at Katsurba Hospital, India.

PATIENTS Eighty hospitalized patients (37 female patients) in a general medical ward. Exclusions were patients with left-sided pleural effusion, history of ascites, or splenomegaly. Mean age was 31.5 years, and weight was 45 ± 8 kg.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

The test performance characteristics of Traube space percussion (Barkun et al¹), and the percussion maneuvers of Castell² and Nixon³ were evaluated at various percussion note thresholds (1 = definitely tympanic, 2 = probably tympanic, 3 = uncertain, 4 = probably dull, 5 = definitely dull¹). Supine palpation and Middleton palpation maneuver⁴ were also assessed on a 5-point scale (1 = spleen definitely not palpable, 2 = spleen probably not palpable, 3 = uncertain, 4 = spleen probably palpable, and 5 = spleen definitely palpable, as previously suggested⁵). The assessments were carried out by a physician blinded to the patient's clinical history and laboratory results. The examination was carried out before or at least 2 hours after the patient had eaten.

Ultrasonography was performed by an independent operator within 24 hours of the clinical examination.

MAIN OUTCOME MEASURES

The sensitivity and specificity of the various maneuvers were described. A spleen was considered enlarged if greater than 13 cm on ultrasonography.

MAIN RESULTS

The prevalence of splenomegaly was 52% (42/80). Mean splenic size was 15 cm among those with splenomegaly and 9.9 cm among those without enlargement. The likelihood ratios for the maneuvers are shown in Table 46-6.

Receiver operating curve (ROC) analyses showed a progressive decline in sensitivity from 98% to 50% as the palpation threshold progressed from 1 to 4 (increasing certainty of feeling a spleen), whereas specificity increased from 58% to 95%.

Nixon percussion maneuver was correlated with splenic size. The ROC area under the curve for varying thresholds on Traube space percussion was 0.74.

CONCLUSIONS

LEVEL OF EVIDENCE Level 3.

STRENGTHS Independent assessment of well-defined physical examination maneuvers.

Table 46-6 Likelihood Ratios of Palpation and Percussion Maneuvers for Splenomegaly

Test	Sensitivity, %	Specificity, %	LR+ (95% CI)	LR- (95% CI)	DOR (95% CI) ^a
Palpation Maneuvers					
Supine 1-handed palpation	79	92	10 (3.7-29)	0.23 (0.13-0.40)	43 (11-163)
Middleton hooking palpation maneuver	86	87	6.5 (3.1-15)	0.16 (0.08-0.32)	40 (11-138)
Percussion Maneuvers					
Nixon percussion	67	82	3.6 (1.8-7.3)	0.41 (0.26-0.64)	8.9 (3.1-25)
Traube space percussion	76	63	2.1 (1.4-3.3)	0.38 (0.20-0.66)	5.5 (2.1-14)
Castell percussion	86	32	1.2 (0.98-1.6)	0.45 (0.19-1.1)	2.8 (0.92-8.30)

Abbreviations: CI, confidence interval; DOR, diagnostic odds ratio; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

^aDOR calculated from data provided in the article.

LIMITATIONS The order in which the examinations were performed is not described. Furthermore, the generalizability may be questioned, considering the patient population characteristics (the mean Quetelet index of the studied patients was low by western standards, at $17.8 \pm 2.6 \text{ kg/m}^2$). The overall prevalence of splenomegaly suggests that this population may differ considerably from others or that there may have been some selection bias.

The palpation maneuvers appeared to perform appreciably better than percussion methods, as evidenced by the high diagnostic odds ratios. We do not know whether “leanness” as evidenced by a low body mass index creates a bias that favors palpation or percussion.

REFERENCES FOR THE EVIDENCE

- Barkun AN, Camus M, Meagher TW, et al. How useful is Traube's space percussion in assessing splenic enlargement? *Am J Med*. 1989;87(5):562-566.
- Castell DO. The spleen percussion sign: a useful diagnostic technique. *Ann Intern Med*. 1967;67(6):1265-1367.
- Nixon RK Jr. The detection of splenomegaly by percussion. *N Engl J Med*. 1954;250(4):166-167.
- Shaw MT, Dvorak V. Palpation of slightly enlarged spleens. *Lancet*. 1973;1(7798):317.
- Barkun AN, Camus M, Green L, et al. The bedside assessment of splenic enlargement. *Am J Med*. 1991;91(5):512-518.

Reviewed by Alan N. Barkun, MD

TITLE Splenic Palpation for the Evaluation of Morbidity Due to Schistosomiasis Mansoni.

AUTHORS Gerspacher-Lara R, Pinto-Silva RA, Serufo JC, Rayes AAM, Drummond SC, Lambertucci JR.

CITATION *Mem Inst Oswaldo Cruz (Rio de Janeiro)*. 1998;93(suppl I):245-248.

QUESTION What are the reliability and validity of 2 methods of palpation in detecting ultrasonographically identified splenomegaly?

DESIGN Prospective assessment of 2 near-complete communities with an independent assessment by ultrasonography.

SETTING Two Brazilian rural communities.

PATIENTS The study population was recruited from 551 individuals (92% of the local population) from Queixadinho, in the district of Caraí, located in the northeast of the State of Minas Gerais, Brazil, an area known to be highly endemic for schistosomiasis. An additional 517 individuals (89% of the total population) were recruited from Capão, a rural community in the district of Presidente Juscelino in the center of the state, where, for unknown reasons, transmission of schistosomiasis probably does not occur and in which other tropical diseases that can cause splenomegaly have never been identified.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

Abdominal palpation was performed with patients in the decubitus position during deep inspiration, by 2 independent physicians in a blinded fashion. The greatest distance between the splenic border and the costal margin was also independently measured by the examiners.

MAIN OUTCOME MEASURES

The 2 examination maneuvers were considered positive for splenomegaly when

- the spleen was palpable by both examiners; and
- the distance between the splenic border and the costal margin was greater than 4 cm, as measured by both examiners.

Splenomegaly was defined as a splenic length greater than 120 mm by ultrasonography. Only patients aged 18 years or older were included in the categorization of splenic enlargement because of the lack of widely accepted quantitative criteria in children.

MAIN RESULTS

The prevalence of splenomegaly in this patient population was 7%. A spleen was palpated by both physicians in 37 cases (discordance between examiners occurred in 5 cases). Mean splenic lengths in patients with and without palpable spleen were 10.4 cm and 7.1 cm, respectively ($P < .001$). Table 46-7 shows the likelihood ratios for the results where a positive test required agreement between the examining physicians.

CONCLUSIONS

LEVEL OF EVIDENCE Level 1.

STRENGTHS The study used a sound design and an accepted gold standard.

LIMITATIONS The methods of palpation are not adequately described.

The results are interesting in that a “positive” result required 2 examiners’ agreement. This suggests that clinicians

Table 46-7 Likelihood Ratio for the Presence of Palpable Splenomegaly vs the Presence of a Large Palpable Spleen^a

Test	Sensitivity	Specificity	LR+ (95% CI)	LR- (95% CI)	DOR (95% CI)
Palpable spleen	0.72	0.91	7.6 (4.5-12)	0.31 (0.14-0.56)	25 (8.5-73)
Distance between splenic border and costal margin > 4 cm	0.28	0.98	14 (4.6-41)	0.74 (0.50-0.89)	19 (5.2-69)

Abbreviations: CI, confidence interval; DOR, diagnostic odds ratio; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

^aThe test was considered positive only when both examiners agreed on the finding.

might have better accuracy when they ask for a second opinion about the palpation findings.

Reviewed by Alan N. Barkun, MD

TITLE Percussion of Traube's Space—A Useful Index of Splenic Enlargement.

AUTHORS Dubey S, Swaroop A, Jain R, Verma K, Garg P, Agarwal S.

CITATION *J Assoc Physicians India*. 2000;48(3):326-328.

QUESTION Is Traube space percussion useful in assessing splenic enlargement?

DESIGN Prospective, nonconsecutive patients, with an independent assessment by ultrasonography.

SETTING An Indian University hospital.

PATIENTS One hundred patients were medical inpatients.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

After Traube space percussion,¹ the findings were labeled as tympanic (resonant) or dull. Dullness to percussion is an abnormal finding that suggests splenomegaly. In addition, the spleen was also palpated with the patient positioned in the supine and right lateral decubitus positions. The clinician assessed the spleen as palpable or not palpable. Each patient was subsequently sent for ultrasonography. Splenomegaly was defined as a splenic longitudinal measurement of 12 cm or more on ultrasonography.

MAIN OUTCOME MEASURES

Sensitivity and specificity.

MAIN RESULTS

The prevalence of splenomegaly in this patient population was 36%. The splenic lengths among patients with ultrasonographically diagnosed splenomegaly were 13.1 ± 0.96 cm vs 9.42 ± 1.06 cm for those without splenic enlargement.

The results of Traube space percussion are shown in Table 46-8. The Quetelet index (a measure of body size) was higher among patients who had false-negative findings.

The diagnostic odds ratios (DORs), calculated from data provided in article, suggest that palpation (DOR, 25; 95% confidence interval [CI], 5.2-117) might be more accurate than percussion (DOR, 6.0; 95% CI, 2.4-15).

Table 46-8 Likelihood Ratios for Percussion, Palpation, and a Combination of the 2 Findings for Splenomegaly

Test	Sensitivity, %	Specificity, %	LR+ (95% CI)	LR- (95% CI)
Traube space percussion	67	75	2.7 (1.7-4.3)	0.44 (0.27-0.68)
Palpation	44	97	14 (3.5-58)	0.57 (0.43-0.77)
Palpation and percussion				
Both positive			14 (0.85-245)	
One positive			2.9 (2.0-4.4)	
Both negative			0.18 (0.10-0.33)	

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

CONCLUSIONS

LEVEL OF EVIDENCE Level 3.

STRENGTHS The study uses, overall, a sound design and an accepted gold standard.

LIMITATIONS The exact sequence of physical examination maneuvers is unclear. Therefore, a comparison of the performance of one technique to another cannot be done precisely, because we do not know the order in which the maneuvers were done. However, the maximal clinical utility appeared to be achieved when both percussion and palpation were considered.

These results in a different patient population appear to confirm the findings that Traube space percussion and palpation are useful for identifying splenomegaly. The authors did not report a standardized order of percussion and then palpation or vice versa. However, in considering the results of the findings, clinicians can conclude that both maneuvers should be performed. The presence of dullness to percussion confirms the importance of a palpable spleen. The absence of dullness to percussion modulates the importance of a palpable spleen and suggests that for some patients either the spleen is palpable, though not enlarged, or the examiner's findings of a palpable spleen are in error. If neither finding is present, the likelihood of splenomegaly is decreased.

REFERENCE FOR THE EVIDENCE

- Barkun AN, Camus M, Meagher TW, et al. How useful is Traube's space percussion in assessing splenic enlargement? *Am J Med*. 1989;87(5):562-566.

Reviewed by Alan N. Barkun, MD

This page intentionally left blank

CHAPTER 47

Does This Patient Have Strep Throat?

Mark H. Ebell, MD

Mindy A. Smith, MD

Henry C. Barry, MD

Kathy Ives, BS

Mark Carey, BS

CLINICAL SCENARIOS

In each of the following cases, the physician must decide whether the patient has group A β -hemolytic streptococcal pharyngitis (strep throat). In case 1, a 7-year-old boy presents in March without a cough but with 1 day of sore throat accompanied by fever, headache, moderate cervical adenopathy, and a markedly exudative and erythematous pharynx. His brother was recently diagnosed as having strep throat. In case 2, a 16-year-old presents with severe sore throat and anterior adenopathy for 3 days but no tonsillar enlargement, exudate, fever, or cough. In case 3, a 42-year-old woman presents with 5 days of sore throat and cough but no adenopathy, tonsillar enlargement, recent exposure to strep, or exudate.

WHY IS THE DIAGNOSIS IMPORTANT?

The 1995 National Ambulatory Medical Care Survey¹ found that sore throat is the third most common presenting complaint in office-based practice, accounting for 4.3% of visits. Sore throat is usually caused by direct infection of the pharyngeal tissue (pharyngitis). The differential diagnosis of pharyngitis is summarized in Table 47-1. Sore throat can also be caused by conditions such as gastroesophageal reflux disease, acute thyroiditis, persistent cough, and postnasal drainage because of allergic rhinitis or sinusitis. However, reliable estimates for the likelihood of these conditions among patients with sore throat are not available.

Untreated group A β -hemolytic streptococcal pharyngitis typically lasts 8 to 10 days. Patients are infectious during the period of acute illness and for approximately 1 week after. Antibiotic treatment decreases the severity of symptoms, reduces their duration by approximately 1 day,⁵ reduces the risk of transmission to others after 24 hours of treatment, and reduces the likelihood of suppurative complications and rheumatic fever.⁶ Suppurative complications include peritonsillar abscess (occurring in <1% of patients treated with antibiotics), retropharyngeal abscess, suppurative cervical lymphadenitis, bacteremia, and, by direct extension, otitis media, sinusitis, and mastoiditis.⁷ Rarely, the infection may lead to meningitis, pneumonia, or bacteremia. Rheumatic fever is a serious sequela of strep throat. Between 1 and 5 weeks after an episode of strep throat, a nonsuppurative inflammatory reaction results in fever, carditis, subcutaneous nodules, chorea, or migratory polyarthritis. Acute rheumatic fever now occurs infrequently in the United States, with a reported annual incidence of approximately 1 case per 1 000 000 population.⁸

Always doing a throat culture or rapid antigen test can lead to overtreatment of low-risk patients because of excessive false-positive results and undertreatment of high-risk patients because of excessive false-negative results. This

Table 47-1 Differential Diagnosis of Pharyngitis²⁻⁴

Etiology	Probability, %
Viral	50-80
Streptococcal	5-36
Epstein-Barr virus	1-10
<i>Chlamydia pneumoniae</i>	2-5
<i>Mycoplasma pneumoniae</i>	2-5
<i>Neisseria gonorrhoeae</i>	1-2
<i>Haemophilus influenzae</i> type b	1-2
Candidiasis	<1
Diphtheria	<1

approach also leads to increased cost.^{9,10} By using the preexamination likelihood of strep throat and the clinical examination, patients can potentially be divided into 3 groups: those with a high probability of strep throat, who could receive empiric antibiotic therapy (case 1, above); those with an intermediate probability of disease, who may require further diagnostic testing (case 2); and those with a low probability of disease, who may require only symptomatic therapy and appropriate follow-up rather than further diagnostic testing or treatment (case 3).¹¹

Pathophysiology

Group A β-hemolytic streptococci trigger an inflammatory response in pharyngeal cells that is responsible for many of the signs and symptoms of pharyngitis. Interleukins 1 and 6, tissue necrosis factor, and prostaglandins cause the febrile response; prostaglandins and bradykinin cause pain; and prostaglandins and nitric oxide cause vasodilation and edema, manifested as erythema and swelling of the tonsillar pillars, uvula, and soft palate. Lysosomal enzymes and oxygen free radicals, although part of the body's response to infection, also cause tissue damage. This tissue damage, in addition to the pustular nature of the group A β-hemolytic streptococcal infection, results in a creamy exudate from the tonsillar pillars. The

pharynx is drained primarily by the anterior cervical nodes, which may become tender and enlarged during infection.¹²

Although group A β-hemolytic streptococcus is not part of the normal flora of the human throat, the asymptomatic carrier rate is 5.0% to 21% in children between the ages of 3 and 15 years. It is lower in children younger than 3 years (1.9%-7.1%) and in older adolescents and adults (2.4%-3.7%).¹³

METHODS

Search Strategy and Quality Review

For the evaluation of individual signs and symptoms, we identified studies of the diagnosis of group A β-hemolytic streptococcal pharyngitis in patients complaining of sore throat. All studies included at least 300 patients, collected data prospectively, and used throat culture as the reference standard. Examiners were unaware of the results of rapid tests or throat cultures for strep when they performed the medical history and physical examination. All articles therefore represent level 1 evidence according to previously published criteria for the evaluation of study quality.¹⁴ The results for a variable are reported only if more than 1 study reported data for that variable.

The MEDLINE search used the following Medical Subject Headings: (“sensitivity and specificity” or “predictive value of tests” or “medical history taking” or “physical examination”) and “pharyngitis.” This search identified 917 articles. In 2 cases, authors were contacted to provide additional information or to clarify a point in the article. Unpublished data were not sought. Seventeen studies (15 in English, 1 in German, and 1 in Spanish) met all of the inclusion criteria described above except study size. Nine studies included at least 300 patients; they are shown in Table 47-2.¹⁵⁻²³ Each study was reviewed independently by 2 clinical investigators, and discrepancies were resolved by discussion. In addition, any articles developing or validating a clinical prediction rule were identified. The included studies reported data for 5453 patients, whereas the 8 excluded studies reported data for only 1182 patients.

Table 47-2 Studies Included

Source, y	Setting	Population Presenting With Complaint of Sore Throat	Patients, No.	Prevalence of Strep Throat Pharyngitis, %
McIsaac et al, ²² 1998	Office	Adults and children	520	29
Kljakovich, ²⁰ 1993	Office	Adults and children	329	12
Reed et al, ²¹ 1990	Urgent care	Adults and children	806	25
Crawford et al, ²³ 1979	Outpatient	Adults and children	472	11
Komaroff et al, ¹⁸ 1986	Office	Adults only	693	10
Walsh et al, ¹⁹ 1975	Office	Adults only	418	15
Steinhoff et al, ¹⁵ 1997	Office	Children only	450	24
Kaplan et al, ¹⁶ 1971	Emergency department	Children only	624	35
Stillerman and Bernstein, ¹⁷ 1961	Office	Children only	1141	36

Statistical Methods

The positive likelihood ratio ($LR+$) and negative likelihood ratio ($LR-$) were calculated for medical history and physical examination findings. The $LR+$ ($LR+ = \text{sensitivity}/[100 - \text{specificity}]$) is a measure of how well a positive result rules in disease, whereas the $LR-$ ($LR- = [100 - \text{sensitivity}]/\text{specificity}$) is a measure of how well a negative result rules out disease. A random-effects estimate was calculated for the sensitivity, specificity, $LR+$, and $LR-$ when the χ^2 statistic suggested homogeneity ($P > .05$) or when random- and fixed-effects models gave similar values. Otherwise, a range for each variable is shown.

The area under the receiver operating characteristic (ROC) curve is a measure of the diagnostic accuracy of a test.²⁴ Specifically, a greater area corresponds to a greater ability to discriminate between patients with and without strep throat. An area of 1.0 under the ROC curve means the test is perfect, whereas an area of 0.5 means it is no better than chance. In this study, the area was calculated with the method of Moses et al.²⁵ It was not possible to generate an ROC curve for some signs and symptoms if fewer than 3 studies reported their sensitivity and specificity.

PRECISION AND ACCURACY

Symptoms

Classically, the streptococcal sore throat is of abrupt onset in older children and adults. Symptoms may be less focal and more gradual in younger children.²⁶ Throat pain is typically described as severe and is associated with difficulty in swallowing. Fever is moderate (reported temperature range, 39°C to 40.5°C). Chills may be present but rigor is not typical. Strep throat is also classically associated with malaise, headache, mild neck stiffness, and gastrointestinal symptoms such as anorexia, nausea, vomiting, and abdominal pain. However, these features may be present in only 35% to 50% of patients and have not been verified by objective studies of diagnosis. Abdominal symptoms may be more common in younger patients, although more recent studies have not confirmed this as an independent predictor.²⁷

Signs

Examination of the throat may reveal erythema and edema of the pharynx and uvula and diffuse erythema and hypertrophy of the lymphoid tissue in the posterior pharynx. The posterior pharynx and tonsillar pillars may be covered with a gray-white membrane or exudate. The pharynx is often described as beefy or bright red, with the color ending abruptly at the soft palate. Petechiae may be present on the soft palate. Tonsils are commonly swollen and erythematous and covered with a punctate or confluent gray-white exudate. The breath is characteristically foul.

The anterior cervical lymph nodes are often tender and enlarged, especially at the angle of the jaw. This sign occurs

early in the course of infection. When present, the characteristic scarlatiniform (“scarlet fever”) rash is one of fine erythematous papules beginning on the trunk and spreading to the extremities but sparing the palms and soles. The rash blanches to pressure and has a sandpaper feel. It is associated with enlarged papillae on a coated tongue that may later become denuded (“strawberry tongue”), circumoral pallor and hyperpigmentation, or accentuation of the rash in the skin creases. This is especially prominent in the antecubital fossae (Pastia sign). In young children, there may be excoriations around the nares. The rash typically subsides in 6 to 9 days and may be followed by desquamation of the palms and soles. Pharyngeal vesicles and ulcers are associated with viral upper respiratory tract infections; their presence reduces the likelihood that a sore throat is caused by group A β -hemolytic streptococci.

Properly viewing the pharynx can be challenging. Adequate examination of the throat requires elevation of the soft palate and uvula and depression of the posterior tongue. Although a tongue blade can help, patients often gag, cough, or bite. The pharynx can sometimes be viewed without a tongue blade by having the patient pant. Small children can be asked to imitate a puppy as a way of encouraging them to pant.

Precision of Symptoms and Signs

Although only limited data are available on the precision of symptoms and signs of streptococcal pharyngitis, these data suggest that observer reliability is high. Komaroff et al¹⁸ had 2 blinded observers examine the same randomly sampled patients and found 88% agreement on 187 medical history and physical examination items, although the κ test was not used to evaluate agreement beyond chance. In another study, a physician and physician assistant both examined the ears, nose, throat, cervical nodes, and chest of 63 patients. Only 1 discrepancy, in examination of cervical adenopathy, was observed.¹⁹ No data were available regarding the ability of physicians to distinguish tonsillar from pharyngeal exudate or regarding the precision of the examination in patients who have undergone tonsillectomy.

Diagnostic Accuracy of Symptoms and Signs

The sensitivity, specificity, $LR+$, and $LR-$ for variables that are reported in at least 2 studies are shown in Table 47-3. The variables with the greatest area under the ROC curve, and hence the best ability to discriminate between patients with and without strep throat, were pharyngeal or tonsillar exudate, fever by history, tonsillar enlargement, tenderness or enlargement of the anterior cervical lymph nodes, and absence of cough.

Findings that were similar across studies, had the greatest $LR+$, and were therefore best at ruling in disease were the presence of tonsillar exudate ($LR+, 3.4$), pharyngeal exudate ($LR+, 2.1$), and strep throat exposure in the previous 2 weeks ($LR+, 1.9$). The absence of findings was not efficient at ruling out disease, with the lowest $LR-$ found for the absence of tender anterior cervical nodes ($LR-, 0.60$),

Table 47-3 Accuracy for Medical History and Physical Examination Elements in the Diagnosis of Strep Throat^a

Symptoms and Signs	Patients, No.	Accuracy	Sensitivity (95% CI) or Range	Specificity (95% CI) or Range	LR+ (95% CI) or Range	LR- (95% CI) or Range
Any exudates ^{15,16,18-21}	3268	0.68	0.21-0.58	0.69-0.92	1.5-2.6	0.66-0.94
Reported fever ^{15,17,20,21}	3232	0.68	0.3-0.92	0.23-0.90	0.97-2.6	0.32-1.0
Measured temperature >37.8°C ^{15,17,18,21}	3091	0.68	0.11-0.84	0.43-0.96	1.1-3.0	0.27-0.94
Anterior cervical nodes swollen/enlarged ^{15,16,18,20-23}	3831	0.67	0.55-0.82	0.34-0.73	0.47-2.9	0.58-0.92
Pharyngeal exudates ^{18,22,23}	1673	0.65	0.03-0.48	0.76-0.99	2.1 (1.4-3.1)	0.90 (0.75-1.1)
Tonsillar swelling/enlargement ¹⁸⁻²²	2703	0.65	0.56-0.86	0.56-0.86	1.4-3.1	0.63 (0.56-0.72)
Tonsillar or pharyngeal exudates ^{15,16,19,21}	2246	0.65	0.28-0.61	0.62-0.88	1.8 (1.5-2.3)	0.74 (0.66-0.82)
Anterior cervical nodes tender ^{15,16,18,22}	2280	0.64	0.32-0.66	0.53-0.84	1.2-1.9	0.60 (0.49-0.71)
Tonsillar exudates ^{20,22}	840	0.64	0.36 (0.21-0.52)	0.71-0.98	3.4 (1.8-6.0)	0.72 (0.60-0.88)
No cough ^{15-19,21,23}	5122	0.63	0.51-0.79	0.36-0.68	1.1-1.7	0.53-0.89
No coryza ^{15-19,22}	3846	0.57	0.42-0.84	0.20-0.70	0.86-1.6	0.51-1.4
Myalgias ^{18,21,22}	2003	0.57	0.49 (0.43-0.56)	0.52-0.69	1.4 (1.1-1.7)	0.93 (0.86-1.0)
History of sore throat ^{16,17,21,22}	3090	0.57	0.18-0.93	0.09-0.86	1.0-1.1	0.55-1.2
Headache ^{17,18,22}	2350	0.56	0.48 (0.42-0.53)	0.50-0.80	0.81-2.6	0.55-1.1
Pharynx injected ^{16,18,19,22}	2939	0.54	0.43-0.99	0.03-0.62	0.66-1.6	0.18-6.4
Measured temperature ≥38.3°C ^{16,22,23}	1096	0.53	0.22-0.58	0.53-0.92	0.68-3.9	0.54-1.3
Nausea ^{17,21}	1941	0.52	0.26 (0.12-0.43)	0.52-0.98	0.76-3.1	0.91 (0.86-0.97)
Duration <3 d ^{20,22}	824	0.43	0.26-0.93	0.59 (0.54-0.64)	0.72-3.5	0.15-2.2
Male sex ^{21,22}	1325	0.39	0.11-0.56	0.39-0.86	0.87 (0.72-1.0)	1.1 (0.93-1.2)
Palatine petechiae ^{18,22}	1202	NC	0.07 (0.02-0.14)	0.95 (0.92-0.96)	1.4 (0.48-3.1)	0.98 (0.92-1.1)
Strep exposure previous 2 wk ^{18,19,22,23}	2091	NC	0.19 (0.12-0.27)	0.87-0.94	1.9 (1.3-2.8)	0.92 (0.86-0.99)
Rash ^{17,21,22}	2356	NC	0.04 (0.03-0.06)	0.79-0.99	0.06-35	0.90-1.1

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio; NC, receiver operating characteristic curve not calculable.

^aWhen one of these operating characteristics was homogeneous ($P > .05$ for the χ^2 test), the summary value and a 95% CI are given. Where they are heterogeneous, only the range is given. Variables are given in the order of the area under the receiver operating characteristic curve, where one could be drawn.

tonsillar enlargement (LR-, 0.63), and tonsillar or pharyngeal exudate (LR-, 0.74). A physician's overall estimate of the probability of strep throat was measured in 2 small studies, which found LR+ values of 3.0 and 1.7 and LR- values of 0.36 and 0.60.^{28,29}

Estimating the Pretest Probability of Strep Throat

Clinical decision-making requires an estimate of the pretest probability, in this case the probability that the patient has strep throat before examination. This estimate should be based primarily on the patient's age, the clinical setting, and the season.

In general, strep throat is more common in children than among infants and adults: group A β-hemolytic streptococcal bacteria can be isolated by throat culture in 24% to 36% of children¹⁵⁻¹⁷ and in 5% to 24% of adults with sore throat.^{4,18,19,28,30,31} Breese³² reports that the likelihood of strep throat peaks between the ages of 5 and 10 years, although it may occur somewhat earlier now because more children are in daycare.

A reasonable estimate of the pretest probability of strep throat in an unselected office-based adult population is 5% to 10% and in an unselected pediatric population, 20% to 25% (Table 47-2). The prevalence of strep throat is also higher among patients treated in emergency departments or urgent care centers than in office practice.^{33,34} Because strep throat is more common in autumn and winter,^{19,29,32} it may be appropriate to adjust these estimates upward during those seasons and downward in spring and summer.

Clinical Prediction Rules

Because individual signs and symptoms are not accurate enough to make a diagnosis, clinical prediction rules have been developed that use several key elements of the medical history and physical examination to predict the probability of strep throat. Using a clinical prediction rule gives a physician a rational basis for assigning a patient to a low-risk category (requires neither testing nor treatment), a high-risk category (empiric antibiotic therapy may be indicated), or a moderate-risk category (may require further diagnostic testing).^{9,10}

Table 47-4 Clinical Prediction Rules for the Prediction of Strep Throat in Patients With Sore Throat

Source, y	Description	Population	Accuracy	Comment
Centor et al, ²⁸ 1981	Simple 4-variable additive score	236 US adult patients in the emergency department	Area under the ROC curve 0.79 (good accuracy)	Successfully validated in 3 new adult populations ³⁴⁻³⁶
Dobbs, ²⁹ 1996	Bayesian score with 14 variables	206 Patients >4 y in a British general practice	71% Sensitive, 71% specific	No prospective validation, relies on one physician's examination skills
Breese, ³² 1977	Algorithm based on 4 signs and symptoms and patient age (see Figure 47-2)	670 US children in original study, 892 children in validation study ³⁷	LR+ = 2, LR- = 0.75 in pediatric validation study ³⁷	The lowest-risk group still has a 6% ³² to 16% ³⁷ risk of strep throat; similar results in study of adults and children ²¹
Clancy et al, ³⁸ 1988	3-Item additive score based on medical history alone	1237 US adult patients in 2 emergency departments and 189 patients at a student health service	Area under the ROC curve 0.70 to 0.74, depending on setting; 85% sensitive and 42% specific	No prospective validation
Hoffman, ³ 1992	7-Item score, including age	1783 Patients in a Danish general practice	LR+ = 1.3, LR- = 0.2	Rule has low specificity (26%)
Komaroff et al, ¹⁸ 1986	6-Item additive score	693 US adult outpatients	Results presented as nomogram only	Not prospectively validated
Walsh et al, ¹⁹ 1975	Algorithm based on 4 signs and symptoms (see Figure 47-1)	418 US adults with sore throat at an HMO outpatient clinic	High risk = 28% strep; moderate risk = 15% strep; low risk = 4% strep	In a prospective study in adults and children by Crawford et al, ²³ 23% of high-risk patients, 12% of moderate-risk patients, and 3% of low-risk patients had a positive throat culture result for strep
Meland et al, ³⁵ 1993	Algorithm based on 4 signs and symptoms	133 Norwegian adults and children with sore throat	High risk = 62% strep; moderate risk = 34% strep; low risk = 10% strep	Not prospectively validated
McIsaac et al, ³⁷ 2000	Algorithm based on 4 signs and symptoms and patient age	621 Canadian adults and children >3 y old and presenting to family physicians	Risk stratified into 5 levels, from 1%-51%	Good prospective validation in primary care practices, including both children and adults

Abbreviations: HMO, health maintenance organization; LR+, positive likelihood ratio; LR-, negative likelihood ratio; ROC, receiver operating characteristic curve.

Table 47-4 summarizes previous efforts to develop or validate clinical prediction rules for the diagnosis of strep throat. One of the best validated is a simple 4-item clinical prediction rule developed by Centor et al.²⁸ The Centor score has been validated in 3 distinct adult populations³⁴⁻³⁶ and considers 4 signs and symptoms: tonsillar exudate, swollen tender anterior cervical nodes, absence of cough, and a history of fever. The rule is accurate, with an area under the ROC curve of 0.79. One point is assigned for each of the patient's signs and symptoms, and the sum is used to determine the likelihood of strep throat (Figure 47-1).²⁸ The presence of 3 or 4 findings increases the probability of strep throat. For example, a patient with a pretest probability of 10% and a Centor score of 4 would have a 41% probability of strep throat. Patients with none or 1 of the cardinal findings have a very low risk of strep throat, and it may be appropriate to forgo testing or treatment in this group. The Centor clinical prediction rule has not been validated in younger patients. Recently, McIsaac et al³⁷ have modified Centor's score and validated it prospectively in a mixed group of adults and children (Figure 47-2). Another rule, developed by Walsh et al,¹⁹ has been validated prospectively in a mixed population of adults and children (Figure 47-3).

The Breese score has been prospectively validated in a large pediatric population.^{32,38,39} However, a low Breese score does not rule out strep: 14% of children with a score of 20

1. Assign 1 point for each of the following clinical characteristics: (1) history of fever, (2) anterior cervical adenopathy, (3) tonsillar exudate, and (4) absence of cough.

2. Find the column that most closely matches the pretest probability of strep in the patient and look down the column to the row that matches the patient's number of points to determine the probability of strep.

Points, No.	Likelihood Ratio	Pretest Probability of Strep Throat, %						
		5	10	15	20	25	40	50
0	0.16	1	2	2	3	5	10	14
1	0.3	2	3	5	7	9	17	23
2	0.75	4	8	12	16	20	33	43
3	2.1	10	19	27	34	41	58	68
4	6.3	25	41	53	61	68	81	86

Figure 47-1 Centor Clinical Prediction Rule for the Diagnosis of Strep Throat in Adults

or less had a positive throat culture result. In addition, it requires the results of a white blood cell count, which may not be immediately available in many outpatient practices.

1. Add Up Points for Patient		Points
Symptom or Sign		
History of fever or measured temperature $> 38^{\circ}\text{C}$		1
Absence of cough		1
Tender anterior cervical adenopathy		1
Tonsillar swelling or exudates		1
Age < 15 y		1
Age ≥ 45 y		-1

Points	Likelihood Ratio	% With Strep
		(Patients With Strep/Total)
-1 or 0	0.05	1 (2/179)
1	0.52	10 (13/134)
2	0.95	17 (18/109)
3	2.5	35 (28/81)
4 or 5	4.9	51 (39/77)

Figure 47-2 McIsaac Modification of the Centor Strep Score

Data from a group of 167 children aged 3 years or older and 453 adults in Ontario, Canada. Baseline risk of strep 17% in this population. Reprinted with permission from McIsaac et al.³⁷

CLINICAL SCENARIOS—RESOLUTIONS

Case 1 describes a child with a high likelihood (51%) of streptococcal pharyngitis according to the McIsaac clinical rule (Figure 47-2). In fact, the likelihood of strep throat is probably even higher because of his recent exposure. The physician might wish to treat without further diagnostic confirmation. Children with a low or intermediate probability of strep and a negative rapid antigen test result should still have a backup throat culture.

In case 2, an adolescent has a pretest probability (estimate, 15%) falling between that of adults and children. In this age group, infectious mononucleosis is also a relatively common cause of sore throat. Assuming a pretest probability of 15% and 2 points on the Centor score, he has a 12% probability of strep throat (Figure 47-1). The physician should decide whether to recommend a rapid antigen test to clarify the need for treatment. Newer rapid tests have a sensitivity (85%) and specificity (93%) close to that of throat culture.⁴⁰ If a patient with a 12% probability of strep throat has a negative rapid test result, the likelihood of strep decreases to only 2%, whereas if the results are positive, it increases to 62%.

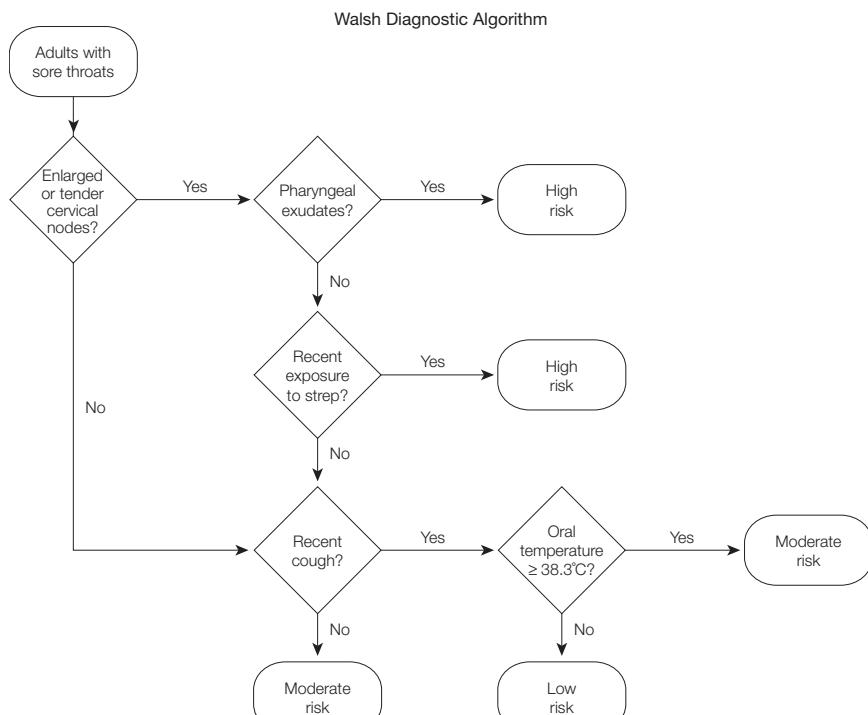


Figure 47-3 Walsh Algorithm for Evaluating Cases of Adults With Sore Throats

Risk Group	% Strep by Culture	
	Original ¹⁹	Validation ²³
High	28	23
Moderate	15	12
Low	4	3

Finally, the woman in case 3 has none of the cardinal characteristics of strep throat in the Centor score and the Walsh algorithm and therefore has a low (2%-3%) probability of strep throat. It may be appropriate to reassure this patient that strep throat is unlikely and to consider non-bacterial causes of sore throat. Some would argue that a throat culture is always necessary and that treatment should be delayed until culture results become available.⁴¹ However, this approach ignores the accuracy of rapid antigen tests, particularly when used in tandem with an accurate assessment of the pretest probability with a clinical score. A recent study of 30000 episodes of sore throat found that changing from a policy encouraging throat culture to one encouraging the use of a rapid antigen test only decreased the percentage of patients with sore throat receiving a culture from 65% to 13%, without an adverse increase in suppurative complications.⁴²

THE BOTTOM LINE

This study further confirms that no single element of the medical history or physical examination is powerful enough to confirm the probability of streptococcal pharyngitis. Instead, physicians should consider a combination of findings, including tonsillar exudate, tender or enlarged anterior cervical nodes, the absence of cough, and a history of fever (or measured temperature >38°C). A rational approach to therapy integrates these findings with the patient's age and the clinical setting, the information from Figures 47-1 and 47-2, the results of rapid antigen testing or throat culture, and the clinician's own judgment.

Author Affiliations at the Time of the Original Publication

Michigan State University, East Lansing (Drs Ebell, Smith, and Barry, Ms Ives, and Mr Carey); First Consulting Group, Okemos, Michigan (Ms Ives).

Acknowledgments

This work was supported by the Michigan Consortium for Family Practice Research, a research center funded by the American Academy of Family Physicians.

We acknowledge the assistance and detailed feedback of Peter Blomgren, MD, Matthew Gillman, MD, and David L. Simel, MD, MHS. We also thank Robert Centor, MD, for his assistance with the calculation of the area under the receiver operating characteristic curve in his clinical prediction rule.

REFERENCES

- National Center for Health Statistics. 1995 National Ambulatory Medical Care Survey [CD-ROM Series 13, No. 11]. Hyattsville, MD: National Center for Health Statistics; 1995.
- Komaroff A, Aronson MD, Pass TM, Ervin CT, Branch WT. Serologic evidence of chlamydial and mycoplasmal pharyngitis in adults. *Science*. 1983;222(4626):927-928.
- Hoffman S. An algorithm for a selective use of throat swabs in the diagnosis of group A streptococcal pharyngo-tonsillitis in general practice. *Scand J Prim Health Care*. 1992;10:295-300.
- Seppala H, Lahtonen R, Ziegler T, et al. Clinical scoring system in the evaluation of adult pharyngitis. *Arch Otolaryngol Head Neck Surg*. 1993;119(3):288-291.
- Denny FW. Effect of treatment on streptococcal pharyngitis: is the issue really settled? *Pediatr Infect Dis*. 1985;4(4):352-354.
- Snellman L, Stang H, Stang J, et al. Duration of positive throat cultures for group A streptococci after initiation of antibiotic therapy. *Pediatrics*. 1993;91(6):1166-1170.
- Dajani A, Taubert K, Ferrieri P, et al. Treatment of acute streptococcal pharyngitis and prevention of rheumatic fever: a statement for health professionals. *Pediatrics*. 1995;96(4 pt 1):758-764.
- Centers for Disease Control and Prevention. Summary of notifiable diseases, United States, 1997. *Morb Mortal Wkly Rep CDC Surveill Summ*. 1998;46(54):1-87.
- McIsaac WJ, Goel V. Effect of an explicit decision-support tool on decisions to prescribe antibiotics for sore throat. *Med Decis Making*. 1998;18(2):220-228.
- Tompkins RK, Burnes DC, Cable WE. An analysis of the cost-effectiveness of pharyngitis management and acute rheumatic fever prevention. *Ann Intern Med*. 1977;86(4):481-492.
- Pauker SG, Kassirer J. The threshold approach to clinical decision-making. *N Engl J Med*. 1980;302(20):1109-1117.
- Bisno AL. Streptococcus pyogenes. In: Mandell GL, Bennett JE, Dolin R, eds. *Principles and Practice of Infectious Diseases*. 4th ed. New York, NY: Churchill-Livingstone; 1995.
- Gunnarsson RK, Holm SE, Soderstrom M. The prevalence of beta-haemolytic streptococci in throat specimens from healthy children and adults: implications for the clinical value of throat cultures. *Scand J Prim Health Care*. 1997;15(3):149-155.
- Holleman DR, Simel DL. Does the clinical examination predict airflow limitation? *JAMA*. 1995;273(4):313-319.
- Steinhoff MC, Khalek MK, Khalaf N, et al. Effectiveness of clinical guidelines for the presumptive treatment of streptococcal pharyngitis in Egyptian children. *Lancet*. 1997;350(9082):918-921.
- Kaplan EL, Top FH, Dudding BA, Wannamaker LW. Diagnosis of streptococcal pharyngitis: differentiation of active infection from the carrier state in the symptomatic child. *J Infect Dis*. 1971;123(5):490-501.
- Stillerman M, Bernstein SH. Streptococcal pharyngitis: evaluation of clinical syndromes in diagnosis. *Am J Dis Child*. 1961;101:476-489.
- Komaroff AL, Pass TM, Aronson MD, et al. The prediction of streptococcal pharyngitis in adults. *J Gen Intern Med*. 1986;1(1):1-7.
- Walsh BT, Bookheim WW, Johnson RC, Tompkins RK. Recognition of streptococcal pharyngitis in adults. *Arch Intern Med*. 1975;135(11):1493-1497.
- Kljakovac M. Sore throat presentation and management in general practice. *N Z Med J*. 1993;106(963):381-383.
- Reed BD, Huck W, French T. Diagnosis of group A beta-hemolytic streptococcus using clinical scoring criteria, Directigen 1-2-3 group A streptococcal test, and culture. *Arch Intern Med*. 1990;150(8):1727-1732.
- McIsaac WJ, White D, Tannenbaum D, Low DE. A clinical score to reduce unnecessary antibiotic use in patients with sore throat. *CMAJ*. 1998;158(1):75-83.
- Crawford G, Brancato F, Holmes KK. Streptococcal pharyngitis: diagnosis by gram stain. *Ann Intern Med*. 1979;90(3):293-297.
- Kardaun JW, Kardaun OJ. Comparative diagnostic performance of three radiological procedures for the detection of lumbar disk herniation. *Methods Inf Med*. 1990;29(1):12-22.
- Moses LE, Shapiro D, Littenberg B. Combining independent studies of a diagnostic test into a summary ROC curve: data-analytic approaches and some additional considerations. *Stat Med*. 1993;12(14):1293-1316.
- Breese BB, Disney FA. The accuracy of diagnosis of beta streptococcal infections on clinical grounds. *J Pediatr*. 1954;44(6):670-673.
- Kreher NE, Hickner JM, Barry HC, Messimer SR. Do gastrointestinal symptoms accompanying sore throat predict streptococcal pharyngitis? an UPRNet study. *J Fam Pract*. 1998;46(2):159-164.
- Centor RM, Witherspoon JM, Dalton HP, Brody CE, Link K. The diagnosis of strep throat in adults in the emergency room. *Med Decis Mak*. 1981;1(3):239-246.
- Dobbs F. A scoring system for predicting group A streptococcal throat infection. *Br J Gen Pract*. 1996;46(409):461-464.
- Foong BB, Yassim M, Chia YC, Kang BH. Streptococcal pharyngitis in a primary care clinic. *Singapore Med J*. 1992;33:597-599.

31. Kahan E, Appelbaum T, Bograd H, et al. Should nurses in Israeli primary care clinics be expected to manage streptococcal throat infections? *Public Health.* 1995;109(5):347-351.
32. Breese BB. A simple scorecard for the tentative diagnosis of streptococcal pharyngitis. *Am J Dis Child.* 1977;131(5):514-517.
33. Centor RM, Meier FA, Dalton HP. Throat cultures and rapid tests for diagnosis of group A streptococcal pharyngitis. *Ann Intern Med.* 1986;105(6):892-899.
34. Wigton RS, Connor JL, Centor RM. Transportability of a decision rule for the diagnosis of streptococcal pharyngitis. *Arch Intern Med.* 1986;146(1):81-83.
35. Meland E, Digranes A, Skjaerven R. Assessment of clinical features predicting streptococcal pharyngitis. *Scand J Infect Dis.* 1993;25(2):177-183.
36. Poses RM, Cebul RD, Collins M, Fager SS. The importance of disease prevalence in transporting clinical prediction rules. *Ann Intern Med.* 1986;105(4):586-589.
37. McIsaac WJ, Goel V, To T, Low DE. The validity of a sore throat score in family practice. *CMAJ.* 2000;163(7):811-815.
38. Clancy CM, Centor RM, Campbell MS, Dalton HP. Rational decision making based on history: adult sore throats. *J Gen Intern Med.* 1988;3(3):213-217.
39. Funamura JL, Berkowitz CD. Applicability of a scoring system in the diagnosis of streptococcal pharyngitis. *Clin Pediatr (Phila).* 1983;22(9):622-626.
40. Pichichero ME, Disney FA, Green JL, et al. Comparative reliability of clinical, culture, and antigen detection methods for the diagnosis of group A beta-hemolytic streptococcal tonsillopharyngitis. *Pediatr Ann.* 1992;21(12):798-805.
41. Pichichero ME. Cost-effective management of sore throat: it depends on the perspective. *Arch Pediatr Adolesc Med.* 1999;153(7):672-680.
42. Webb KH, Needham CA, Kurtz SB. Use of a high-sensitivity rapid strep test without culture confirmation of negative results. *J Fam Pract.* 2000;49(1):34-38.

Prepared by David L. Simel, MD, MHS
Reviewed by Jane Kim, MD, and Kathy Vanenkevort, BS

CLINICAL SCENARIO

A 19-year-old college student has a severe sore throat and a mild fever (temperature 38.3°C), and he feels bad. The symptoms have been present for 4 days and initially started with a dry cough. There is a pharyngeal exudate, but only on the left side of the posterior pharynx. His neck reveals tender adenopathy. Should you assume he has streptococcal pharyngitis and start treatment?

UPDATED SUMMARY ON STREPTOCOCCAL PHARYNGITIS

Original Review

Ebell MH, Smith MA, Barry HC, Ives KY, Carey M. Does this patient have strep throat? *JAMA*. 2000;284(22):2912-2918.

UPDATED LITERATURE SEARCH

Our literature search used the parent search for The Rational Clinical Examination series, combined with the search term “pharyngitis,” for the years 2000 to August 2005. Because the original publication supported the use of the Centor score, we reviewed studies that further explored and validated the use of this score in patient populations that included adults, rather than children alone, with pharyngitis. We identified 27 potentially relevant articles for further review, of which 6 warranted closer assessments. Three of those studies¹⁻³ validated the impression that the individual symptoms and signs for streptococcal pharyngitis are not diagnostically useful, so multivariate models that include combinations of findings must be used. However, none of these studies included prospective model validation.

NEW FINDINGS

- The Centor score alone is inadequate for making a correct diagnosis in patients with 2 or more symptoms.
- The clinical examination improves markedly when the Centor score is used to identify patients for rapid streptococcal tests.

- Treatment decisions based on the Centor score, without rapid testing, depend more on the prevalence of disease and benefit/risk of treatment rather than useful likelihood ratios (LRs).

Details of the Update

A study of Israeli adults older than 16 years and with pharyngitis provided a unique patient sample by including those with a Centor score of only 0 to 1.² Most other studies exclude these patients from their analysis, focusing only on those with at least 2 of 4 symptoms. However, 38% of the patients had a mild pharyngitis presentation, with a 0 to 1 score; the LR for such a score is 0.16 and only 5% of patients with a score of 0 to 1 will have a positive culture result. When the investigators created a multivariable model (7 symptoms and signs) for this patient group that included a broad spectrum of disease, only pharyngeal exudates remained significant at predicting culture positivity (positive LR, 1.8; 95% confidence interval [CI], 1.5-2.2; negative LR, 0.27; 95% CI, 0.13-0.53).

The Centor score may not work equally well for children. To evaluate this, McIsaac et al⁴ assembled a population of patients aged 3 to 69 years, and they validated their modified Centor score. The score was modified by age, with points added or subtracted as follows: aged 3 to 14 years, add 1 point; aged 15 to 44 years, add 0 points; aged 45 years or older, subtract 1 point. After adjusting for age, the investigators compared the modified Centor score to culture for those patients with a score greater than or equal to 2. The prevalence of disease was 22% for adults but 34% for children younger than 18 years. The modified Centor score did not appreciably change the LR for adults, but it did have an effect on children. At high scores, the LRs differ, with an LR of 1.6 (95% CI, 0.5-5.0) for a modified Centor score of 4 to 5 in adults that improves to an LR of 4.0 (95% CI, 2.7-6.0) for children younger than 18 years. After evaluating a variety of treatment strategies, the authors reported predictive values based on the modified Centor scores (Table 47-5).

A pragmatic study assessed the use of the Centor score in adults (>18 years) to identify patients for point-of-care testing and throat cultures.⁵ The study is retrospective, so we cannot determine the number of patients evaluated for

Table 47-5 Modified Centor Scores

	Centor Score, Adults (≥18 y) ^a	Modified Centor Score, Children (3-17 y)
	LR (95% CI)	LR (95% CI)
Score 4	1.2 (0.62-2.2)	4.0 (2.7-6.0)
Score 2-3	1.3 (0.85-1.9)	0.69 (0.59-0.83) ^b
Score 0-1	0.26 (0.14-0.48)	Uncertain

Abbreviations: CI, confidence interval; LR, likelihood ratio.

^aSummary LR from McIsaac et al,⁴ Atlas et al,⁵ and Chazan et al.²^bSummary LR from McIsaac et al.⁴**Table 47-6** Positive Predictive Value of Treatment According to the Modified Centor Score

Modified Centor Score	Antibiotic Treatment Strategy	Positive Predictive Value of Decision to Treat, % (95% CI)	Negative Predictive Value of Decision to Not Treat, % (95% CI)
Adults (≥18 y) ^a			
4	Treat with antibiotics	84 (73-90)	94 (90-96)
2-3	Rapid test; treat if positive result		
Children (3-17 y)			
2-5 (All children with sore throat)	All get rapid test; treat for a positive rapid test result and culture those with a negative rapid test result	98 (94-99)	100 (98-100)

Abbreviation: CI, confidence interval.

^aMcIsaac et al.⁴

acute pharyngitis who did not undergo point-of-care testing. However, prevalence of pharyngitis is similar to that in other published studies. As in other studies, the Centor score alone had an LR of 1.5 (95% CI, 1.2-1.9) for patients with at least 2 findings, and for those with 0 to 1 finding the LR was 0.35 (95% CI, 0.16-0.75). The value of point-of-care testing highlighted its utility when combined with the Centor score in both ruling in and ruling out disease. A positive rapid streptococcal test result in patients with at least 2 Centor score findings had an LR of 179; the Centor score did not affect the LR when the point-of-care testing result was negative, because the LR was 0.09 for a broad range of 0 to 4 symptoms.

IMPROVEMENTS IN THE DATA PRESENTED IN THE ORIGINAL PUBLICATION

These newer studies confirm the utility of the Centor score for certain patients while highlighting the weaknesses. The disease prevalence is higher in children (<18 years), and an age-adjusted score improves the performance of the Centor score. Point-of-care testing, in combination with the Centor score, is valuable.

CHANGES IN THE REFERENCE STANDARD

The throat culture continues to be the recognized reference standard for the diagnosis of group A β-hemolytic streptococci. However, a positive throat culture or rapid antigen test result provides adequate confirmation of the presence of group A β-hemolytic streptococci in the pharynx and is accepted as a pragmatic reference standard.

RESULTS OF LITERATURE REVIEW

The modified Centor score can direct the antibiotic treatment strategy (Table 47-6).

EVIDENCE FROM GUIDELINES

The diagnosis of acute group A streptococcal pharyngitis should be suspected on clinical and epidemiologic grounds and then supported by performance of a laboratory test.⁶ However, streptococcal pharyngitis is not the etiology of most cases of pharyngitis, so antibiotics are usually unnecessary. This is especially important, given the increasing concerns about antibiotic resistance. For adults, empiric treatment is not recommended for patients with a Centor score less than or equal to 3. For individuals with 2 or more symptoms, rapid antigen testing should guide treatment.

CLINICAL SCENARIO—RESOLUTION

This young college student has what initially seemed like a viral illness, heralded by a dry cough. However, his symptoms seem to have progressed relatively quickly, and he now has a Centor score of 3 (fever, exudates, and tender cervical adenopathy). Although this seems like streptococcal pharyngitis, he could also have mononucleosis or a variety of other infectious etiologies. The Centor score confers an LR not much different from 1, but the results of a rapid streptococcal test would ensure the diagnosis.

STREPTOCOCCAL PHARYNGITIS—MAKE THE DIAGNOSIS

No matter what the patient's age, most cases of pharyngitis will not be attributable to streptococcus. During the general physical examination, clinicians should consider performing a throat culture or rapid antigen test, but only *in tandem* with the Centor score. None of the univariate signs or symptoms associated with pharyngitis has high enough sensitivity and specificity for diagnosis according to clinical grounds alone. The greatest utility for the Centor score is in identifying patients for whom a throat culture or rapid streptococcal test should be performed because the score itself is not sufficient for confirming a diagnosis of streptococcal pharyngitis.

PRIOR PROBABILITY

The prevalence of streptococcal pharyngitis is higher in children than among infants and adults: group A β -hemolytic streptococcal bacteria can be isolated by throat culture in 24% to 36% of children and in 5% to 24% of adults with sore throat. Streptococcal pharyngitis is also more common in autumn and winter; thus, it may be appropriate to adjust the pretest probability upward during those seasons.

POPULATION FOR WHOM STREPTOCOCCAL PHARYNGITIS SHOULD BE CONSIDERED

- Children and adults with sore throat.

DETECTING THE LIKELIHOOD OF STREPTOCOCCAL PHARYNGITIS

The Centor score and modified Centor score perform differently for younger vs older patients (Table 47-7). The Centor score improves greatly when combined with rapid strep test results (Table 47-8).

Table 47-7 Likelihood Ratios for Centor Scores as a Function of Age

Score	LR (95% CI)
Adults >18 y	
Centor score 2-4	≈ 1
Centor score 0-1	0.26 (0.14-0.48)
Children 3-17 y	
Modified Centor score, 4-5	4.0 (2.7-6.0)
Modified Centor score, 2-3	0.69 (0.59-0.83)
Modified Centor score, 0-1	Uncertain

Abbreviations: CI, confidence interval; LR, likelihood ratio.

Table 47-8 Centor Score Combined With Rapid Strep Point-of-Care Test Results, Adults

Centor Score	Point-of-Care Test Result	LR (95% CI)
2-4 Findings	Positive	179 (110-2861)
0-1	Positive	26 (1.4-465)
0-4	Negative	0.09 (0.03-0.24)

Abbreviations: CI, confidence interval; LR, likelihood ratio.

REFERENCE STANDARD TESTS

Streptococcal throat culture, rapid streptococcal antigen tests.

REFERENCES FOR THE UPDATE

- Wong MCK, Chung CH. Group A streptococcal infection in patients presenting with a sore throat at an accident and emergency department: prospective observational study. *Hong Kong Med J*. 2002;8(2):92-98.
- Chazan B, Shaabi M, Bishara E, Colodner R, Raz R. Clinical predictors of streptococcal pharyngitis in adults. *Isr Med Assoc J*. 2003;5(6):413-415.^a
- Nawaz H, Smith DS, Mazhari R, Katz DL. Concordance of clinical findings and clinical judgment in the diagnosis of streptococcal pharyngitis. *Acad Emerg Med*. 2000;7(10):1104-1109.
- McIsaac WJ, Kellner JD, Aufricht P, Vanjaka A, Low DE. Empirical validation of guidelines for the management of pharyngitis in children and adults [published correction appears in *JAMA*. 2005;294:2700]. *JAMA*. 2004;291(13):1587-1595.^a
- Atlas SJ, McDermott SM, Mannone C, Barry MJ. The role of point of care testing for patients with acute pharyngitis. *J Gen Intern Med*. 2005;20(8):759-761.^a
- Cooper RJ, Hoffman JR, Bartlett JG, et al. Principles of appropriate antibiotic use for acute pharyngitis in adults: background. *Ann Intern Med*. 2001;134(6):509-517.

^aFor the Evidence to Support the Update for this topic, see <http://www.JAMAEvidence.com>.

This page intentionally left blank

EVIDENCE TO SUPPORT THE UPDATE:

47

Streptococcal Pharyngitis

TITLE The Role of Point of Care Testing for Patients With Acute Pharyngitis.

AUTHORS Atlas SJ, McDermott SM, Mannone C, Barry MJ.

CITATION *J Gen Intern Med.* 2005;20(8):759-761.

QUESTION Does point-of-care (POC) testing with a rapid test for group A β -hemolytic streptococcus improve the performance of the Centor score?

DESIGN Prospective, nonconsecutive study in which every patient who had POC testing also had a throat culture.

SETTING Two primary care practices with data collected during a 12-month period.

PATIENTS Adults (≥ 18 years) with symptoms of acute pharyngitis. Patients were excluded if their symptom duration was greater than 7 days, they had taken antibiotics within the past 24 hours, they were immunocompromised, or they had an acute pulmonary disease flare-up.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

The providers collected information for the Centor score and collected the samples for POC tests and throat cultures.

MAIN OUTCOME MEASURES

Sensitivity, specificity, and likelihood ratio (LR) of the Centor score and POC tests.

Centor score = history of fever (temperature $>38^\circ\text{C}$), tonsillar exudates, swollen anterior cervical lymph nodes, absence of cough (patient report).

A positive response to each finding is given 1 point, so a maximum Centor score is 4.

MAIN RESULTS

The authors completed data forms on 179 patients. They excluded 29 according to their criteria and had a final sample size of 148 after eliminating 2 patients with incomplete data.

Thirty-eight patients (26%) had group A β -hemolytic streptococcus by culture. The LRs for the Centor scores improved greatly when combined with the rapid strep test (Tables 47-9 and 47-10).

CONCLUSIONS

LEVEL OF EVIDENCE Level 3.

STRENGTHS Pragmatic study that took advantage of clinical decisions to perform POC testing, informed by the Centor score, vs the culture reference standard.

LIMITATIONS Nonconsecutive patients, so we do not know how many patients were evaluated for acute pharyngitis without POC testing.

Table 47-9 Likelihood Ratios for Centor Scores

Centor Score	n	LR+ (95% CI)
4	18	1.4
3	30	1.4
2	45	1.6
0-1 ^a	55	0.35
Collapsed data		
2-4 Findings		1.5 (1.2-1.9)
0-1		0.35 (0.16-0.75)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio.

^aOnly 1 patient had 0 findings.

Table 47-10 Likelihood Ratios for Rapid Strep Tests as a Function of Centor Scores

Centor Score	POC	N	LR+ (95% CI)
2-4	Positive	31	179
0-1	Positive	4	26
2-4	Negative	62	0.07
0-1	Negative	51	0.14
Collapsed data			
2-4 Findings	Positive		179 (110-2861)
0-1	Positive		26 (1.4-465)
0-4	Negative		0.09 (0.03-0.24)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; POC, point-of-care rapid strep test.

Because not all patients with acute pharyngitis were enrolled, it is possible that the providers selected patients for whom the diagnosis was not clear, making the Centor score perform with lower efficiency. However, the prevalence of streptococcal pharyngitis in this group is comparable to that in other studies. The Centor score did not perform well for identifying affected patients, but the presence of no more than 1 finding reduces the prior probability of 25% to 10%. The Centor score alone was not adequate for making a diagnosis.

The power of POC testing is highlighted by the results. A Centor score modulates the LR in that individuals with a score of 2 to 4 and a positive POC test result have an even higher LR than those with a score of 0 to 1. At the lower end of prior probability of group A β -hemolytic streptococcus by culture in adults with acute pharyngitis (10%), the probability of disease with a Centor score of 0 to 1 and a positive POC test result increased to 74%. A negative POC test result, at any level of Centor score, effectively ruled out group A β -hemolytic streptococcus by culture, with a posttest probability of less than 2%.

Reviewed by David L. Simel, MD, MHS

TITLE Clinical Predictors of Streptococcal Pharyngitis in Adults.

AUTHORS Chazan B, Shaabi M, Bishara E, Colodner R, Raz R.

CITATION *Isr Med Assoc J*. 2003;5(6):413-415.

QUESTION What are the clinical features that predict sore throat caused by group A β -hemolytic streptococcus?

DESIGN Consecutive patients with sore throat.

SETTING Israeli primary care clinics during 4 consecutive winter months.

PATIENTS Adults (>16 years) with sore throat.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

The clinical symptoms were recorded and a throat swab for culture was obtained. Rapid testing was not done.

MAIN OUTCOME MEASURES

A multivariate model was used to analyze the independent symptoms and signs. The outcome of Centor score vs the culture was reported.

Table 47-11 Likelihood Ratios for Centor Scores

Centor Score	Total (%)	LR+ (95% CI)
4	14 (5)	0.51 (0.12-2.2)
2-3	112 (55)	2.0 (1.6-2.4)
0-1	78 (40)	0.16 (0.06-0.43)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio.

MAIN RESULTS

A total of 207 patients were enrolled, with only 3 dropped because of missing data; 24% of patients had a positive throat culture result.

A multivariate analysis of 7 symptoms and 10 signs showed that only pharyngeal exudate was significantly predictive of a positive culture result (positive likelihood ratio [LR], 1.8; 95% confidence interval [CI], 1.5-2.2; negative LR, 0.27; 95% CI, 0.13-0.53). A low Centor score made strep throat much less likely (Table 47-11).

CONCLUSIONS

LEVEL OF EVIDENCE Level 2.

STRENGTHS Consecutive adults, including those who had a Centor score of 0 to 1. These patients are missing in most other validation studies.

LIMITATIONS The study had few patients who had all 4 findings of the Centor; thus, the results are unstable for this group.

This prospective study allows us to assess the usefulness of a Centor score of 0 to 1. Only 5% of patients with 1 or no symptoms had a positive culture result. Given that the probability of a positive culture result in this group was at the upper range of probabilities observed in adults, a Centor score of 0 to 1 decreases the probability of strep throat to less than 5% for most adults. The data support the recommendation for obtaining a rapid test for those with a score of 2 to 3, rather than treating empirically, because the LR was only 2. There were so few patients with a Centor score of 4 that the results are not useful for making conclusions about this group of patients.

The multivariate analysis selected exudates as the only useful sign or symptom. This finding suggests that the presence of exudates in the Centor score might be dominated by the results of this single finding when a population of all patients with sore throats is included, rather than just those with more than 1 Centor finding.

Reviewed by David L. Simel, MD, MHS

TITLE Empirical Validation of Guidelines for the Management of Pharyngitis in Children and Adults.

AUTHORS McIsaac WJ, Kellner JD, Aufricht P, Vanjaka A, Low DE.

CITATION JAMA. 2004;291(13):1587-1595.

QUESTIONS What is the likelihood of a group A β -hemolytic streptococcus culture according to a modified Centor score adjusted for patient age? Among a modified Centor score (adjusted for age) (Box 47-1), rapid tests, and the throat culture (the reference standard), which single or combined approach results in the most correct treatment decisions with the fewest rapid tests and cultures?

DESIGN The data were collected prospectively during a 3-year study period, and then the strategies were analyzed retrospectively.

SETTING Family practice clinic in Canada.

PATIENTS Patients with a chief complaint of sore throat, who ranged in age from 3 to 69 years. Patients were enrolled if they had a modified Centor score of 2 or greater and the physician or study nurse believed that a throat swab was necessary.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

Various treatment strategies that included the modified Centor score alone, rapid flu tests, or culture were assessed retrospectively to determine whether the strategy led to unnecessary tests or antibiotics. See Box 47-1.

MAIN OUTCOME MEASURES

The frequency of culture positivity as a function of the modified Centor score allowed us to calculate likelihood ratios (LRs) for the score in predicting culture positivity for group A β -hemolytic streptococcus. We transformed the sensitivity and specificity of each strategy to LRs and predictive values.

Box 47-1 Modified Centor Score (Range 0 to 5)

History of fever (temperature $>38^{\circ}\text{C}$), tonsillar exudates, swollen anterior cervical lymph nodes, absence of cough (patient report)

(A positive response to each finding is given 1 point, and then modified by age)

Age	Modification for Age
3-14 y	+1
15-44 y	0
≥ 45 y	-1

Because a culture is the reference standard test, we focused only on strategies that used initial combinations of the modified Centor Score or rapid tests, rather than strategies that went straight to culture.

MAIN RESULTS

A total of 918 patients were screened, with complete data available for 787 patients. Among the 333 adults, the prevalence of disease was 22%. The children had a prevalence of 34%. The modified Centor score performed differently, depending on the patient's age (Table 47-12). Treatment could be guided by the score (Table 47-13).

CONCLUSIONS

LEVEL OF EVIDENCE Level 3.

STRENGTHS This is a large study, conducted during a 3-year study period. A large distribution of patient ages helps us evaluate the generalizability of results.

Table 47-12 Likelihood Ratios of Modified Centor Scores as a Function of Age

Modified Centor Score	LR+ (95% CI)
Adults (≥ 18 y)	
4-5	1.6 (0.5-5.0)
3	1.3 (1.1-1.6)
2	0.53 (0.34-0.82)
Children (3-17 y)	
4-5	4.0 (2.7-6.0)
3	0.73 (0.61-0.88)
2	0.50 (0.31-0.80)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio.

Table 47-13 Management Strategy for Patients With a Modified Centor Score Greater Than or Equal to 2

Modified Centor Score	Antibiotic Treatment Strategy	Positive Predictive Value of Decision to Treat, % (95% CI)	Negative Predictive Value of Decision to Not Treat, % (95% CI)
Adults (≥ 18 y)			
4	Treat with antibiotics	84 (73-90)	94 (90-96)
2-3	Rapid test, treat if positive result		
Children (3-17 y)			
2-5 (All children with sore throat)	All get rapid test; treat for a positive rapid test result and culture for those with a negative rapid test result	98 (94-99)	100 (98-100)

Abbreviation: CI, confidence interval.

LIMITATIONS The study entrance criteria required that the physician or nurse determine that a throat swab was warranted. Although data were not given on the number of eligible patients who were not enrolled, clinicians should understand that these patients had a chief complaint of sore throat. The inference is that patients with sore throat who were more concerned about other symptoms (eg, fever or nasal congestion) were not enrolled. Furthermore, adults who had a sore throat but only 1 symptom would not have been included, because they had a modified Centor score of 0 to 1. The treatment strategies were not studied prospectively but instead were evaluated after the data were collected.

The modified Centor score (adjusted for age) did not work much better than the original Centor score for adults. For children, the LR of a modified Centor score of 4 to 5 increases the likelihood of group A streptococcus 4-fold, but current treatment recommendations require a rapid test for all children and cultures for those with negative results.^{1,2} This strategy leads to almost 100% accuracy for treatment decisions in children with sore throats.

For adults, the data apply only to patients with a modified Centor score of at least 2. Those patients with all 4 symptoms can be treated empirically with antibiotics, and those with a score of 2 to 3 can have treatment guided by a rapid test. With this strategy, about 16% of treated patients will not have group A β -hemolytic streptococcus, whereas 6% of patients with infection will not be treated. The only way to eliminate the 6% of patients who go untreated would be to use a strategy that required culture whenever the rapid strep test result is negative.

Reviewed by David L. Simel, MD, MHS

REFERENCES FOR THE EVIDENCE

1. Bisno AL, Gerber MA, Gwaltney JM Jr, Kaplan EL, Schwartz RH. Practice guidelines for the diagnosis and management of group A streptococcal pharyngitis. *Clin Infect Dis*. 2002;35(2):113-125.
2. Cooper RJ, Hoffman JR, Bartlett JG, et al. Principles of appropriate antibiotic use for acute pharyngitis in adults: background. *Ann Intern Med*. 2001;134(6):509-517.

CHAPTER 48

Is This Patient Having a Stroke?

Larry B. Goldstein, MD

David L. Simel, MD, MHS

CLINICAL SCENARIO

The wife of a 58-year-old right-handed man calls emergency medical services because her husband abruptly developed difficulty speaking and moving his right arm. Figure 48-1 presents the diagnostic flow of a patient who experiences neurologic symptoms that suggest a stroke.

WHY IS THE CLINICAL EXAMINATION OF PATIENTS WITH SUSPECTED STROKE IMPORTANT?

Since the original review of stroke published as part of The Rational Clinical Examination series more than a decade ago, much has changed.¹ What has not changed is the staggering cost of the personal, societal, and economic consequences of strokes. The estimated direct and indirect cost of stroke in 2005 is \$56.8 billion in the United States alone.² More than 700000 people in the United States have a stroke each year, of which nearly one-third represent recurrent events.³ About 163000 annual stroke deaths make it the third leading cause of death in the United States. Between 15% and 30% of stroke survivors become permanently disabled, whereas 20% remain in institutional care 3 months after their stroke. Not too long ago, the clinical examination functioned primarily to catalog a patient's neurologic impairments that in turn correlated with the stroke's vascular territory and likely cause. The inferences about the anatomy and etiology guided secondary preventive strategies and established the prognosis, rather than directing immediate treatment.

Despite the advent of modern noninvasive neuroimaging technologies, the clinical examination for stroke is more important than ever because therapeutic interventions for patients with acute stroke and sophisticated approaches to prevent recurrent strokes now exist. Appropriate treatment and prevention depend on accurate interpretation of the patient's symptoms and clinical examination findings. For example, the risk/benefit balance for carotid endarterectomy requires an accurate assessment of symptoms to identify those with a transient ischemic attack (TIA) or nondisabling stroke.⁴

The rapid screening of patients with neurologic symptoms begins with prehospital care personnel⁵ because the effectiveness of reperfusion strategies for acute ischemic stroke are time dependent. The brain can withstand profound ischemia for only limited periods, and the benefits of intravenous tissue plasminogen activator (tPA) lessens as the time from the onset of the patient's symptoms increases.⁶ Public education programs have stressed the need to call emergency medical responders (eg, 911) for persons experiencing stroke symptoms. Patients, family members, and prehospital care personnel such as emergency medical technicians must recognize the symptoms and signs of strokes to minimize treatment delays. Arrival to the hospital by emergency medical transport has been associated with more rapid treatment and

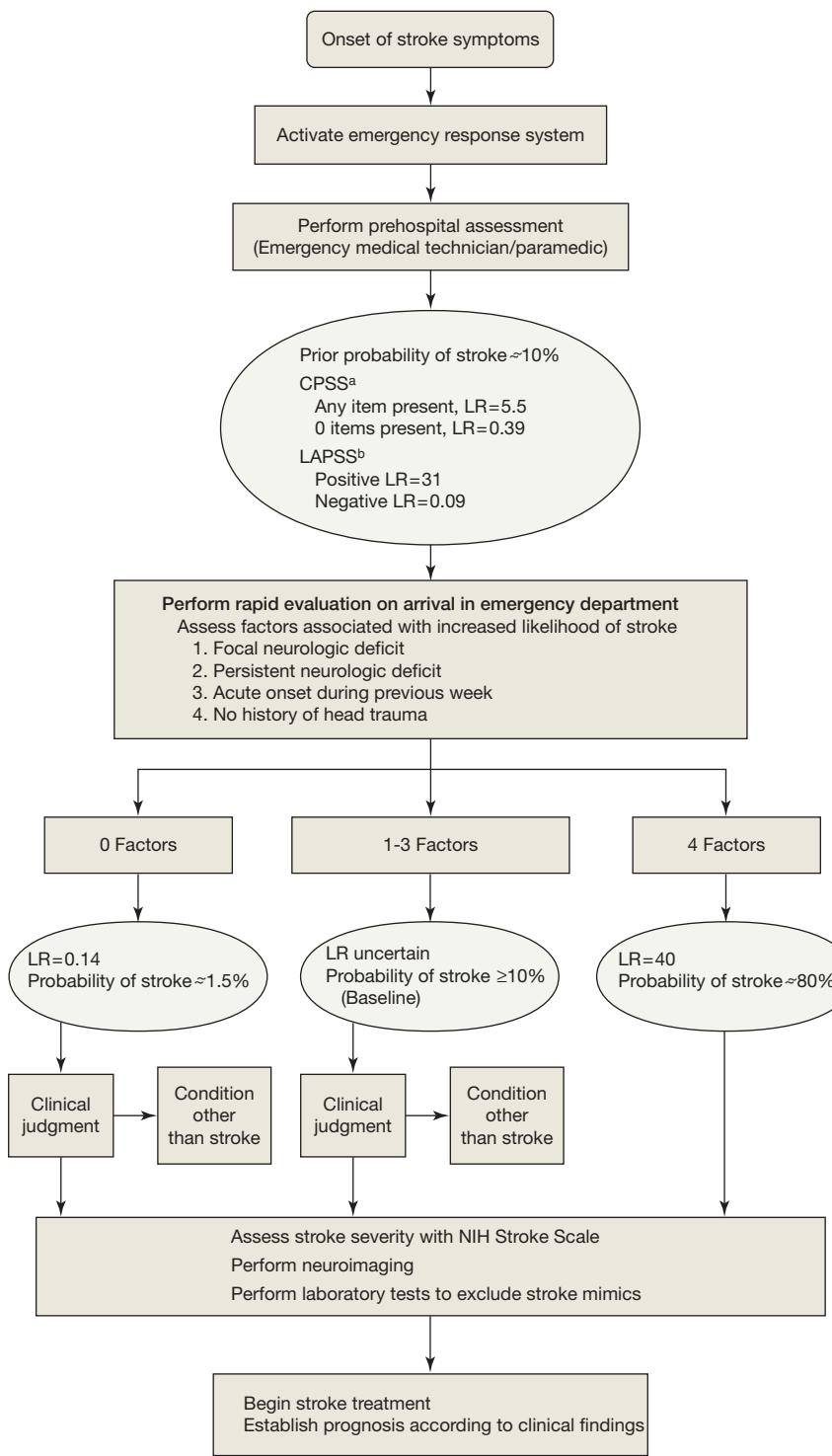


Figure 48-1 Diagnostic Flow of a Patient Who Experiences Neurologic Symptoms That Suggest a Stroke

^aCincinnati Prehospital Stroke Scale (CPSS); facial droop, arm drift, and abnormal speech. ^bLos Angeles Prehospital Stroke Scale (LAPSS); medical history (age >45 y, no history of seizures, symptoms < 24 h, not wheelchair bound), blood glucose 60–400 mg/dL (3.3–22 mmol/L), and examination showing unilateral facial weakness, grip weakness, and arm weakness. Abbreviations: LR, likelihood ratio; NIH, National Institutes of Health.

thereby presumably improved outcomes.^{7–10} Thus, the accuracy of the clinical examination becomes relevant not just for stroke specialists and emergency physicians but also for paramedics, nursing personnel, and emergency medical technicians who may be the first responders. When patients with stroke symptoms arrive at the hospital, a standardized neurologic examination, combined with neuroimaging results, determines subgroups of patients who might benefit from

intravenous thrombolysis vs those who may be at increased risk from thrombolytic-related bleeding.^{11–13}

Experienced examiners tailor the neurologic examination to address specific clinical questions because a stroke produces different symptoms and signs, depending on the area of affected brain. A variety of other conditions complicate diagnostic efforts by causing symptoms and signs similar to stroke (stroke mimics). In the patient example, emergency

medical services were called for a patient with new focal neurologic symptoms. We will observe the example patient through the emergency evaluation and highlight the clinical questions and features of the examination that increase the likelihood of accurately and reliably identifying a stroke, the stroke subtype, and the patient's prognosis.

METHODS

This review updates a 1994 report on clinical assessment of stroke¹ and is based on relevant studies identified through MEDLINE, restricted to the time since the last review. Information on the physical examination and neurologic examination is difficult to identify because the Medical Subject Headings for the articles typically do not include obvious terms. For example, searching the terms "cerebrovascular disorders" limited to human research studies, English-language articles (1994–2005) yields 9029 articles. However, when the results of this global search are crossed with the term "neurological examination," there are 176 articles, and when crossed with "physical examination," only 19 articles remain. Eliminating review articles and case reports from this reduced set left only 4 potentially relevant articles. Because of the low yield, we relied heavily on searches of the bibliographies of textbook chapters, review articles, and personal files to identify additional relevant literature for updating the role of the clinical examination since the original Rational Clinical Examination article on stroke in 1994.

To examine the accuracy and reliability of the clinical assessment of stroke for either diagnosis or prognosis, the following

general inclusion criteria were used in assessing articles: (1) the article addressed the issue of accuracy or reliability of medical history or physical examination for diagnosis or estimation of short-term prognosis (mortality or functional disability); (2) the study site or participants (clinicians or patients) were described; (3) the data were not limited to case reports or reviews of other studies; and (4) the primary data or appropriate summary statistics were presented.

For assessment of the accuracy of diagnosis, references included articles that also described a final diagnosis established by an expert who reviewed all clinical data, neuroimaging, and other relevant laboratory tests. These articles were evaluated for quality according to whether the clinical examination was performed masked to the neuroimaging results (see Table 1-7).¹⁴ Articles describing prognosis in terms of functional status were included if the outcome was measured with a scale that is either comparable to a scale in common use or was validated in the context of the study.

The sensitivity (how often a diagnostic procedure detects a condition when it is present), specificity (how often a diagnostic procedure result is negative when the condition is absent), and likelihood ratios (LRs) (the odds favoring the diagnosis or outcome vs not having the diagnosis) for each finding or scale were recorded from each article or were calculated according to primary data as necessary.^{15,16} Table 48-1 summarizes the included studies that gave sensitivity and specificity data for the diagnosis of stroke or TIA. For studies of precision, the κ statistic (describes the agreement between paired observers beyond that predicted by chance) or the intraclass correlation coefficient (when there are more than 2 examiners) is given. Intraclass coefficients range from 0 to 1,

Table 48-1 Summary of Included Studies With Sensitivity/Specificity Data

Source, y	Level of Evidence ^a	Country	Setting	No. of Participants	Inclusion Criteria
Kothari et al, ¹⁷ 1997	2	United States	ED	299	Clinical trial and ED patients
Kothari et al, ¹⁸ 1999	3	United States	ED and neurology service	171	Suspected stroke or stroke mimic
Kidwell et al, ¹⁹ 2000	1	United States	Field and ED	441	Suspected stroke
Karanja et al, ²⁰ 1997	2	United States	Neurology clinics	381	Stroke, TIA, or other neurologic condition
von Arbin et al, ²¹ 1980	3	Sweden	Hospital	2252	Medical admissions
von Arbin et al, ²² 1981	3	Sweden	Stroke unit	206	Stroke unit admission
Panzer et al, ²³ 1985	2	United States	Hospital	369	Suspected stroke
Oxbury et al, ²⁴ 1975	3	United Kingdom	Hospital	93	Stroke
Tuthill et al, ²⁵ 1969	3	United States	Stroke unit/community hospital	202	Suspected stroke
Frithz and Werner, ²⁶ 1976	3	Sweden	Hospital	344	Stroke, <70 y
Allen, ²⁷ 1984	3	United Kingdom	Hospital	148	Stroke, <76 y
Henley et al, ²⁸ 1988	2	United Kingdom	Hospital	172	Stroke
Fullerton et al, ²⁹ 1988	3	Ireland	Hospital	206	Acute stroke
Britton et al, ³⁰ 1980	2	Sweden	Stroke unit	200	Suspected stroke

Abbreviations: ED, emergency department; TIA, transient ischemic attack.

^aSee Table 1-7 for a description of Evidence Grades and Levels.

Table 48-2 The National Institutes of Health Stroke Scale^a

Item	Response ^b
1a. Level of consciousness	0 = Alert 1 = Not alert 2 = Obtunded 3 = Unresponsive
1b. Level of consciousness questions	0 = Answers both correctly 1 = Answers 1 correctly 2 = Answers neither correctly
1c. Level of consciousness commands	0 = Performs both tasks correctly 1 = Performs 1 task correctly 2 = Performs neither task
2. Gaze	0 = Normal 1 = Partial gaze palsy 2 = Total gaze palsy
3. Visual fields	0 = No visual loss 1 = Partial hemianopsia 2 = Complete hemianopsia 3 = Bilateral hemianopsia
4. Facial palsy	0 = Normal 1 = Minor paralysis 2 = Partial paralysis 3 = Complete paralysis
5. Motor arm	0 = No drift a. Left 1 = Drift before 5 s b. Right 2 = Falls before 10 s 3 = No effort against gravity 4 = No movement
6. Motor leg	0 = No drift a. Left 1 = Drift before 5 s b. Right 2 = Falls before 5 s 3 = No effort against gravity 4 = No movement
7. Ataxia	0 = Absent 1 = One limb 2 = Two limbs
8. Sensory	0 = Normal 1 = Mild loss 2 = Severe loss
9. Language	0 = Normal 1 = Mild aphasia 2 = Severe aphasia 3 = Mute or global aphasia
10. Dysarthria	0 = Normal 1 = Mild 2 = Severe
11. Extinction/inattention	0 = Normal 1 = Mild 2 = Severe

^aThe actual form for recording the data contains detailed instructions for the use of the scale. Available at http://www.ninds.nih.gov/doctors/NIH_Stroke_Scale.pdf (accessed June 13, 2008). An online course for provider education is available at http://www.ninds.nih.gov/doctors/stroke_scale_training.htm (accessed June 13, 2008).

^bScore = sum of scores from each item.

with 0 indicating random agreement and 1 indicating perfect agreement. Random-effects estimates were used for the LR summary measures.

RESULTS

Prehospital Assessment

Accuracy

According to a prospective observational cohort study, when examination was performed by a physician, the presence of any of 3 physical examination findings (facial paresis, arm drift, and abnormal speech) was selected from the National Institutes of Health Stroke Scale (NIHSS) as the most useful. These 3 items, selected by statistical recursive partitioning techniques, identified patients with stroke with 100% sensitivity (lower 95% confidence limit, 95%) and 88% specificity (95% confidence interval [CI], 82%-91%) (positive LR [LR+], 7.9; 95% CI, 5.6-11; negative LR [LR-], 0; 95% CI, 0-0.12), although the sensitivity decreased to 66%, with a similar specificity when this instrument was validated in the hospital setting.¹⁷ Several schemes facilitate the rapid, accurate identification of stroke patients by emergency medical personnel.

The Cincinnati Prehospital Stroke Scale (CPSS) uses the 3 most important items (facial paresis, arm drift, and abnormal speech) derived from the NIHSS (Table 48-2).¹⁷ In a prospective study, one of 2 emergency physicians certified in the use of the full NIHSS evaluated 171 patients (selected by a neurologist from either the emergency department or inpatient neurology service) with chief symptoms that suggested a stroke.¹⁸ The examining physicians were aware of the patient's chief report but not the presenting clinical signs or final diagnosis. Each patient also had separate examinations by 4 of 24 emergency medical personnel, masked to all the clinical data. According to data provided in the article, we calculated the LRs for increasing numbers of findings (0-3) for the physicians (Table 48-3). The same calculations can be done for the emergency medicine personnel, although the CIs are overstated because the findings are presented for the total number of examinations rather than unique patients. Nonetheless, the diagnostic accuracy for the emergency department physician compared with the emergency medical personnel was identical, with the area under each receiver operating characteristic = 0.88. The presence of any single finding of the 3 created a sharp increase in the likelihood of stroke. After collapsing the data at a threshold of greater than or equal to 1 finding vs 0 findings, the physician had an LR of greater than or equal to 1 finding = 5.5 (95% CI, 3.3-9.1) and an LR of 0 findings = 0.39 (95% CI, 0.25-0.61); the emergency medical personnel had an LR of greater than or equal to 1 finding = 5.4 (95% CI, 4.1-7.0) and an LR of 0 findings = 0.46 (95% CI, 0.38-0.56). Although this study did not evaluate the emergency medical personnel's diagnostic accuracy according to examinations performed in the field, this method of identifying patients with acute stroke is being widely used throughout the country and can be performed in less than a minute.

The Los Angeles Prehospital Stroke Screen (LAPSS) assesses for a unilateral arm drift, handgrip strength, and facial paresis.¹⁹

The screen was evaluated prospectively on all noncomatose, nontrauma patients with neurologic complaints compatible with stroke, who were transported by emergency medical technicians to a single hospital. The relevant neurologic signs were altered consciousness, focal neurologic signs, seizure, syncope, head pain, or a cluster category of weakness/dizziness/sick. The criteria for an in-the-field stroke diagnosis by the emergency medical technician were met when the patients were older than 45 years, had no seizure history, had symptoms for fewer than 24 hours, were not wheelchair bound or bedridden, had a blood glucose level between 60 and 400 mg/dL (3.3 and 22 mmol/L), and a unilateral deficit in one of the 3 findings previously listed. A reviewer, masked to the emergency medical personnel's evaluation, determined the final discharge diagnosis according to the emergency department chart. Compared with the final diagnosis, the LAPSS had a sensitivity of 91% (95% CI, 76%-98%), specificity of 97% (95% CI, 93%-99%), LR+ of 31 (95% CI, 13-75), and LR- of 0.09 (95% CI 0.03-0.27) for patients with possible stroke (Table 48-4). An analysis that included all ambulance runs showed even better specificity (and therefore a much higher LR+), with only a slight decrement in sensitivity, attributed to 2 stroke patients who were not correctly identified in the field as having a possible stroke, of 1092 total ambulance runs (0.19%). Among all patients with neurologically relevant signs, the prevalence of stroke was 10%, which represents a useful anchor for prior probability estimates (Figure 48-1).

Reliability

The data assessing the CPSS compare emergency medical personnel with physicians for examinations performed in a controlled hospital setting rather than in the field.¹⁸ The intraclass correlation coefficient (Pearson r) for the total score was 0.89 (95% CI, 0.87-0.92) among the prehospital care personnel and 0.92 (95% CI, 0.89-0.93) between the physician and the pre-hospital personnel. The greatest agreement was for arm drift (Pearson $r = 0.91$; 95% CI, 0.89-0.93), followed by abnormal speech (Pearson $r = 0.87$; 95% CI, 0.34-0.90) and facial palsy (Pearson $r = 0.78$; 95% CI, 0.74-0.83).

Scenario

With either the CPSS or the LAPSS, the patient would have been identified as likely to have had a stroke, triggering rapid transport to the nearest appropriate emergency department for further evaluation and treatment. Physicians should feel confident with the medical history and brief screening examination for stroke that is obtained by appropriately trained emergency first responders.

In the case of this patient scenario, the patient arrives at the emergency department and his wife reports that her husband has hypertension. He has no history of diabetes, seizures, or recent head trauma. He is being treated with aspirin and a diuretic. He continues to have difficulty moving his right arm, along with trouble speaking.

Is This Patient Having a Transient Ischemic Attack or Stroke?

In the LAPSS study previously discussed, only 8% of 441 patients transported to the hospital for nontraumatic, non-

Table 48-3 Comparison of Physician Assessment With That of Emergency Medicine Personnel^{a,b}

No. of Findings Present	Stroke	Nonstroke Diagnosis	LR (95% CI)
Physician Assessment ^c			
3	4	1	14 (1.6-121)
2	6	5	4.2 (1.4-13)
1	15	10	5.2 (2.6-11)
0	13	117	0.39 (0.25-0.61)
Emergency Medical Personnel ^d			
3	20	10	7.0 (3.3-14)
2	22	10	7.6 (3.7-16)
1	49	39	4.4 (3.0-6.4)
0	63	476	0.46 (0.38-0.56)

Abbreviations: CI, confidence interval; LR, likelihood ratio.

^aBased on data from Kothari et al.¹⁸

^bCollapsing data into a 2 × 2 table yields an LR of ≥1 finding = 5.5; 95% CI, 3.3-9.1.

^cData represent unique patients and stratum-specific LR.

^dData represent 4 examinations for each patient.

Table 48-4 Performance of Emergency Medicine Technicians on Stroke Assessment in the Field^a

	Stroke Frequency/All Patients	LR+ (95% CI)	LR- (95% CI)
Ambulance runs for patients with target symptoms for possible stroke ^b	34/206	31 (13-75)	0.09 (0.03-0.27)
All patients transported by ambulance	36/1298	217 (90-526)	0.14 (0.06-0.31)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

^aBased on data from Kidwell et al.¹⁹

^bNeurologic signs were altered consciousness, focal neurologic signs, seizure, syncope, head pain, and a cluster category of weakness/dizziness/sick.

comatose, neurologically relevant complaints had a final diagnosis of acute symptomatic cerebrovascular disease.¹⁹ A variety of conditions can mimic TIA or stroke. Seizures,^{31,32} neoplasms,³³ infection,³⁴ intracranial hemorrhage,³⁵ and hypoglycemia³⁶ and other metabolic abnormalities are among the conditions that can simulate a TIA and stroke. In another series, among 821 consecutive patients initially diagnosed with stroke, 13% were finally determined to have other conditions.³⁷ The most frequent causes of misdiagnosis were unrecognized seizures, confusional states, syncope, toxins, neoplasms, and subdural hematomas.

Transient Ischemic Attack

Transient ischemic attack is traditionally defined as a focal neurologic deficit of ischemic origin of less than 24 hours' duration.³⁸ Because most TIAs last fewer than 4 hours, the diagnosis is usually based on medical history rather than findings on

examination.³⁹ However, many patients previously diagnosed with TIA actually had cerebral infarcts demonstrated on magnetic resonance imaging (MRI).⁴⁰ Clinically silent infarcts (and potentially infarcts associated with a classically defined TIA)

Table 48-5 Estimates of the Accuracy of Classification of Stroke Type Based Solely on Clinical Data^{a,b}

Diagnosis	References	LR+ (95% CI)	LR- (95% CI)
Stroke vs not stroke ^c	21	40 (29-55)	0.14 (0.10-0.20)
TIA vs not TIA ^d	22	21 (10-42)	0.09 (0.02-0.34)
Hemorrhagic vs non-hemorrhagic stroke ^e	22, 23, 43	3.1 (2.1-4.6)	0.61 (0.48-0.76)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio; TIA, transient ischemic attack.

^aBased on data from Goldstein and Matchar.¹

^bOnly studies for which sensitivity, specificity, and LRs could be calculated are represented in the table.

^cPersistent neurologic deficit of acute onset during the previous week, without a history of head trauma, according to medical history and examination alone.

^dFocal neurologic deficit with a duration of less than 24 hours, according to medical history and examination alone.

^eEstimates for hemorrhagic vs nonhemorrhagic stroke are summary estimates from random-effects measures.

Table 48-6 Precision of Elements of the Neurologic Examination of Stroke Patients

Finding	κ Score or Range ^a	References ^b
Medical History		
Seizure at onset	0.39	45
Previous stroke	0.31	45
Transient ischemic attack	0.11	45
Vomiting at onset	0.35	45
Headache	0.36	45
Examination		
Level of consciousness	0.38-1.00	45-51
Orientation	0.19-1.00	45-47, 51-53
Gaze preference	0.33-1.00	45, 46, 48-50, 53
Visual field defect	0.16-0.81	45, 46, 48, 50, 53, 54
Facial paresis	0.13-1.00	45, 46, 50, 52, 53, 54
Arm strength	0.42-1.00	45-54
Leg strength	0.40-0.84	45-54
Limb ataxia	-0.16-0.69	46, 48, 50, 51
Sensation	0.27-0.89	45, 46, 48, 50, 51, 53
Language	0.54-0.84	45-48, 51, 53
Dysarthria	0.29-1.00	45-48, 51
Neglect	0.58-0.89	46, 48, 50, 51
Pupillary response	0.95	48
Plantar response	0.67	48
Gait	0.91	49

^aThe values of the κ statistic may be interpreted similar to the interpretation of correlation coefficients ($\kappa = 0.20$, slight; $0.21-0.40$, fair; $0.41-0.60$, moderate; $0.61-0.80$, substantial; $0.81-1.00$, almost perfect agreement⁵⁵).

^bAmong the cited studies, individual items were measured by different observers with various experience.

may contribute to vascular dementia.⁴¹ Traditionally defined TIA is an important marker of short- and long-term vascular risk. Of 1707 patients from a large health care plan in the United States, evaluated in the emergency department and diagnosed with TIA, 5.3% had a stroke within 2 days, whereas 10.5% had a stroke within 90 days.⁴² The diagnosis of a stroke or TIA indicates the need for urgent management.

Accuracy of a Transient Ischemic Attack Diagnosis

Among patients admitted to a stroke unit for evaluation of an acute neurologic deficit, a clinical diagnosis of TIA increased the odds of a final TIA diagnosis by about 20-fold (Table 48-5) (LR+, 21; 95% CI, 10-42), whereas an alternate diagnosis greatly decreased the odds of a TIA (LR, 0.09; 95% CI, 0.02-0.34). The excellent performance of the clinical examination in this filtered population with a high probability of stroke probably does not extrapolate to the emergency setting in which patients with neurologically relevant complaints have a broader differential diagnosis. In another study, about one-third of patients initially diagnosed with TIA were eventually given a different diagnosis, with TIA being definitely not established in an additional one-third.⁴⁴

Reliability of a Transient Ischemic Attack Diagnosis

Despite its clinical importance, the reliability of the diagnosis of TIA can be poor. Agreement among experienced physicians for a patient's history of TIA is barely greater than chance ($\kappa = 0.11$; see the footnote to Table 48-6 for a guide to interpreting κ scores).⁴⁵ Some of the imprecision is due to differences in categorizing patients as having minor stroke or TIA, a distinction that has little influence on patient management. Even with a standardized protocol, disagreements frequently occur with regard to the features of the TIA. In one study, medical histories were obtained from 28 patients by pairs of neurologists.⁵⁶ Agreement in the number of TIAs was observed in about half of the cases. In two-thirds of the cases, there was agreement in the time of onset for the first TIA and the duration of the episode; there was agreement less than half the time in the frequency and type of symptoms. A new definition of TIA shortens the duration for qualifying episodes: "a brief episode of neurological dysfunction caused by a focal disturbance of brain or retinal ischemia, with clinical symptoms typically lasting less than 1 hour, and without [radiographic] evidence of infarction."⁵⁷ The accuracy of symptoms, signs, and the overall clinical impression using this new definition has not been studied.

There is fair agreement when minor stroke and TIA are considered together as a previous ischemic episode ($\kappa = 0.60$).⁴⁵ Other studies suggest that substantial diagnostic agreement can be achieved when a standardized protocol is used for the 2 diagnoses ($\kappa = 0.65$ ⁵⁸ and 0.77 ⁵⁹). The Asymptomatic Carotid Atherosclerosis Study compared an algorithm for the diagnoses to both an on-site neurologist's diagnosis and that of an external panel of reviewers with expertise in stroke (ie, the gold standard).²⁰ The key symptoms were sudden change in speech, visual loss, diplopia, numbness or tingling, paralysis or weakness, and nonorthostatic dizziness. Comparing stroke or TIA vs no vascular event, there was 80% agreement

between the external panel and the algorithm ($\kappa = 0.60$; 95% CI, 0.52-0.68).

Stroke

The operational definition of stroke requires relevant, focal neurologic symptoms with no other potential etiologies. Guideline statements from several professional societies recommend excluding systemic or other neurologic processes that might cause the patient's acute deficit as part of the evaluation of the appropriateness of administering acute thrombolytic therapy.⁶⁰⁻⁶²

Accuracy of a Stroke Diagnosis

The results of studies on the accuracy of stroke diagnosis are given in Table 48-5. In one study, patients with the presence of 4 findings were considered to have had a stroke if their medical history included a persistent, focal neurologic deficit of acute onset during the previous week but no history of head trauma.²¹ This study, done before modern neuroimaging, relied on autopsy or stroke unit evaluations to establish the diagnosis in 39% of 2034 patients and consensus agreement for the remaining patients. Using this rule, emergency department physicians correctly identified 152 of 176 consecutive patients with stroke (sensitivity, 86%; 95% CI, 81%-91%) and 1818 of 1858 patients without stroke (specificity, 98%; 95% CI, 97%-99%). Thus, the odds of having a diagnosis of stroke increase dramatically when patients satisfy this classification rule are LR+ of 4 findings is 40 (95% CI, 29-55); LR- is 0.14 (95% CI, 0.10-0.20). Although this LR- is low, neuroimaging studies may still be required to help diagnose conditions that mimic stroke. Differences in the accuracy of the diagnosis of stroke according to either the interval between the onset of symptoms and time of presentation or the likelihood that the patients' symptoms and signs could be assigned to a specific vascular territory were not addressed. Data concerning the accuracy of the diagnosis for patients evaluated soon after the beginning of symptoms were lacking in this study, but the accuracy is particularly relevant, given the advent of reperfusion therapies such as intravenous tPA that necessitate treatment within 3 hours of symptom onset.⁶⁰⁻⁶²

Reliability of a Stroke Diagnosis

High-quality studies of the reliability of the diagnosis of stroke are lacking.

Scenario

The prior probability of a stroke among patients with neurologically relevant symptoms is 10%. According to the patient's focal neurologic symptoms, the LAPSS study¹⁹ suggests that the LR for stroke is 31. According to the clinical information obtained in the field and before the complete emergency department evaluation, the posterior probability for stroke is 78% (from posterior odds = $[0.1/0.9] \times 31 = 3.4$). The emergency physician's confirmation that the patient had an abrupt onset of focal neurologic symptoms and no known conditions that would increase the chances of a stroke mimic increases the likelihood of a stroke.

In the case of this patient scenario, the patient's blood pressure reading is 150/95 mm Hg. His pulse rate is 84/min and

regular. He is alert, knows his age and the current month, and is able to follow simple verbal commands (NIHSS item 1a, 1b, 1c; Table 48-2). He has dysarthric speech (NIHSS item 10) and had difficulty naming common objects (ie, dysnomia; NIHSS item 9), but his speech is understandable. At rest, the patient tends to look only to the left, but on command he is able to look to the right (ie, a left gaze preference; NIHSS item 2). On asking him to identify your fingers at the periphery of his visual fields, you discover that he sees nothing to his right (a right homonymous visual field defect; NIHSS item 3). The right side of his face droops (a right lower facial paresis; NIHSS item 4), and when he holds his arms straight out with the palms facing up, his right arm drifts downward (a right-sided drift; NIHSS item 5). His right leg is slightly weak to motor testing but does drift by a count of 5 (NIHSS item 6). He has no limb ataxia (the smoothness of movements is consistent with the amount of limb weakness; NIHSS item 7) but has diminished pain sensibility in his right arm (pinprick is described as feeling dull in his right arm compared with his left; NIHSS item 8). There is no evidence of spatial neglect (he is able to recognize being touched on his right arm and leg when touched on the right and left sides simultaneously; NIHSS item 11). A glucose level obtained by fingerstick was 110 mg/dL (6.1 mmol/L).

What Is the Vascular Distribution of the Stroke?

Accuracy of Determining the Stroke Distribution

Historical and objective data help localize the affected portions of the nervous system, providing clues about the likely pathophysiology and etiology (essential for rational secondary prevention).⁶¹ Clinicians must recognize that computed tomographic (CT) scan results are frequently negative during the first hours after ischemic stroke and technical limitations often impair CT imaging of posterior fossa structures. These limitations in early neuroimaging of the evolving stroke serve to emphasize the importance of the clinical examination. MRI scans, with greater sensitivity than CT, are often not available for immediate, routine patient evaluations.⁶³

Reliability of Determining the Stroke Distribution

The clinical examination is most important despite its less than perfect accuracy early in the course of a stroke episode, when the initial imaging studies may not reveal the abnormality. An understanding of the reliability of the examination helps identify the clinical features that have potential utility. Clinical experience suggests that the reliability of individual elements of the neurologic history and examination is important for the description of the stroke patient's neurologic deficits (Table 48-6).^{45-52,54,64} Obtaining historical data from stroke patients can be hampered because of the communication deficits caused by the stroke. Only 1 of these studies assessed the reliability of historical data.⁴⁵

The reliability of historical items is generally low, ranging from slight to fair agreement between observers,⁴⁵ which is particularly noteworthy because so much of diagnosis, particularly of transient events such as TIAs, depends on the patient's medical history. The reliability of specific neurologic examination findings improves when the examination

is performed with knowledge of the patient's medical history and when a full examination is performed in contrast to an examination aimed at a particular finding.⁶⁵ Several specific findings are assessed, with high degrees of reliability (Table 48-6). However, in practice, anatomic diagnosis for neurologic conditions requires recognition of the pattern of abnormal and normal findings, rather than a single finding.

Experienced physicians consider their own views of the reliability of given findings (ie, subjective sensory abnormalities tend to be unreliable) when arriving at a specific anatomic diagnosis. Although neuroanatomic diagnosis can be complex, schemes have been developed that can be generally applied. For example, the Oxfordshire classification (used primarily in research settings) assigns one of the 4 anatomic distributions (Box 48-1).⁶⁶ When caused by ischemia, the total anterior circulation infarction syndrome (TACS) reflects proximal occlusion of the internal carotid artery or trunk of the middle cerebral artery; the partial anterior circulation infarction syndrome suggests a branch artery occlusion in the middle cerebral artery distribution; a lacunar infarction syndrome indicates occlusion of a small penetrating vessel; and posterior circulation infarction syndrome is consistent with vertebrobasilar distribution stroke. The reliability of this classification is moderate to good ($\kappa = 0.54$; 95% CI, 0.39–0.68).⁶⁷

Box 48-1 Oxfordshire Classification of Subtypes of Cerebral Infarction^a

TOTAL ANTERIOR CIRCULATION INFARCTION SYNDROME (TACS)

A combination of new higher cerebral dysfunction (ie, dysphasia, dyscalculia, visuospatial disorder); homonymous visual field defect; and ipsilateral motor or sensory deficit of at least 2 areas of the face, arm, and leg.

PARTIAL ANTERIOR CIRCULATION INFARCTION SYNDROME (PACS)

Only 2 of the 3 components of the TACS syndrome are present, with higher cerebral dysfunction alone or with a motor/sensory deficit more restricted than those classified as LACS (ie, confined to 1 limb or to face and hand, but not to the whole arm).

LACUNAR INFARCTION SYNDROME (LACS)

Pure motor stroke, pure sensory stroke, sensorimotor stroke, or ataxic hemiparesis.

POSTERIOR CIRCULATION INFARCTION SYNDROME (POCS)

Any of the following: ipsilateral cranial nerve palsy with contralateral motor or sensory deficit; bilateral motor or sensory deficit; disorder of conjugate eye movement; cerebellar dysfunction without ipsilateral long-tract deficit (ie, ataxic hemiparesis); or isolated homonymous visual field defect.

^aBased on data from Bamford et al.⁶⁶

Scenario

This right-handed patient with unilateral right facial and limb weakness might have a lesion affecting contralateral central motor pathways at any level of the neuraxis above the midpons. However, when these findings are combined with an aphasia (manifest as a dysnomia), the patient's deficit is likely the result of a lesion affecting the dominant hemisphere. The greater involvement of face and arm as compared with his leg suggests an abnormality extending from the region of the Sylvian fissure toward the convexity, consistent with ischemia in the distribution of the left middle cerebral artery.

In this patient scenario, the examination result is consistent with a left middle cerebral artery distribution cerebral infarction, fulfilling criteria for TACS. However, a neuroimaging study is necessary to help exclude a stroke mimic and to determine whether the patient may have had a brain hemorrhage. You request a brain CT scan.

Assigning Stroke Severity

According to the information provided in Table 48-6, it is apparent that the reliability of specific items varies widely. During the course of care, and to guide prognosis, standardized assessments of a stroke patient's deficits improve the reliability of the routine neurologic examination. Examples with supportive reliability and validity data include the Canadian Neurological Scale,⁵² the Copenhagen Stroke Scale,⁵⁴ the Scandinavian Neurological Stroke Scale,⁴⁹ the Unified Neurological Stroke Scale,⁶⁸ and the NIHSS.⁴⁸ Of these, the NIHSS has been widely adopted for clinical care and research in the United States and other countries (Table 48-2). The scale and instructions are available as an online resource (http://www.ninds.nih.gov/doctors/stroke_scale_training.htm; accessed June 13, 2008). The reliability of the scale's individual items has been studied extensively (Table 48-7); data from some of these studies are included in the ranges given in Table 48-6. With the highest values within each range, most items can have substantial to almost perfect levels of agreement. With the lowest values, reliability can be as low as slight to moderate.

Recognition of the potential for limited reliability of some items has led to the development of a free online training and certification program sponsored by the American Stroke Association, in conjunction with the American Academy of Neurology and the National Institute of Neurological Disorders and Stroke (<http://nihss-english.trainingcampus.net/uas/modules/trees/windex.aspx>; accessed June 13, 2008). With training, the NIHSS can be used reliably by nonneurologist physicians, as well as nurses.^{50,69} The NIHSS can also be scored with high reliability by remote observers via telemedicine (correlation between bedside and remote scores, $r = 0.955$; $P < .001$).⁷⁰

The NIHSS scores correlate well with the size of the stroke as measured by MRI.⁷¹ Therapeutically, a secondary analysis of the National Institutes of Health (NIH) tPA trial data found that the risk of intracerebral hemorrhage was independently associated with baseline stroke severity as assessed

with the NIHSS, divided into 5 categories (0-5, 6-10, 11-15, 16-20, and >20; odds ratio [OR], 1.8; 95% CI, 1.2-2.9).¹² After tPA treatment, 17% of patients with a baseline NIHSS score greater than 20 developed an intracerebral hemorrhage vs 3% to 5% with less severe strokes. Overall, those in the most severe category had the overall worst prognosis for recovery by 3 months, yet they were also the most likely to improve with tPA (OR, 4.3; 95% CI, 1.6-12). This information, derived from clinical observations, is helpful when discussing the risks and benefits of the treatment with patients and families.

Scenario

In this case, the example patient had an NIHSS score of 9 (item 2 = 1, item 3 = 2, item 4 = 2, item 5 = 1, item 8 = 1, item 9 = 1, item 10 = 1; Table 48-2).

He has a 22% risk of death or a poor outcome without reperfusion therapy. You need to determine whether he had a hemorrhagic or ischemic stroke to assess the appropriateness of thrombolysis.

Classifying the Stroke

Accuracy of Stroke Classification

It is not enough to determine whether the patient with an acute focal neurologic deficit has had a stroke. Treatment with a thrombolytic or an antithrombotic drug is contraindicated in patients with hemorrhage. Three studies that provide information about the accuracy of medical history and physical examination in distinguishing hemorrhagic from ischemic strokes indicate that clinical judgment can be used to increase or decrease the likelihood of hemorrhage, but diagnostic errors occur (Table 48-5). In one study, a multivariate model showed that initial depressed level of consciousness, vomiting, severe headache, warfarin therapy, systolic blood pressure above 220 mm Hg, and glucose level above 170 mg/dL (9.4 mmol/L) in a patient without diabetes increased the likelihood of hemorrhagic stroke.²³ The presence of any of these features more than doubles the odds of hemorrhage (LR+, 2.4; 95% CI, 1.8-3.2) and the absence of any of these features decreases the odds by one-third (LR-, 0.35; 95% CI, 0.18-0.68). The other 2 studies described the accuracy of the physician's overall assessment without the use of a predictive model and produced results that performed similarly to those of the multivariate model (the results were statistically homogenous for the diagnostic OR; $P = .99$). Thus, the clinical judgment that a stroke is hemorrhagic has an LR = 3.1 (95% CI, 2.1-4.6), whereas the assessment that the stroke is not hemorrhagic decreases the likelihood (LR, 0.61; 95% CI, 0.48-0.76). The use of a complex discriminant score (based on specific historical and objective physical factors) modestly improves accuracy relative to clinician judgment but is cumbersome and not clinically useful.⁷² A neuroimaging study is mandatory before the patient is given a thrombolytic agent or anticoagulant.⁶⁰⁻⁶²

Reliability of Stroke Classification

Examining neurologists show only slight agreement on classifying a stroke as due to an infarct vs a hemorrhagic stroke ($\kappa = 0.38$).⁷³

Table 48-7 Reliability of National Institutes of Health Stroke Scale Items^a

Item	κ Range
1a. Level of consciousness	0.46 to 0.68
1b. LOC questions	0.44 to 0.94
1c. LOC commands	0.41 to 0.94
2. Gaze	0.33 to 0.82
3. Visual fields	0.57 to 0.90
4. Facial palsy	0.22 to 0.74
5. Arm strength	0.77 to 0.97
6. Leg strength	0.39 to 0.98
7. Limb ataxia	-0.16 to 0.69
8. Sensation	0.39 to 0.89
9. Language	0.60 to 0.84
10. Dysarthria	0.29 to 0.72
11. Extinction/neglect	0.53 to 0.89

Abbreviation: LOC, level of consciousness.

^aBased on published data.^{35,48,50,51}

Scenario

The patient was alert, not nauseated, did not have a headache, and was not receiving warfarin. His blood pressure was not severely increased, and his blood glucose level was normal. The chance of an intracerebral hemorrhage is low but cannot be excluded without a neuroimaging study.

In this case, recognizing that the neuroimaging results may be inconclusive, you must consider whether there might be some other cause for his stroke and his likely stroke subtype diagnosis.

Ischemic Stroke Subtype Diagnosis

Accuracy of Ischemic Stroke Subtype

Ischemic stroke may be caused by a variety of pathophysiological conditions and mechanisms. The distinction between ischemic stroke subtypes is important to guide specific secondary prevention measures such as treatment with anticoagulants that are useful in patients with cardiogenic embolism. In contrast, anticoagulants are not useful for patients with atherosclerotic stroke.^{74,75} Patients with carotid artery distribution symptoms who have an ipsilateral high-grade extracranial carotid artery stenosis benefit from carotid endarterectomy.⁷⁶ Simple clinical features useful at the bedside can help. For example, the acute onset of a focal neurologic deficit in a patient with a cardiac or arterial embolic source increases the odds of embolic stroke up to nearly 11-fold (LR+, 11; 95% CI, 5.7-21), whereas the absence of these features decreases the odds of embolic stroke by approximately one-quarter to one-half (LR-, 0.36; 95% CI, 0.24-0.56).⁷⁷

Reliability

Only a few studies have considered the reliability of classification of stroke type based solely on clinical findings. The available data indicate that a physician's assessment of ischemic stroke subtype according to medical history and physical

examinations alone is not reliable. For example, the Stroke Data Bank Investigators found that agreement on classification of stroke subtypes (cardiogenic embolism, large artery atherosclerosis, tandem arterial pathology, lacunar stroke, infarct of unknown cause, parenchymatous hemorrhage, and subarachnoid hemorrhage) was poor ($\kappa = 0.15$).⁷³ The combined poor accuracy and reliability means that radiographic and other tests are required to help identify the ischemic stroke subtype. The combination of the clinical findings and the neuroimaging results serves as the reference standard for determining the presence of an ischemic stroke.^{77,78}

Scenario

After the clinical examination, accurate ischemic stroke subtype diagnosis typically requires neuroimaging and other studies (ie, echocardiography to identify a source for possible emboli).

In this case, the brain CT scan result was interpreted as being normal. He was observed to have paroxysmal atrial fibrillation on a heart monitor during his CT examination. After careful review of the inclusion/exclusion criteria for

intravenous tPA, he was treated beginning 2 hours after the onset of his symptoms for a presumed ischemic stroke. Non-invasive studies later showed no evidence of extracranial carotid artery stenosis.

Prognosis

Patients with any combination of impaired consciousness, hemiplegia, and conjugate gaze palsy have a relatively higher mortality rate during the first 3 weeks after their stroke. Data from the prethrombolytic era showed that the presence of any of these findings had an LR of 1.8 for death (95% CI, 1.2-2.8), whereas the absence of all 3 had an LR of 0.36 (95% CI, 0.13-1.0).²⁴ Thirty-seven percent of those whose consciousness was initially impaired died, compared with no deaths among patients initially alert.²⁴ Several multivariable scoring systems have been developed to aggregate those findings believed by clinicians to reflect stroke severity and predict mortality (Table 48-8). These scores are calculated by adding points for abnormal clinical findings.

Table 48-8 Prognosis After Stroke According to Clinical Data^a

	24	25 ^a	26	27 ^a	28 ^a	29 ^a	79 ^a	Reference
Score Components								
Orientation		+	+	+	+	+	+	
Level of consciousness	+	+	+	+	+	+	+	
Neglect			+	+	+			
Language	+			+		+	+	
Gaze preference	+	+				+	+	
Visual field defect			+	+			+	
Facial paresis		+	+			+	+	
Dysarthria			+					
Arm strength	+	+	+		+	+	+	
Leg strength	+	+	+		+	+	+	
Ambulation				+				
Plantar response							+	
Sensation		+	+	+		+	+	
General function	+			+		+		
Accuracy								
Mortality								
Sensitivity (95% CI), %	80 (55-93)	72 (63-79)	...	60 (48-72)	85 (76-91)	42 (26-61)	86 (77-92)	
Specificity (95% CI), %	56 (41-70)	99 (97-100)		94 (87-97)	80 (72-86)	95 (90-98)	60 (51-68)	
LR+ (95% CI)	1.8 (1.2-2.8)	77 (19-305)	...	9.4 (4.5-20)	4.3 (2.9-6.2)	9.0 (3.7-22)	2.1 (1.7-2.7)	
LR- (95% CI)	0.36 (0.13-1.0)	0.29 (0.22-0.38)		0.42 (0.31-0.58)	0.19 (0.12-0.31)	0.61 (0.44-0.84)	0.23 (0.13-0.42)	
Disability^c								
Sensitivity (95% CI), %	91 (83-95)	78 (64-88)	73 (60-83)	14 (7-26)	...	
Specificity (95% CI), %			86 (73-93)	86 (74-93)	77 (60-88)	97 (91-99)		
LR+ (95% CI)	6.4 (3.2-13)	5.5 (2.8-11)	3.1 (1.7-5.8)	4.5 (1.2-16)	...	
LR- (95% CI)			0.11 (0.05-0.21)	0.25 (0.15-0.44)	0.36 (0.22-0.56)	0.89 (0.79-0.99)		

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

^aBased on data from Goldstein and Matchar.¹

^bEllipses indicate study did not provide relevant data.

^cPredictions concerning disability refer to the chance of returning to independence in activities of daily living according to dichotomization into less vs more severe from impairment level scores, including the indicated items as reflected in each total score's definitions provided in the source articles.

Predicting functional outcome among stroke survivors is more complicated than predicting survival.⁸⁰ The results of functional outcome assessments vary, depending on when the assessments are performed and how outcome is measured. As with mortality, multivariate discriminant scores have also been used to predict dependency in activities of daily living (Table 48-8).^{24-29,79} The NIHSS score not only provides a numeric summary of a patient's neurologic impairments that allows monitoring for changes in the extent of deficits but also helps determine prognosis and the use of specific therapies. One study found that each additional point on the NIHSS, within 24 hours of stroke onset, was associated with a decrease in the likelihood of an excellent outcome at 7 days by 24% (OR, 0.76; 95% CI, 0.72-0.80) and at 3 months by 17% (OR, 0.83; 95% CI, 0.81-0.86).⁷⁹ As described above, the NIHSS predicts a patient's prognosis. Less than 20% of untreated patients with an NIHSS score of more than 15 at baseline recover to the point of having little or no disability.⁷⁹ Approximate point estimates predicting outcome at 3 months are based on NIHSS scores obtained within the first 24 hours of ischemic stroke (Table 48-9).^{81,82}

Scenario

The example patient was alert, was not hemiplegic, and did not have a conjugate gaze palsy. He has a low likelihood of in-hospital mortality related to the stroke. According to his NIHSS score of 9, he has an approximately 78% chance of having a good or excellent recovery by 3 months without treatment (Table 48-9). Twenty-four hours after he received tPA, a brain CT scan showed no evidence of hemorrhage, and he was administered warfarin for secondary stroke prophylaxis for an atrial fibrillation-related cardioembolic stroke. He was able to ambulate independently by the time of hospital discharge, and his speech disturbance improved (NIHSS score of 4). He received outpatient physical, occupational, and speech therapy and had an NIHSS score of 2 after 3 months.

THE BOTTOM LINE

The medical history and neurologic examination are critical tools for the identification and treatment of patients with suspected cerebrovascular disease. This is especially true in patients being evaluated soon after the onset of symptoms, before neuroimaging results are available, and in patients with transient symptoms in whom no parenchymal abnormality on brain neuroimaging may develop.

Among noncomatose patients without head trauma who have neurologically relevant symptoms for which stroke is a consideration, the prior probability of a TIA or stroke is approximately 10%.

The likelihood of stroke increases with the following acute neurologic deficits: facial droop, arm drift, or a speech disturbance. Despite the increased odds of stroke in patients who satisfy this simple clinical rule (using the CPSS, LR+, 5.5; 95% CI, 3.3-9.1), appropriate neuroimaging and other tests are still required to exclude other potentially treatable etiologies and to better define the stroke subtype.

Table 48-9 Prognosis at 3 Months According to the Baseline NIHSS for Patients With Ischemic Stroke^{a,b}

	NIHSS Score					
	0-3	4-6	7-10	11-15	16-22	≥23
Dead	1	2	4	9	18	34
Poor ^b	3	10	18	35	40	48
Good ^b	15	25	32	34	25	12
Excellent ^b	80	63	46	22	17	6

Abbreviation: NIHSS, National Institutes of Health Stroke Scale.

^aBased on data from Adams et al.⁷⁹ Data are presented as percentages.

^bOutcome was determined according to the Glasgow Outcome Scale⁸¹ and Barthel Index⁸²: poor, Glasgow Outcome Scale score <2 and Barthel Index score <60; good, Glasgow Outcome Scale score <2 or Barthel Index score <60; excellent, Glasgow Outcome Scale score ≥2 and Barthel Index score >60.

Reliability is lowest for historical items and subjective findings (ie, the sensory examination). Reliability is higher for objective findings such as motor impairment. The astute clinician is aware of these differences when weighing the relative diagnostic implications.

The NIHSS is widely used for recording the clinical findings because it improves reliability and provides information helpful for determining a patient's prognosis and management. Reliability improves with experience, and Web-based resources are available for training and certification (<http://nihss-english.trainingcampus.net/uas/modules/trees/windex.aspx>; accessed March 12, 2008).

Clinical findings may be suggestive of stroke type, but reliability is poor when the diagnosis is based solely on medical history and physical examination. Neuroimaging is required to exclude hemorrhage and other tests are necessary to help identify the ischemic stroke subtype. Ischemic stroke subtype is often never established with certainty during the process of care, so acute therapeutic decisions must sometimes be made with the knowledge that the ischemic stroke subtype diagnosis may be unreliable.

The severity of a patient's initial neurologic impairments provides a useful guide for prognosis.

Author Affiliations at the Time of the Original Publication

Department of Medicine (Neurology), Duke Center for Cerebrovascular Disease, Durham, North Carolina (Dr Goldstein); Center for Clinical Health Policy Research and Education, Duke University, and Department of Veterans Affairs Medical Center, Durham, North Carolina (Drs Goldstein and Simel); and Department of Medicine, Duke University, Durham, North Carolina (Dr Simel).

Acknowledgments

Dr Goldstein has received research grants from the National Institutes of Health (NIH), the Veterans Administration, and the Centers for Disease Control and Prevention/University of North Carolina at Chapel Hill; is the clinical trial site principal investigator, PROFESS (Boehringer Ingelheim), RESPECT (AGA Corporation); is a member of the steering committee for the SPARCL Trial (Pfizer-Parke-Davis); is a

member of the neurology executive committee, RESPECT Trial (AGA Corporation); has received speaking honoraria from Bayer and Pfizer-Parke-Davis; and is a consultant for and is a member of the advisory board of AstraZeneca, BMS/Sanofi, CuraGen Corp, DPharm, GlaxoSmithKline, Johnson & Johnson, Merck Research, Pfizer-Parke-Davis, and Proneuron Biotechnologies.

Dr Goldstein was supported in part through an NIH Mid-career award in patient-oriented research (k24-ns-01265).

We appreciate the contributions and advice offered by Brian Stettler, MD, and David Matchar, MD, on earlier versions of the manuscript.

REFERENCES

- Goldstein LB, Matchar DB. Clinical assessment of stroke. *JAMA*. 1994;271(14):1114-1120.
- American Heart Association. Heart disease and stroke statistics—2008 update. <http://216.185.112.5/presenter.jhtml?identifier=3000090>. Accessed June 13, 2008.
- Williams GR, Jiang JG, Matchar DB, Samsa GP. Incidence and occurrence of total (first-ever and recurrent) stroke. *Stroke*. 1999;30(12):2523-2528.
- Goldstein LB. Extracranial carotid artery stenosis. *Stroke*. 2003;34(11):2767-2773.
- Katzan IL, Gruber TM, Furlan AJ, et al. Cuyahoga County Operation Stroke speed of emergency department evaluation and compliance with National Institutes of Neurological Disorders and Stroke time targets. *Stroke*. 2003;34(4):994-998.
- Hacke W, Donnan G, Fieschi C, et al. Association of outcome with early stroke treatment: pooled analysis of ATLANTIS, ECASS, and NINDS rt-PA stroke trials. *Lancet*. 2004;363(9411):768-774.
- Menon SC, Pandey DK, Morgenstern LB. Critical factors determining access to acute stroke care. *Neurology*. 1998;51(2):427-432.
- Kothari R, Jauch E, Broderick J, et al. Acute stroke: delays to presentation and emergency department evaluation. *Ann Emerg Med*. 1999;33(1):3-8.
- Morris DL, Rosamond W, Madden K, Schultz C, Hamilton S. Prehospital and emergency department delays after acute stroke—the Genentech Stroke Presentation Survey. *Stroke*. 2000;31(11):2585-2590.
- Lacy CR, Suh DC, Bueno M, Kostis JB. Delay in presentation and evaluation for acute stroke. Stroke Time Registry for Outcomes Knowledge and Epidemiology (STROKE). *Stroke*. 2001;32(1):63-69.
- NINDS tPA Study Group. Generalized efficacy of t-PA for acute stroke: subgroup analysis of the NINDS t-PA stroke trial. *Stroke*. 1997;28(11):2119-2125.
- NINDS tPA Study Group. Intracerebral hemorrhage after intravenous t-PA therapy for ischemic stroke. *Stroke*. 1997;28(11):2109-2118.
- Brown DL, Johnston KC, Wagner DP, Haley CEJ. Predicting major neurological improvement with intravenous recombinant tissue plasminogen activator treatment of stroke. *Stroke*. 2004;35(1):147-150.
- Holleman DR, Simel DL. Does the clinical examination predict airflow limitation? *JAMA*. 1995;273(4):313-319.
- Sackett DL. A primer on the precision and accuracy of the clinical examination. *JAMA*. 1992;267(19):2638-2644.
- Simel DL, Samsa GP, Matchar DB. Likelihood ratios with confidence: sample size estimation for diagnostic test studies. *J Clin Epidemiol*. 1991;44(8):763-770.
- Kothari R, Hall K, Brott T, Broderick J. Early stroke recognition: developing an out-of-hospital NIH Stroke Scale. *Acad Emerg Med*. 1997;4(10):986-990.
- Kothari RU, Pancioli A, Liu T, Brott T, Broderick J. Cincinnati Prehospital Stroke Scale: reproducibility and validity. *Ann Emerg Med*. 1999;33(4):373-378.
- Kidwell CS, Starkman S, Eckstein M, Weems K, Saver JL. Identifying stroke in the field. *Stroke*. 2000;31(1):71-76.
- Karanja PN, Nelson JJ, Lefkowitz DS, et al. Validation of the ACAS TIA/stroke algorithm. *Neurology*. 1997;48(2):346-351.
- von Arbin M, Britton M, De Faire U, Helmers C, Miah K, Murray V. Validation of admission criteria to a stroke unit. *J Chronic Dis*. 1980;33(4):215-220.
- von Arbin M, Britton M, De Faire U, Helmers C, Miah K, Murray V. Accuracy of bedside diagnosis in stroke. *Stroke*. 1981;12(3):288-293.
- Panzer RJ, Feibel JH, Barker WH, Griner PF. Predicting the likelihood of hemorrhage in patients with stroke. *Arch Intern Med*. 1985;145(10):1800-1803.
- Oxbury JM, Greenhall RCD, Grainger KMR. Predicting the outcome of stroke: acute stage after cerebral infarction. *BMJ*. 1975;3(5976):125-127.
- Tuthill JE, Pozen TJ, Kennedy FB. A neurologic grading system for acute strokes. *Am Heart J*. 1969;78(1):53-57.
- Frithz G, Werner I. Studies on cerebrovascular strokes II: clinical findings and short-term prognosis in a stroke material. *Acta Med Scand*. 1976;199(1-2):133-140.
- Allen CMC. Predicting the outcome of acute stroke: a prognostic score. *J Neurol Neurosurg Psychiatry*. 1984;47:475-480.
- Henley S, Pettit S, Todd-Pokroppek A, Tupper A. Who goes home? predictive factors in stroke recovery. *J Neurol Neurosurg Psychiatry*. 1985;48(1):1-6.
- Fullerton KJ, Mackenzie G, Stout RW. Prognostic indices in stroke. *Q J Med*. 1988;66(250):147-162.
- Britton M, De Faire U, Miah K, Helmers C. Prognostication in acute cerebrovascular disease: subjective assessment and test of a prognostic score. *Acta Med Scand*. 1980;207(1-2):37-42.
- Suzuki I, Shimizu H, Ishijima B, Tani K, Sugishita M, Adachi N. Aphasic seizure caused by focal epilepsy in the left fusiform gyrus. *Neurology*. 1992;42(11):2207-2210.
- Lee H, Lerner A. Transient inhibitory seizures mimicking crescendo TIA. *Neurology*. 1990;40(1):165-166.
- Ross RT. Transient tumor attacks. *Arch Neurol*. 1983;40(10):633-636.
- Chari CR, Rao NS. Transient neurologic deficit as a presentation of tuberculosis of the central nervous system. *Neurology*. 1987;37(12):1884-1885.
- Britton M, Hindmarsh T, Murray V, Tyden SA. Diagnostic errors discovered by CT in patients with suspected stroke. *Neurology*. 1984;34(11):1504-1507.
- Foster JW, Hart RG. Hypoglycemic hemiplegia: two cases and a clinical review. *Stroke*. 1987;18(5):944-946.
- Norris JW, Hachinski VC. Misdiagnosis of stroke. *Lancet*. 1982;1(8267):328-331.
- Ziegler DK, Hassanein R. Prognosis in patients with transient ischemic attacks. *Stroke*. 1973;4(4):666-673.
- Werdelin L, Juhler M. The course of transient ischemic attacks. *Neurology*. 1988;38(5):677-680.
- Kidwell CS, Alger JR, Di Salle F, et al. Diffusion MRI in patients with transient ischemic attacks. *Stroke*. 1999;30(6):1174-1180.
- Vermeer SE, Prins ND, Den Heijer T, Hofman A, Koudstaal PJ, Breteler MMB. Silent brain infarcts and the risk of dementia and cognitive decline. *N Engl J Med*. 2003;348(13):1215-1222.
- Johnston SC, Gress DR, Browner WS, Sidney S. Short-term prognosis after emergency department diagnosis of TIA. *JAMA*. 2000;284(22):2901-2906.
- Allen CMC. Clinical diagnosis of the acute stroke syndrome. *Q J Med*. 1983;52:515-523.
- Calanchini PR, Swanson PD, Gotshall RA, et al. Cooperative study of hospital frequency and character of transient ischemic attacks IV: reliability of diagnosis. *JAMA*. 1977;238(19):2029-2033.
- Shinar D, Gross CR, Mohr JP, et al. Interrater variability in the assessment of neurological history and examination in the Stroke Data Bank. *Arch Neurol*. 1985;42(6):557-565.
- Goldstein LB, Bertels C, Davis JN. Interrater reliability of the NIH Stroke Scale. *Arch Neurol*. 1989;46(6):660-662.
- Cote R, Battista RN, Wolfson C, Boucher J, Adams J, Hachinski VC. The Canadian Neurological Scale: validation and reliability assessment. *Neurology*. 1989;39(5):638-643.
- Brott T, Adams HPJ, Olinger CP, et al. Measurements of acute cerebral infarction: a clinical examination scale. *Stroke*. 1989;20(7):864-870.
- Lindenstrom E, Boysen G, Christiansen LW, Hansen BR, Nielsen PW. Reliability of Scandinavian Neurological Stroke Scale. *Cerebrovasc Dis*. 1991;1:103-107.
- Dewey HM, Donnan GA, Freeman EJ, et al. Interrater reliability of the National Institutes of Health Stroke Scale: rating by neurologists and nurses in a community-based stroke incidence study. *Cerebrovasc Dis*. 1999;9(6):323-327.
- Meyer BC, Hemmen TM, Jackson CM, Lyden PD. Modified National Institutes of Health Stroke Scale for use in stroke clinical trials. *Stroke*. 2002;33(5):1261-1266.
- Cote R, Hachinski VC, Shurvell BL, Norris JW, Wolfson C. The Canadian Neurological Scale: a preliminary study in acute stroke. *Stroke*. 1986;17(4):731-737.

53. Gelmers HJ, Gorter K, De Weerd CJ, Wiezer HJA. Assessment of inter-observer variability in a Dutch multicenter study on acute ischemic stroke. *Stroke*. 1988;19(6):709-711.
54. Olsen J, Simonsen K, Norgaard B, et al. Reproducibility and utility of a simple neurological scoring system for stroke patients (Copenhagen Stroke Scale). *J Neurol Rehabil*. 1988;2:59-63.
55. Kramer MS, Feinstein AR. Clinical biostatistics LIV: the biostatistics of concordance. *Clin Pharmacol Ther*. 1983;29(1):111-123.
56. Tomasello F, Mariani F, Fieschi C, et al. Assessment of inter-observer differences in the Italian Multicenter Study on Reversible Cerebral Ischemia. *Stroke*. 1982;13(1):32-35.
57. Easton JD, Albers GW, Caplan LR, Saver JL, Sherman DG. Reconsideration of TIA terminology and definitions. *Neurology*. 2004;62(8 suppl 6):S29-S34.
58. Kraaijeveld CL, Van Gijn J, Schouten HJA, Staal A. Interobserver agreement for the diagnosis of transient ischemic attacks. *Stroke*. 1984;15(4):723-725.
59. Koudstaal PJ, Van Gijn J, Staal A, Duivenvoorden HJ, Gerritsma JGM, Kraaijeveld CL. Diagnosis of TIA: improvement of interobserver agreement by a checklist in ordinary language. *Stroke*. 1986;17(4):723-728.
60. Quality Standards Subcommittee of AAN. Thrombolytic therapy for acute ischemic stroke—summary statement. *Neurology*. 1996;47(3):835-839.
61. Adams HP, Adams RJ, Brott T, et al. Guidelines for the early management of patients with ischemic stroke: a scientific statement from the stroke council of the American Stroke Association. *Stroke*. 2003;34(4):1056-1083.
62. Albers GW, Amarenco P, Easton JD, Sacco RL, Teal P. Antithrombotic and thrombolytic therapy for ischemic stroke: the 7th ACCP Conference on Antithrombotic and Thrombolytic Therapy. *Chest*. 2004;126(3 suppl):483S-512S.
63. Camilo O, Goldstein LB. Statewide assessment of hospital-based stroke prevention and treatment services in North Carolina. *Stroke*. 2003;34(12):2945-2950.
64. Adams RJ, Meador KJ, Sethi KD, Grotta JC, Thomson DS. Graded neurological scale for use in hemispheric stroke treatment protocols. *Stroke*. 1987;18(3):665-669.
65. Vogel HP. Influence of additional information on interrater reliability in the neurologic examination. *Neurology*. 1992;42(11):2076-2081.
66. Bamford J, Sandercock P, Dennis M, Burn J, Warlow C. Classification and natural history of clinically identifiable subtypes of cerebral infarction. *Lancet*. 1991;337(8756):1521-1526.
67. Lindley RI, Warlow CP, Wardlaw JM, Dennis MS, Slattery J, Sandercock PAG. Interobserver reliability of a clinical classification of acute cerebral infarction. *Stroke*. 1993;24(12):1801-1804.
68. Treves TA, Karepov VG, Aronovich BD, Bornstein NM, Korczyn AD. Interrater agreement in evaluation of stroke patients with the Unified Neurological Stroke Scale. *Stroke*. 1994;25(6):1263-1264.
69. Goldstein LB, Samsa GP. Reliability of the National Institutes of Health Stroke Scale: extension to non-neurologists in the context of a clinical trial. *Stroke*. 1997;28(2):307-310.
70. Wang S, Lee SB, Pardue C, et al. Remote evaluation of acute ischemic stroke: reliability of National Institutes of Health Stroke Scale via telestroke. *Stroke*. 2003;34(10):E188-E191.
71. Tong DC, Yenari MA, Albers GW, O'Brien M, Marks MP, Moseley ME. Correlation of perfusion- and diffusion-weighted MRI with NIHSS score in acute (<6.5 hour) ischemic stroke. *Neurology*. 1998;50(4):864-870.
72. Poungvarin N, Viriyavejakul A, Komontri C. Siriraj stroke score and validation study to distinguish supratentorial intracerebral hemorrhage from infarction. *BMJ*. 1991;302(6792):1565-1567.
73. Gross CR, Shinar D, Mohr JP, et al. Interobserver agreement in the diagnosis of stroke type. *Arch Neurol*. 1986;43(9):893-898.
74. Coull BM, Williams LS, Goldstein LB, et al. Anticoagulants and antiplatelet agents in acute ischemic stroke: report of the Joint Stroke Guideline Development Committee of the American Academy of Neurology and the American Stroke Association. *Neurology*. 2002;59(1):13-22.
75. Mohr JP, Thompson JLP, Lazar RM, et al. A comparison of warfarin and aspirin for the prevention of recurrent ischemic stroke. *N Engl J Med*. 2001;345(20):1444-1451.
76. Biller J, Feinberg WM, Castaldo JE, et al. Guidelines for carotid endarterectomy: a statement for healthcare professionals from a special writing group of the Stroke Council, American Heart Association. *Stroke*. 1998;29(2):554-562.
77. Adams HP, Bendixen BH, Kappelle LJ, et al. Classification of subtype of acute ischemic stroke: definitions for use in a multicenter clinical trial. *Stroke*. 1993;24(1):35-41.
78. Goldstein LB, Jones MR, Matchar DB, et al. Improving the reliability of stroke subgroup classification using the Trial of ORG 10172 in Acute Stroke Treatment (TOAST) criteria. *Stroke*. 2001;32(5):1091-1097.
79. Adams HP, Davis PH, Leira EC, et al. Baseline NIH Stroke Scale score strongly predicts outcome after stroke: a report of the Trial of Org 10172 in Acute Stroke Treatment (TOAST). *Neurology*. 1999;53(1):126-131.
80. Jongbloed L. Prediction of function after stroke: a critical review. *Stroke*. 1986;17(4):765-776.
81. Jennett B, Bond M. Assessment of outcome after severe brain damage: a practical scale. *Lancet*. 1975;1:480-484.
82. Mahoney FI, Barthel DW. Functional evaluation: the Barthel Index. *Md State Med J*. 1965;14:61-65.

This page intentionally left blank

Prepared by David L. Simel, MD, MHS

Reviewed by Sheri Keitz, MD

UPDATED SUMMARY ON STROKE

Original Review

Goldstein LB, Matchar DB. Clinical assessment of stroke. *JAMA*. 1994;271(14):1114-1120.

STROKE—MAKE THE DIAGNOSIS

PRIOR PROBABILITY

Among emergency patients with nontraumatic, noncomatose, neurologically relevant complaints, the prevalence of stroke or transient ischemic attack is roughly 10%.

POPULATION FOR WHOM STROKE SHOULD BE CONSIDERED

Stroke can be considered in patients with a variety of symptoms and signs. Patient with acute neurologic findings, especially those associated with acute focal sensory deficits, focal weakness, change in mentation or level of consciousness, or sudden loss of ability to communicate effectively, should be evaluated for a stroke. Headache, seizure, and syncope are also important symptoms that can identify a patient with a stroke.

DETECTING THE LIKELIHOOD OF STROKE

Typically, the physician can rely on just a few findings for identifying the patient with a stroke (Table 48-10).

UPDATED LITERATURE SEARCH AND REVISED ARTICLE

Goldstein LB, Simel DL. Is this patient having a stroke? *JAMA*. 2005;293(19):2391-2402.

The Update was prepared within 12 months of publication of The Rational Clinical Examination article, so the “Make the Diagnosis” section summarizes the findings published in the original review. The “Clinical Scenario” and “Resolution” are published in the article shown in this section.

Table 48-10 Likelihood Ratios for Stroke From Summing Up Combinations of Findings

Combination of Findings ^a	Findings Present	LR+ (95% CI)
Cincinnati Prehospital Stroke Scale ¹		
Facial paresis	3 Present	14 (1.6-121)
Arm drift	2 Present	4.2 (1.4-13)
Abnormal speech	1 Present	5.2 (2.6-11)
	0 Present	0.39 (0.25-0.61)
Hospital Evaluation ²		
Persistent neurologic deficit	4 Present	40 (29-55)
Focal neurologic deficit	1-3 Present	Uncertain LR, but probability of stroke $\geq 10\%$
Acute onset of symptoms during the previous week	0 Present	0.14 (0.10-0.20)

Abbreviations: CI, confidence interval; LR, likelihood ratio; LR+, positive likelihood ratio.

^aThe findings should be applied to patients who have no head trauma.

REFERENCE STANDARD TESTS

Combination of clinical findings with neuroimaging results.

REFERENCES FOR THE UPDATE

- Kothari R, Hall K, Brott T, Broderick J. Early stroke recognition: developing an out-of-hospital NIH Stroke Scale. *Acad Emerg Med*. 1997;4(10):986-990.
- von Arbin M, Britton M, De Faire U, Helmers C, Miah K, Murray V. Validation of admission criteria to a stroke unit. *J Chronic Dis*. 1980;33(4):215-220.

This page intentionally left blank

CHAPTER 49

Does This Patient Have Temporal Arteritis?

Gerald W. Smetana, MD

Robert H. Shmerling, MD

CLINICAL SCENARIOS

CASE 1 A 74-year-old woman has recent onset of daily bitemporal headache but is otherwise well. Her general physical examination results are normal and the erythrocyte sedimentation rate (ESR) is moderately elevated, at 64 mm/h. You wonder whether additional medical history or physical examination findings will modify your suspicion of possible temporal arteritis (TA) or whether the historical features alone warrant proceeding to temporal artery biopsy.

CASE 2 A 53-year-old man has a 1-month history of fever and fatigue and reports a single episode of transient partial loss of vision in one eye. You believe that TA is among the diagnostic considerations but suspect that he is too young for this diagnosis. You wonder if additional medical history, physical examination, or laboratory testing will change the probability of TA sufficiently to alter your decision about the role of temporal artery biopsy, rather than pursuing diagnostic evaluation for carotid artery stenosis or other considerations first.

WHY IS THIS AN IMPORTANT QUESTION TO ANSWER WITH A CLINICAL EXAMINATION?

When faced with a patient with headache, fatigue, or other possible presenting symptom of TA, clinicians must be able to correctly and confidently establish the diagnosis to prevent irreversible vision loss and to minimize the inappropriate evaluation and treatment of alternative diagnoses. Although headache is the most common reason for clinical suspicion of TA, no single type of headache or other clinical presentation is specific for TA, and the disorder is among the diagnostic considerations for many symptom complexes in older individuals. Our review will analyze the diagnostic value of these varied symptoms and signs in predicting the likelihood of TA among patients for whom there is a clinical suspicion of disease.

The first known report of a patient with TA was by Hutchinson in 1890.¹ His case was that of a man who was referred because of "red streaks on his head" that were painful and prevented him from wearing his hat; these proved to be swollen temporal arteries, which over time became firm and pulseless. It was not until 1932 that Horton et al² described the first 2 cases of pathologically confirmed TA; both patients had fever, weakness, anorexia, weight loss, anemia, leukocytosis, and painful tender temporal arteries. Thus, many of the characteristic features of this newly described disease were present in these first few patients. Headache was absent. In 1937, headache was recognized as a common feature,³ and in 1938 vision loss was first reported.⁴ In the modern era, however, clinicians are unlikely to treat patients with such advanced disease and the full array of untreated symptoms.

The mortality of patients with treated TA, during follow-up periods as long as 12 years, is the same as for age-matched individuals without TA. For example, Matteson et al⁵ studied 205 of the patients with TA who formed the initial cohort for the development of the American College of Rheumatology classification criteria. During a mean of 7 years of follow-up, the survival for patients with giant-cell arteritis was nearly identical to that of age-matched controls; the standardized mortality ratio was 1.03. Other authors have also observed that no excess mortality exists among patients with TA during periods ranging from 4.5 to 12 years.⁶⁻⁸ Unrecognized (and therefore untreated) patients may have a higher mortality, but no such natural history studies of untreated patients exist in the modern era.

Although preventing death may not be among the benefits of early diagnosis of TA, timely diagnosis and treatment will prevent vision loss. A prompt decision regarding further evaluation (including referral for temporal artery biopsy) and early initiation of treatment are the primary rationales for improving the clinical prediction of the diagnosis. In addition, clinicians may avoid an extensive evaluation for other causes of symptoms by establishing a proper diagnosis. Because systemic corticosteroids have been the standard therapy for TA for decades, few studies have determined the long-term incidence of vision loss among untreated patients. Several studies, however, have demonstrated a substantial reduction in the incidence of vision loss after institution of corticosteroid therapy. Even among patients with complete unilateral vision loss, prompt recognition and corticosteroid therapy will decrease the risk of vision loss in the contralateral eye.

Aiello et al⁹ reviewed the Mayo Clinic experience of 245 patients diagnosed with TA who had a complete ophthalmologic examination at diagnosis or early in the course of treatment. The estimated 5-year probability of developing vision loss after initiation of corticosteroid therapy was 1%; that of additional vision loss in patients who already had vision loss was 13%. These observations and others emphasize the importance of the early diagnosis and treatment of TA and of the clinical examination in identifying patients at risk for catastrophic vision outcomes.¹⁰⁻¹²

Estimates of the prevalence of TA have been fairly constant. Using population data from Olmsted County, Minnesota, Salvarani et al¹³ estimated the age-adjusted incidence for individuals aged 50 years or older to be 24.2 per 100 000 women and 8.2 per 100 000 men. In another report, prevalence estimates increased by age and were 200 per 100 000 individuals aged 50 years and older, and 1100 per 100 000 individuals aged 85 years and older.¹⁴ These findings are similar to those observed in a Swedish population study, in which the average annual incidence of TA among individuals older than 50 years was 22.2 per 100 000 and the incidence increased with age.¹⁵ In this study of 665 patients with TA proven by biopsy, only 1 patient was younger than 50 years. Other investigators have reported similar incidences.^{16,17} That TA is predominantly a disease of older individuals has importance because of the aging of our society. In the US 2000 census (<http://factfinder.census.gov/servlet/ACSSAFF>

Facts?_submenuId=factsheet_1&_sse=on; accessed June 15, 2008), 35 million individuals (12.4% of the population) were aged 65 years or older and 9 million (3.3% of the population) were aged 80 years or older; these proportions are expected to increase.

The relatively low prevalence of TA does not diminish its importance to clinicians because of the morbidity resulting from overlooking this disorder. In fact, the higher prevalence of TA (1.5%) in one large autopsy series suggests that the disorder may be either unrecognized or clinically occult in many cases.¹⁸ The vision prognosis of occult TA is, of course, unknown, and series that describe the frequency of signs and symptoms include only patients with clinically evident TA.

Pathophysiology

The clinical manifestations of TA are a direct consequence of local (or “arteritic”) and systemic inflammatory disease. Localized arterial inflammation, particularly in the smaller branches of the external carotid artery, cause endovascular damage, vessel stenosis, and occlusion, ultimately leading to tissue ischemia or necrosis. Examples of localized arteritic symptoms include jaw claudication, caused by involvement of the masticatory muscles, and vision loss caused by involvement of the ophthalmic or posterior ciliary arteries. The particular cytokine profile may contribute to the ischemic and prominent constitutional features, such as malaise, fever, or weight loss.^{19,20}

How to Elicit the Signs and Symptoms

The myriad signs and symptoms in patients with TA require familiarity with the most common ones, recognizing that many patients will demonstrate few symptoms and have a normal physical examination result. Headache, jaw claudication, vision complaints, polymyalgia rheumatica (PMR), and constitutional features in a patient older than 55 years are among the most common symptoms. A high index of suspicion will lead the clinician to pursue these features because they may not be part of routine history-taking. Headache quality (typically severe and throbbing; less often sharp, dull, or burning), location (may be diffuse or localized but is bitemporal in half of cases), and onset (typically acute) are key features to assess; however, the headache of TA is often nonspecific in character.²¹ Headache may actually be due to scalp tenderness, reported by the patient as pain when combing the hair or putting on a hat. The headache is a new headache that is either recent in onset or different from previous headaches among patients with a history of chronic headaches. The duration of the headache before seeking medical attention is commonly 2 to 3 months. Jaw claudication refers to pain in the proximal jaw near the temporomandibular joint that develops only after a brief period of chewing, especially food requiring vigorous mastication, such as steak or a bagel.

Clinicians must distinguish jaw claudication from other causes of jaw pain in elderly persons, such as disorders of the temporomandibular joint (in which pain begins right away with chewing) or ill-fitting dentures. Vision complaints

commonly include sudden monocular blindness, but clinicians should ask patients about a stuttering onset of vision loss, amaurosis fugax, a field cut, or diplopia. As an inflammatory polyarthritis with tendon or bursal involvement, PMR typically causes abrupt onset of morning stiffness involving the neck, shoulders, and hips, with referred pain to the proximal arms and thighs; this explains the prominent myalgias.²² Although neoplasm and infection may be highly suspected in the older patient with fever, anorexia, weight loss, and malaise, systemic inflammatory disease such as TA may also cause these symptoms.

The physical examination result is frequently unremarkable in patients with TA, but the detection of certain abnormalities may increase the suspicion of disease. The patient's temperature and general appearance are important first steps. Abnormalities of the temporal arteries, including tenderness, reduced or absent pulsation, erythema, nodularity, or swelling, may be detected by light palpation just anterior and slightly superior to the tragus of the ear; following the pulse anteriorly along the temples and comparison with the contralateral side helps detect findings that may be remarkably focal. Scalp tenderness, usually near the temporal arteries, may also be evident by light palpation. The scalp and tongue should be inspected for ischemic or necrotic skin changes. The funduscopic examination, ideally with pupillary dilation, may reveal a pale or swollen disc (evidence of ischemic optic neuropathy)²³ or retinal artery occlusion, whereas vision field testing may demonstrate a field cut. Joint examination may reveal reduced range of motion in the shoulder or hip because of pain or more distal synovitis, particularly of the wrist.

METHODS

Search Strategy and Quality Review

We performed a MEDLINE search of English-language articles published between January 1966 and July 2000. Search terms included "temporal arteritis," "giant cell arteritis," "clinical features," "diagnosis," "diagnostic tests," "sensitivity and specificity," "medical history taking," "physical examination," "signs and symptoms," and "erythrocyte sedimentation rate." We identified additional references by the use of a previously published search strategy in The Rational Clinical Examination series.²⁴ This strategy combined 10 exploded Medical Subject Headings ("physical examination," "medical history taking," "professional competence," "sensitivity and specificity," "reproducibility of results," "observer variation," "diagnostic tests, routine," "decision support techniques," "Bayes theorem," "mass screening") and 2 text-word categories ("sensitivity and specificity" and "physical examination"), and intersected with "temporal arteritis." We identified additional articles, including those predating MEDLINE, through a hand search of the bibliographies of retrieved articles, previous reviews, monographs, and textbooks. Both authors independently reviewed all retrieved articles to determine their eligibility for our review and included only those articles in which agreement existed that the

study had met our inclusion criteria. We sought no unpublished studies.

The purpose of our review is to determine the value of individual clinical features in predicting the likelihood of positive results from temporal artery biopsy. Eligible studies were, therefore, those in which the authors provided a detailed list of clinical features for patients suspected of having or confirmed to have TA. We excluded articles with limited data on clinical features and those with fewer than 7 patients with positive temporal artery biopsy results. Many early studies classified patients as having TA according to either the authors' own clinical criteria alone or the presence of positive biopsy results. When a study considered both groups of patients as having TA, we required that at least 90% of included patients had undergone temporal artery biopsy and had had a positive result.

We classified each article by the pathologic criteria used to determine the presence of positive biopsy results and by the referral source for recruitment of patients. In some cases, authors published clinical data on the same or overlapping series of patients in more than 1 article. In these cases, if we could not determine with certainty that no overlap existed between the patients in these studies, we excluded all studies except for the report with the largest number of patients. Of 114 studies retrieved using our search strategy, 41 were eligible for our review. Twenty-one studies included patients with both positive and negative temporal artery biopsy results; these form the core of our review.

We determined whether the authors required any predetermined published clinical criteria for patient inclusion, such as the American College of Rheumatology criteria²⁵ for the diagnosis of TA, or other criteria. When studies used such criteria to classify patients as having TA with positive biopsy results or TA with negative biopsy results, we considered a positive biopsy result to be the true reference standard and considered only those patients with such results to have the disease. In our analysis, we included only those clinical features that were cited by at least 2 studies.

We classified the quality of evidence in each study by 2 methods. First, we developed our own criteria that focused on the diagnostic criteria (Table 49-1). This step was necessary to distinguish studies that used biopsy result as a reference standard from those that used established clinical criteria. In addition, we graded the quality of each study with a classification scheme for levels of evidence adapted from that previously developed for The Rational Clinical Examination series (see Table 1-7).²⁶ In this scheme of levels 1 through 5, the highest levels of evidence we found were in level 3 studies.

Statistical Methods

Sensitivity was defined as the proportion of patients with TA who had the particular sign or symptom; specificity was the proportion of patients without TA who did not have the particular sign or symptom. We calculated likelihood ratios (LRs) when authors reported clinical findings of

Table 49-1 Temporal Arteritis Diagnostic Criteria Quality Score

Score	Diagnostic Quality
1	Patients require biopsy confirmation to be classified as having temporal arteritis.
2	Patients are classified according to the presence of predefined established clinical criteria for temporal arteritis and on biopsy results.
3	All patients meet predefined established clinical criteria for temporal arteritis. The authors consider patients with negative biopsy results to have temporal arteritis if they meet established clinical criteria for temporal arteritis.
4	A series of consecutive patients with temporal arteritis proven by biopsy. No controls or patients with negative biopsy results.
5	No use of established clinical criteria. Patients do not require biopsy confirmation to be classified as having temporal arteritis.
6	The investigators require the presence of a particular symptom (eg, visual problems) in all patients with temporal arteritis.

patients suspected of having TA both with positive and negative temporal artery biopsy results. Summary measures for these dichotomous data and for the data reported on a continuous scale (eg, hemoglobin) were obtained with a random effects measure that gives broad 95% confidence intervals (CIs).^{27,28} Uncertainty in these measures is reflected in the broad 95% CIs around the estimates.

RESULTS

Precision and Accuracy

Twenty-one studies that met our inclusion criteria included patients with both positive and negative temporal artery biopsy results and form the basis of our review (Table 49-2). These studies reported clinical findings on a total of 2680 patients, 1050 of whom had positive temporal artery biopsy results. The overall prevalence (prior probability) of positive biopsy results among patients with a clinical suspicion of TA in these studies was 39%. All but 4 of the studies were retrospective chart reviews. Eleven of the studies were of the highest quality (study quality 1) according to our predetermined criteria, and 19 of the studies included all patients who had a temporal artery biopsy during the study period.

Precision of the Medical History and Physical Examination for Temporal Arteritis

No study that met our inclusion criteria evaluated the precision (ie, interobserver or intraobserver variation) of the medical history and physical examination for the diagnosis of TA. Most of the studies cited in this review are retrospective chart reviews and did not use standardized instruments for eliciting signs and symptoms across different observers. We therefore restrict our discussion to the accuracy of clinical findings.

Accuracy of Symptoms for the Diagnosis of Temporal Arteritis

Among the studies that included data on patients both with positive and negative temporal artery biopsy results, 14 symptoms were cited by at least 2 studies (Table 49-3). A limitation of our approach is that authors reported some findings much more frequently than others. However, our review incorporates the full extent of the published experience and presumably these reports include all of the major clinical features. Only 2 symptoms had LRs of sufficient power to be useful to clinicians. Jaw claudication had the highest LR+ (4.2), which is consistent with the traditional clinical teaching that jaw claudication, although somewhat insensitive, is a relatively specific feature for TA. When we pooled the sensitivity data from all eligible studies, including those studies that reported only patients with positive temporal artery biopsy results,^{20,25,53-71} jaw claudication was present in only 34% of patients with disease (Table 49-4).

More surprising was the finding that diplopia was the next most predictive symptom, with an LR+ of 3.4. Although the presence of diplopia substantially increases the likelihood of disease, the absence of diplopia does not significantly modify the probability of disease (LR-, 0.95) because of its low sensitivity (9% among all studies). We derived this value from 5 studies that evaluated this feature; previous reviews and textbooks have not emphasized the importance of diplopia. No other symptom had an LR+ exceeding 2. This includes features often thought to be useful to clinicians, such as fever, PMR, vision loss, and temporal headache. The LR- of all 14 symptoms was near 1. In other words, the absence of any particular symptom did not rule out TA or make the disorder substantially less likely. Patients with positive temporal artery biopsy results had a mean duration of symptoms of 3.5 months before diagnosis; this was 1.5 months (95% CI, 0.4-2.5 months) shorter than those with negative biopsy results, emphasizing the relatively acute onset of symptoms of biopsy-proven TA and that a longer duration of symptoms makes a positive temporal artery biopsy result less likely.

Accuracy of the Physical Examination for the Diagnosis of Temporal Arteritis

Findings on physical examination were more likely to influence the probability of positive temporal artery biopsy results than were historical features (Table 49-5). The presence of synovitis made positive temporal artery biopsy results significantly less likely (LR+, 0.41). The absence of any temporal artery abnormality also made disease substantially less likely (LR, 0.53). Scalp tenderness, a finding often thought to be specific for TA, did not perform well as a predictor of positive biopsy results. Among patients in whom TA was suspected, the frequency of scalp tenderness was similar in patients with and without the disease (LR+, 1.6).

Abnormal findings on examination of the temporal artery increased the probability of positive biopsy results and predicted disease to a greater extent than any other variable. Beading, prominence, or enlargement of the temporal artery

Table 49-2 Characteristics of Studies That Include Patients With Both Positive and Negative Temporal Artery Biopsy Results

Study, y	Diagnostic Quality/Level of Evidence ^a	Study Type	No. of Patients	Positive Biopsy Results, No. (%)	Referral Source ^{b,c}	Pathologic Criteria Used to Establish Positive Biopsy Results ^c	Comments
Gabriel et al, ²⁹ 1995	1/3	Retrospective	525	172 (33)	All	Achkar et al ³⁰	
Hayreh et al, ³¹ 1997	1/3	Prospective	363	106 (29)	All	Author	
McDonnell et al, ³² 1986	1/3	Retrospective	250	42 (17)	Specialty	Author	
Hall et al, ³³ 1983	1/3	Retrospective	134	46 (34)	All	Not stated	
Fernandez-Herlihy, ³⁴ 1988	1/3	Retrospective	107	29 (27)	All	Author	Omitted group C patients with equivocal biopsies
Chmelewski et al, ³⁵ 1992	1/3	Retrospective	98	30 (31)	All	Author	
Fauchald et al, ³⁶ 1972	1/3	Retrospective	94	61 (65)	All	Not stated	Comparison group patients all had PMR
Stuart, ³⁷ 1989	1/3	Retrospective	75	14 (19)	All	Allsop and Gallagher ³⁸	
Kent and Thomas, ³⁹ 1990	1/3	Retrospective	70	8 (11)	All	Not stated	
Roth et al, ⁴⁰ 1984	1/3	Retrospective	51	7 (14)	All	Not stated	
Bevan et al, ⁴¹ 1968	1/4	Retrospective	37	28 (76)	All	Author	Arteritis and giant cells pooled as biopsy-result positive
Duhaut et al, ⁴² 1999	2/3	Prospective	292	207 (71)	All	McDonnell et al ³²	All patients >50 y old, ESR >40 mm/h, response to 72 h of corticosteroids
Baldursson et al, ⁴³ 1994	2/3	Retrospective	133	127 (96)	All	ACR	
Gonzalez et al, ⁴⁴ 1989	2/4	Retrospective	21	10 (48)	All	Not stated	All patients met clinical criteria for GCA
Genereau et al, ⁴⁵ 1999	3/3	Retrospective	37	19 (51)	All	ACR	
Vilaseca et al, ⁴⁶ 1987	3/4	Retrospective	103	45 (44)	All	Allsop and Gallagher ³⁸	
Gur et al, ⁴⁷ 1996	3/4	Retrospective	39	30 (77)	Specialty and PCP	Banks et al ⁴⁸	All patients met ACR criteria for GCA
Brittain et al, ⁴⁹ 1991	5/4	Prospective	31	15 (48)	Not stated	Not stated	
Hedges et al, ⁵⁰ 1983	5/5	Retrospective	91	28 (31)	All	Author	Patients excluded if adequate chart documentation of history-taking was absent
Skaug et al, ⁵¹ 1995	6/3	Retrospective	98	13 (13)	Specialty	Not stated	All patients had eye complaints
Dixon et al, ⁵² 1966	6/4	Prospective	31	13 (42)	Specialty	Author	All patients had PMR

Abbreviations: ACR, 1990 American College of Rheumatology criteria for the diagnosis of giant-cell arteritis²⁵; ESR, erythrocyte sedimentation rate; GCA, giant-cell arteritis; PCP, primary care practices; PMR, polymyalgia rheumatica.

^aDiagnostic quality is described in Table 49-1. Levels of Evidence are those used for the Rational Clinical Examination series (Table 1-7).

^bAll, indicates all patients referred for biopsy; specialty, rheumatology or ophthalmology or other specialty practice; not stated, referral source not stated by authors.

^cAuthor, indicates author's own explicitly stated criteria; not stated, no pathologic criteria stated for a positive temporal artery biopsy.

all conferred LR+_s of greater than 4. A tender temporal artery also suggested an increased probability of positive biopsy results (LR, 2.6). An absent temporal artery pulse showed a trend toward a useful LR+; the value of 2.7 was, however, not statistically different from 1. The LRs for "any temporal artery abnormality" may underestimate their

power. If eligible studies did not list clinical features separately for each patient, it was not possible to determine whether specific temporal artery abnormalities overlapped; in such cases, we made the most conservative calculation about the actual number of patients with any temporal artery abnormality.

Table 49-3 Summary Likelihood Ratios for Symptoms Among Patients With Suspected Temporal Arteritis

Symptom/References	No. of Patients With Data on Variable ^a	LR+ (95% CI)	LR- (95% CI)
Jaw claudication ^{29,31-35,37,39,40,42,44-46,50-52}	2314	4.2 (2.8-6.2)	0.72 (0.65-0.81)
Diplopia ^{33,34,42,50,51}	703	3.4 (1.3-8.6)	0.95 (0.91-0.99)
Temporal headache ^{36,42}	386	1.5 (0.78-3.0)	0.82 (0.64-1.0)
Weight loss ^{31,34,36,37,39,41,42,46,47}	1417	1.3 (1.1-1.5)	0.89 (0.79-1.0)
Anorexia ^{34,37,39,41,42,46}	674	1.2 (0.96-1.4)	0.87 (0.75-1.0)
Fatigue ^{31,33,37,39,41,42,44,46}	1095	1.2 (0.98-1.4)	0.94 (0.86-1.0)
Fever ^{29,31,34-37,40-42,46,47}	1708	1.2 (0.98-1.4)	0.92 (0.85-0.99)
Any headache ^{29,31-35,37,39-47,50,51}	2475	1.2 (1.1-1.4)	0.7 (0.57-0.85)
Arthralgia ^{33,34,37,39,40,44,46,52}	582	1.1 (0.86-1.4)	1.0 (0.92-1.1)
Any vision symptom ^{29,32-37,39-42,44-47,51,52}	2083	1.1 (0.93-1.3)	0.97 (0.9-1.0)
Polymyalgia rheumatica ^{29,34,35,37,39,40,42,44,45,47,50}	1383	0.97 (0.76-1.2)	0.99 (0.83-1.2)
Myalgia ^{31,36,39,40,46}	681	0.93 (0.81-1.1)	1.1 (0.87-1.3)
Unilateral vision loss ^{32,50}	341	0.85 (0.58-1.2)	1.2 (1.0-1.3)
Vertigo ^{34,36,44}	212	0.71 (0.38-1.3)	1.1 (0.93-1.2)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

^aIncludes only studies that report results for patients with both positive and negative biopsy results.

LRs approaching 1 suggest that, among patients with a clinical suspicion for TA, the feature was as common among those with positive biopsy results as it was among those with negative results. We separately determined the sensitivity of physical examination features among all studies, including those restricted to patients with positive biopsy results

(sensitivity only studies, Table 49-6). In each study, physicians would have referred patients for a temporal artery biopsy when they believed the diagnosis to be sufficiently likely to justify a biopsy. These patients represent a selected sample who often manifested several clinical features of interest, including those analyzed in this review. Patients who lacked features commonly considered suggestive of TA were presumably less likely to have a temporal artery biopsy. This verification bias makes the value of those few findings with the highest positive LRs even greater because they help predict biopsy results among patients with a significant clinical suspicion of disease.

TA is more common among women than men and among whites than blacks. The LRs do not reflect this observation, perhaps because referring physicians incorporated this knowledge into their decisions about which patients to refer for biopsy. However, if one pools the data from all eligible studies, including those that reported only patients with positive temporal artery biopsy results, TA was 2.1 times more common in women than men (Table 49-6). TA among black patients in published reports is restricted largely to small case series,⁶⁶ and white patients constituted 86% of all patients with positive biopsy results.

Among patients referred for biopsy, the average age of those with positive results was 73 years; this was only 3.8 years (95% CI, 2.1-5.4 years) older than the average age of patients with negative results. Age was, however, a valuable criterion for predicting the likelihood of TA. When data for all eligible studies were reviewed, including those that reported only patients with positive biopsy results, 26 studies provided sufficient data to determine the age range of patients with biopsy-proven TA. Only 2 patients among a total of 1435 patients were younger than 50 years; this

Table 49-4 Summary Sensitivity of Symptoms Among All Patients With Positive Temporal Artery Biopsy Results^a

Variable	No. of Studies	Sensitivity (95% CI)
Any headache	32	0.76 (0.72-0.79)
Temporal headache	8	0.52 (0.36-0.67)
Weight loss	19	0.43 (0.35-0.53)
Fever	26	0.42 (0.33-0.52)
Fatigue	19	0.39 (0.28-0.52)
Myalgia	8	0.39 (0.23-0.56)
Any vision symptom	35	0.37 (0.30-0.44)
Anorexia	12	0.35 (0.23-0.48)
Polymyalgia rheumatica	30	0.34 (0.28-0.41)
Jaw claudication	35	0.34 (0.29-0.41)
Arthralgia	13	0.30 (0.21-0.40)
Unilateral vision loss	11	0.24 (0.14-0.36)
Facial pain	4	0.17 (0.12-0.23)
Bilateral vision loss	7	0.15 (0.07-0.27)
Vertigo	4	0.11 (0.05-0.19)
Diplopia	14	0.09 (0.07-0.13)

Abbreviation: CI, confidence interval.

^aIncludes results of all eligible studies, including those that reported clinical features for patients with positive biopsy results only.

Table 49-5 Summary Likelihood Ratios for Signs and Demographics and Laboratory Data Among Patients With Suspected Temporal Arteritis

Variable/References	No. of Patients With Data on Variable ^a	LR+ (95% CI)	LR- (95% CI)
Signs and Demographics			
Beaded temporal artery ^{42,52}	323	4.6 (1.1-18.4)	0.93 (0.88-0.99)
Prominent or enlarged temporal artery ^{36,39,42,44,52}	508	4.3 (2.1-8.9)	0.67 (0.5-0.89)
Absent temporal artery pulse ^{41,52}	68	2.7 (0.55-13.4)	0.71 (0.38-1.3)
Tender temporal artery ^{36,39-42,50-52}	755	2.6 (1.9-3.7)	0.82 (0.74-0.92)
Any temporal artery abnormality ^{29,31-33,37,43,46b}	1559	2.0 (1.4-3.0)	0.53 (0.38-0.75)
Scalp tenderness ^{31,33-35,52}	923	1.6 (1.2-2.1)	0.93 (0.86-1.0)
Optic atrophy or ischemic optic neuropathy ^{40,50}	142	1.6 (1.0-2.5)	0.8 (0.58-1.1)
Any funduscopic abnormality ^{29,35,50,52}	745	1.1 (0.8-1.4)	1.0 (0.92-1.1)
White race ^{32,35,37,40,50}	565	1.1 (0.99-1.2)	
Male sex ^{29,31-37,40-43,45-47,49-52}	2565	0.83 (0.72-0.96)	
Synovitis ^{29,37,46,52}	734	0.41 (0.23-0.72)	1.1 (1.0-1.2)
Laboratory Data			
ESR			
>100 mm/h ^{35,49,50}	220	1.9 (1.1-3.3)	0.8 (0.68-0.95)
>50 mm/h ^{35,47,49,50}	259	1.2 (1.0-1.4)	0.35 (0.18-0.67)
Abnormal ^{32,37,42,46,49-51}	941	1.1 (1.0-1.2)	0.2 (0.08-0.51)
Anemia ^{31,32,34,35,37,46,47,49}	1057	1.5 (0.82-2.9)	0.79 (0.6-1.0)

Abbreviations: CI, confidence interval; ESR, erythrocyte sedimentation rate; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

^aIncludes only studies that report results for patients with both positive and negative biopsy results.

^bIncludes only abnormalities that are not classified more specifically by the cited studies. The true incidence of any abnormality is presumably higher but cannot be calculated from the primary data.

resulted in a sensitivity of 99% for the criterion of age older than 50 years. This outcome suggests that clinicians should consider TA only as a diagnostic possibility in a person younger than 50 years if multiple characteristic or high-probability features are present.

Accuracy of the Laboratory Evaluation for the Diagnosis of Temporal Arteritis

Although the primary purpose of this analysis was to determine the operating characteristics of the medical history and physical examination in diagnosis, clinicians usually obtain an ESR before determining which patients have sufficient likelihood of TA to justify a referral for biopsy. We therefore chose to evaluate the test characteristics of the ESR. The mean value for patients with disease was 88 mm/h; that for patients without disease was a mean of 10 mm/h lower (95% CI, 4-25 mm/h). This difference was not statistically significant.

Results of the ESR measurement were a valuable guide to clinicians; a low or normal level was more likely to rule out disease than a high value was likely to rule in disease. Previously, Miller et al⁷² had determined normal ESR values among 27912 adults without apparent disease and suggested defining the upper limit of normal ESR as either age/2 (for men) or as (age + 10)/2 (for women). In our source studies, authors most commonly did not define "normal" ESR; it was not possible to determine whether these normal values were adjusted for age. With this caveat, a normal ESR made TA unlikely; the LR for a normal ESR was 0.2 (Table 49-5).

Table 49-6 Summary Sensitivity of Signs and Demographics and Laboratory Data Among All Patients With Positive Temporal Artery Biopsy Results^a

Variable	No. of Studies With Data on Variable	Sensitivity (95% CI)
Signs and Demographics		
White race	11	0.86 (0.62-0.97)
Any temporal artery abnormality	16	0.65 (0.54-0.74)
Prominent or enlarged temporal artery	6	0.47 (0.40-0.54)
Absent temporal artery pulse	6	0.45 (0.26-0.66)
Tender temporal artery	13	0.41 (0.30-0.52)
Male sex	40	0.32 (0.29-0.35)
Any funduscopic abnormality	6	0.31 (0.14-0.54)
Scalp tenderness	13	0.31 (0.20-0.44)
Optic atrophy or ischemic optic neuropathy	4	0.29 (0.10-0.57)
Beaded temporal artery	3	0.16 (0.07-0.28)
Laboratory Data		
ESR		
Abnormal	24	0.96 (0.93-0.97)
>50 mm/h	14	0.83 (0.75-0.90)
>100 mm/h	10	0.39 (0.29-0.50)
Anemia	22	0.44 (0.34-0.54)

Abbreviations: CI, confidence interval; ESR, erythrocyte sedimentation rate.

^aIncludes results of all eligible studies, including those that reported clinical features for patients with positive biopsy results only.

When we separately analyzed the pooled data from all studies, only 4% of patients with positive temporal artery biopsy results and data on ESR had a normal value. If one uses a less strict cutoff point, even an ESR of less than 50 mm/h substantially reduces the probability of disease (LR, 0.35). This value is lower than the LR₋ of any symptom or sign.

In contrast to clinical lore, a high ESR was less useful in identifying those with TA among all patients referred for biopsy, which likely relates to the verification bias inherent in patient selection for the eligible studies because referring physicians would have had knowledge of the ESR before recommending a biopsy. Although an ESR of greater than 100 mm/h conferred an LR₊ of 1.9, this value is less than the most useful symptoms and signs. In contrast, mean ESR values were similar for patients with and without positive temporal artery biopsy results.

Anemia was present in 44% of patients with biopsy-proven TA. This finding was present in a similar number of patients who had negative biopsy results. Mean hemoglobin levels were similar between patients with positive and negative biopsy results (11.6 g/dL vs 12.4 g/dL, respectively); the lack of anemia was not helpful in ruling out disease.

ARE THESE CLINICAL FEATURES EVER NORMAL?

The presence of particular symptoms or signs in patients with negative temporal artery biopsy results does not imply that these findings are “normal” or common in patients without disease. Rather, it suggests that other conditions that clinicians may initially confuse for TA have overlapping clinical features. The frequency of such findings in randomly selected individuals of the same age would likely be lower than the frequency among patients in this review with negative biopsy results.

Several studies have followed patients with negative biopsy results to determine their ultimate or correct diagnoses. Chmelewski et al³⁵ reported the outcomes of 98 patients undergoing temporal artery biopsies during a 5-year period at their institution. Among the 68 patients with negative biopsy results, 15 proved to have neurologic disorders (including migraine, stroke, and optic neuropathy), 14 had PMR, 10 had other rheumatologic disorders (including vasculitis other than TA, rheumatoid arthritis, and CREST [calcinosis, Raynaud disease, esophageal dysmotility, sclerodactyly, telangiectasia] syndrome), and 4 had fever of unknown origin. Miscellaneous diagnoses included sinusitis, endocarditis, amyloidosis, and malignancy. In another biopsy series, Roth et al⁴⁰ studied 33 patients with a clinical suspicion of TA but negative biopsy results. The most common diagnoses, in descending order, were joint disease (degenerative or rheumatoid), malignant lymphoma, arteriosclerotic carotid artery disease, diabetes mellitus, and ischemic optic neuropathy.

In our first clinical scenario, the history of bitemporal headache and a modestly increased ESR would be among those factors that may lead a clinician to suspect TA. In this setting, one would seek the potential additional history of jaw claudication or diplopia and determine the presence of a

prominent, tender, or beaded temporal artery. If present, these factors would substantially increase the likelihood of positive temporal artery biopsy results.

In the second scenario, TA is among the diagnostic considerations for transient partial monocular vision loss in the setting of a constitutional illness. The history in this case is sufficiently compelling to justify a temporal artery biopsy. Given the high prior probability and the poor performance of historical and examination features in excluding disease, an otherwise normal medical history and physical examination result would not sufficiently reduce the likelihood of TA to avoid the need for a temporal artery biopsy. A normal ESR would, however, reduce the likelihood of disease by a factor of 0.2 and should prompt consideration of alternative diagnoses.

THE BOTTOM LINE

Available data suggest that many of the clinical features commonly found in patients with the disease are unhelpful in predicting the likelihood of positive temporal artery biopsy results. Our study evaluates the predictive value of clinical features among patients who are already clinically suspected of having the disease, as determined by the clinicians who referred them for biopsy. Although we could not determine, from the primary studies, the factors that went into the decision to refer for biopsy, certain clinical features modified the likelihood of disease among these patients. It is likely that these same clinical factors would be useful to consider at initial evaluation, even before the decision to proceed to biopsy. In addition, the verification bias inherent in this analysis makes the significance of our results greater because they help to predict biopsy results even among patients who have a higher prior probability of disease than do unselected patients with any particular clinical feature.

When a medical history is taken in a patient with possible TA, jaw claudication and diplopia substantially increase the probability of positive biopsy results (LR₊s, 4.2 and 3.4, respectively). No symptoms help rule out the diagnosis by their absence. Among physical examination findings, synovitis makes the diagnosis of TA less likely, whereas beaded, prominent, enlarged, and tender temporal arteries increase the likelihood of positive biopsy results. Beaded, prominent, or enlarged arteries confer the highest positive LRs of any clinical or laboratory feature and substantially increase the probability that a patient with suspected TA will have positive biopsy results. Although these findings increase the chance of having TA, they are variably sensitive, from 16% (beaded temporal artery) to 65% (any temporal artery abnormality).

The results of tests of ESR alter the likelihood of positive biopsy results. A normal ESR (LR, 0.2) or ESR less than 50 mm/h (LR, 0.35) makes positive biopsy results less likely, but setting the ESR threshold at 100 mm/h is less efficient because patients with an ESR less than 100 mm/h have an LR (0.8) that only slightly decreases the likelihood of disease. Among patients clinically suspected of having disease, those

with an ESR greater than 100 mm/h have a modestly increased likelihood of biopsy-proven TA (LR, 1.9).

The clinician faced with a patient who may have TA has a difficult diagnostic challenge. The goal is to rule out other morbid conditions that may mimic TA, to avoid unnecessary evaluation, and to quickly and correctly identify and treat patients who do in fact have the disorder. Given the extreme difference in prevalence of TA between the general population (<1%) vs those referred for temporal artery biopsy (39%), we infer that clinicians are adept at identifying patients at high risk for disease. Many clinicians choose to treat patients they have referred for biopsy with corticosteroids, in the absence of contraindication, pending biopsy results. Although this strategy would appear particularly wise in the presence of a factor that we have shown predicts likelihood of disease, this approach deserves further study.

Our review of clinical series of patients with suspected TA does not allow a determination of the predictive value of selected combinations of clinical and laboratory features. In addition, it is not possible to determine from our data whether certain combinations of features would sufficiently increase the likelihood of disease that a clinician should treat presumptively for TA and not perform a biopsy at all. The morbidity of a prolonged course of corticosteroids, however, is such that most clinicians would favor confirmation of disease by biopsy even if the clinical probability is high.

Our analysis demonstrates that a limited number of clinical features substantially modify the probability of the diagnosis of TA among patients suspected of having the disease. Ultimately, the clinician must integrate multiple clinical factors to optimize diagnostic and therapeutic strategies for patients with suspected TA.

Author Affiliations at the Time of the Original Publication

Division of General Medicine and Primary Care (Dr Smetana) and Division of Rheumatology (Dr Shmerling), Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, Massachusetts.

Acknowledgments

We appreciate the expert advice offered by Stephanie Studenski, MD, and Kenneth Schmader, MD, during the preparation of our manuscript. We also wish to thank David Simel, MD, MHS, for his thoughtful guidance, review of the manuscript, and statistical advice.

REFERENCES

1. Hutchinson J. Diseases of the arteries: on a peculiar form of thrombotic arteritis of the aged which is sometimes productive of gangrene. *Arch Surg.* 1890;1:323-329.
2. Horton B, Magath T, Brown G. An undescribed form of arteritis of the temporal vessels. *Proc Staff Mtg Mayo Clin.* 1932;7:700-701.
3. Horton B, Magath T. Arteritis of the temporal vessels: report of seven cases. *Proc Staff Mtg Mayo Clin.* 1937;12:548-553.
4. Jennings G. Arteritis of the temporal vessels. *Lancet.* 1938;1:424-428.
5. Matteson EL, Gold KN, Block DA, Hunder GG. Long-term survival of patients with giant cell arteritis in the American College of Rheumatology giant cell arteritis classification criteria cohort. *Am J Med.* 1996;100(2):193-196.
6. Bengtsson B, Malmvall B. Prognosis of giant cell arteritis including temporal arteritis and polymyalgia rheumatica. *Acta Med Scand.* 1981;209(5):337-345.
7. Gonzalez-Gay MA, Blanco R, Abraira V, et al. Giant cell arteritis in Lugo, Spain, is associated with low longterm mortality. *J Rheumatol.* 1997;24(11):2171-2176.
8. Gouet D, Marechaud R, Alcalay M, et al. Survival in giant cell arteritis: a survey of 87 patients [letter]. *J Rheumatol.* 1985;12(6):1209-1210.
9. Aiello PD, Trautman JC, McPhee TJ, Kinselman AR, Hunder GG. Visual prognosis in giant cell arteritis. *Ophthalmology.* 1993;100(4):550-555.
10. Font C, Cid MC, Coll-Vinent B, Lopez-Soto A, Grau JM. Clinical features in patients with permanent visual loss due to biopsy-proven giant cell arteritis. *Br J Rheumatol.* 1997;36(2):251-254.
11. Liu GT, Glaser JS, Schatz NJ, Smith JL. Visual morbidity in giant cell arteritis: clinical characteristics and prognosis for vision. *Ophthalmology.* 1994;101(11):1779-1785.
12. Myles AB. Prognosis of polymyalgia rheumatica and giant cell arteritis. *Baillieres Clin Rheumatol.* 1991;5(3):493-503.
13. Salvarani C, Gabriel SE, O'Fallon WM, Hunder GG. The incidence of giant cell arteritis in Olmsted County, Minnesota: apparent fluctuations in a cyclic pattern. *Ann Intern Med.* 1995;123(3):192-194.
14. Lawrence RC, Helmick CG, Arnett FC, et al. Estimates of the prevalence of arthritis and selected musculoskeletal disorders in the United States. *Arthritis Rheum.* 1998;41(5):778-799.
15. Petrusdottir V, Johansson H, Nordborg E, Nordborg C. The epidemiology of biopsy-positive giant cell arteritis: special reference to cyclic fluctuations. *Rheumatology (Oxford).* 1999;38(12):1208-1212.
16. Boesen P, Sorensen S. Giant cell arteritis, temporal arteritis, and polymyalgia rheumatica in a Danish county: a prospective investigation. *Arthritis Rheum.* 1987;30(3):294-299.
17. Gran JT, Myklebust G. The incidence of polymyalgia rheumatica and temporal arteritis in the county of Aust Agder, South Norway: a prospective study 1987-94. *J Rheumatol.* 1997;24(9):1739-1743.
18. Ostberg G. Temporal arteritis in a large necropsy series. *Ann Rheum Dis.* 1971;30(3):224-235.
19. Roche NE, Fulbright JW, Wagner AD, Hunder GG, Gorony JJ, Weyand CM. Correlation of interleukin-6 production and disease activity in polymyalgia rheumatica and giant cell arteritis. *Arthritis Rheum.* 1993;36(9):1286-1294.
20. Weyand CM, Tetzlaff N, Bjornsson J, Brack A, Younge B, Gorony JJ. Disease patterns and tissue cytokine profiles in giant cell arteritis. *Arthritis Rheum.* 1997;40(1):19-26.
21. Solomon S, Cappa KG. The headache of temporal arteritis. *J Am Geriatr Soc.* 1987;35(2):163-165.
22. Salvarani C, Cantini F, Olivieri I, et al. Proximal bursitis in active polymyalgia rheumatica. *Ann Intern Med.* 1997;127(1):27-31.
23. Hayreh SS, Podhajsky PA, Zimmerman B. Ocular manifestations of giant cell arteritis. *Am J Ophthalmol.* 1998;125(4):509-520.
24. Lederle FA, Simel DL. Does this patient have abdominal aortic aneurysm? *JAMA.* 1999;281(1):77-82.
25. Hunder GG, Bloch DA, Michel BA, et al. The American College of Rheumatology 1990 criteria for the classification of giant cell arteritis. *Arthritis Rheum.* 1990;33(8):1122-1128.
26. Holleman DR, Simel DL. Does the clinical examination predict airflow limitation? *JAMA.* 1995;273(4):313-319.
27. Eddy D, Hasselblad V, Shachter R. *Meta-analysis by the Confidence Profile Method: The Statistical Synthesis of Evidence.* San Diego, CA: Academic Press; 1992.
28. Eddy D, Hasselblad V. *Fast*Pro Software for Meta-analysis by the Confidence Profile Method.* 1.8 ed. San Diego, CA: Academic Press; 1992.
29. Gabriel SE, O'Fallon WM, Achkar AA, Lie JT, Hunder GG. The use of clinical characteristics to predict the results of temporal artery biopsy among patients with suspected giant cell arteritis. *J Rheumatol.* 1995;22(1):93-96.
30. Achkar A, Lie JT, Hunder GG, O'Fallon WM, Gabriel SE. How does prior corticosteroid treatment affect the biopsy findings in giant cell arteritis? *Ann Intern Med.* 1994;120(12):987-992.
31. Hayreh SS, Podhajsky PA, Raman R, Zimmerman B. Giant cell arteritis: validity and reliability of various diagnostic criteria. *Am J Ophthalmol.* 1997;123(3):285-296.
32. McDonnell PJ, Moore GW, Miller NR, Hutchins GM, Green WR. Temporal arteritis: a clinicopathologic study. *Ophthalmology.* 1986;93(4):518-530.
33. Hall S, Persellin S, Lie JT, O'Brien PC, Kurland LT, Hunder GG. The therapeutic impact of temporal artery biopsy. *Lancet.* 1983;2(8361):1217-1220.

34. Fernandez-Herlihy L. Temporal arteritis: clinical aids to diagnosis [published correction appears in *J Rheumatol.* 1989;16:260]. *J Rheumatol.* 1988;15(12):1797-1801.
35. Chmielewski WL, McKnight KM, Agudelo CA, Wise CM. Presenting features and outcomes in patients undergoing temporal artery biopsy: a review of 98 patients. *Arch Intern Med.* 1992;152(8):1690-1695.
36. Fauchald P, Rygvol O, Oystese B. Temporal arteritis and polymyalgia rheumatica: clinical and biopsy findings. *Ann Intern Med.* 1972;77(6):845-852.
37. Stuart RA. Temporal artery biopsy in suspected temporal arteritis: a five year survey. *N Z Med J.* 1989;102(874):431-433.
38. Allsop C, Gallagher P. Temporal artery biopsy in giant cell arteritis: a reappraisal. *Am J Surg Pathol.* 1981;5(4):317-323.
39. Kent R, Thomas L. Temporal artery biopsy. *Am Surg.* 1990;56(1):16-21.
40. Roth AM, Milsow L, Keltnor JL. The ultimate diagnoses of patients undergoing temporal artery biopsies. *Arch Ophthalmol.* 1984;102(6):901-903.
41. Bevan AT, Dunnill MS, Harrison MJ. Clinical and biopsy findings in temporal arteritis. *Ann Rheum Dis.* 1968;27(3):271-277.
42. Duhaut P, Pinede L, Bornet H, et al, for the Groupe de Recherche sur l'Arterite a Cellules Geantes. Biopsy proven and biopsy negative temporal arteritis: differences in clinical spectrum at the onset of the disease. *Ann Rheum Dis.* 1999;58(6):335-341.
43. Baldursson O, Steinsson K, Bjornsson J, Lie JT. Giant cell arteritis in Iceland: an epidemiologic and histopathologic analysis. *Arthritis Rheum.* 1994;37(7):1007-1012.
44. Gonzalez EB, Varner WT, Lisse JR, Daniels JC, Hokanson JA. Giant-cell arteritis in the southern United States: an 11-year retrospective study from the Texas Gulf Coast. *Arch Intern Med.* 1989;149(7):1561-1565.
45. Genereau T, Lortholary O, Guillemin L, et al. Temporal 67-gallium uptake is increased in temporal arteritis. *Rheumatology (Oxford).* 1999;38(8):709-713.
46. Vilaseca J, Gonzalez A, Cid MC, Lopez-Vivancos J, Ortega A. Clinical usefulness of temporal artery biopsy. *Ann Rheum Dis.* 1987;46(4):282-285.
47. Gur H, Rapman E, Ehrenfeld M, Sidi Y. Clinical manifestations of temporal arteritis: a report from Israel. *J Rheumatol.* 1996;23(11):1927-1931.
48. Banks P, Cohen M, Ginsburg W, Hunder GG. Immunohistologic and cytochemical studies of temporal arteritis. *Arthritis Rheum.* 1983;26(10):1201-1207.
49. Brittain G, McIlwaine G, Bell J, Gibson J. Plasma viscosity or erythrocyte sedimentation rate in the diagnosis of giant cell arteritis? *Br J Ophthalmol.* 1991;75(11):656-659.
50. Hedges TR III, Gieger GL, Albert DM. The clinical value of negative temporal artery biopsy specimens. *Arch Ophthalmol.* 1983;101(8):1251-1254.
51. Skaug TR, Midelfart A, Jacobsen G. Clinical usefulness of biopsy in giant cell arteritis. *Acta Ophthalmol Scand.* 1995;73(6):567-570.
52. Dixon A, Beardwell C, Kay A, Wanka J, Wong Y. Polymyalgia rheumatica and temporal arteritis. *Ann Rheum Dis.* 1966;25(3):203-208.
53. Brack A, Martinez-Taboada V, Stanson A, Goronzy JJ, Weyand CM. Disease pattern in cranial and large-vessel giant cell arteritis. *Arthritis Rheum.* 1999;42(2):311-317.
54. Branum G, Massey EW, Rice J. Erythrocyte sedimentation rate in temporal arteritis. *South Med J.* 1987;80(12):1527-1528.
55. Cid MC, Font C, Oristrell J, et al. Association between strong inflammatory response and low risk of developing visual loss and other cranial ischemic complications in giant cell (temporal) arteritis. *Arthritis Rheum.* 1998;41(1):26-32.
56. Dare B, Byrne E. Giant cell arteritis: a five-year review of biopsy proven cases in a teaching hospital. *Med J Aust.* 1980;1(8):372-373.
57. Desmet G, Knockaert D, Bobbaers H. Temporal arteritis: the silent presentation and delay in diagnosis. *J Intern Med.* 1990;227(4):237-240.
58. Dimant J, Grob D, Brunner NG. Ophthalmoplegia, ptosis, and miosis in temporal arteritis. *Neurology.* 1980;30(10):1054-1058.
59. Fainaru M, Friedman G, Friedman B. Temporal arteritis in Israel: a review of 47 patients. *J Rheumatol.* 1979;6(3):330-335.
60. Glutz von Blotzheim S, Borruat FX. Neuro-ophthalmic complications of biopsy-proven giant cell arteritis. *Eur J Ophthalmol.* 1997;7(4):375-382.
61. Gonzalez-Gay MA, Blanco R, Rodriguez-Valverde V, et al. Permanent visual loss and cerebrovascular accidents in giant cell arteritis: predictors and response to treatment. *Arthritis Rheum.* 1998;41(8):1497-1504.
62. Hauser W, Ferguson R, Holley K, Kurland L. Temporal arteritis in Rochester, Minnesota, 1951 to 1967. *Mayo Clin Proc.* 1971;46(9):597-602.
63. Healey LA, Wilske KR. Presentation of occult giant cell arteritis. *Arthritis Rheum.* 1980;23(6):641-643.
64. Jacobson DM, Slamovits TL. Erythrocyte sedimentation rate and its relationship to hematocrit in giant cell arteritis. *Arch Ophthalmol.* 1987;105(7):965-967.
65. Jonasson F, Cullen JF, Elton RA. Temporal arteritis: a 14-year epidemiological, clinical and prognostic study. *Scott Med J.* 1979;24(2):111-117.
66. Love DC, Rapkin J, Lesser GR, et al. Temporal arteritis in blacks. *Ann Intern Med.* 1986;105(3):387-389.
67. Machado EB, Michet CJ, Ballard DJ, et al. Trends in incidence and clinical presentation of temporal arteritis in Olmsted County, Minnesota, 1950-1985. *Arthritis Rheum.* 1988;31(6):745-749.
68. Mambo NC. Temporal (granulomatous) arteritis: a histopathological study of 32 cases. *Histopathology.* 1979;3(3):209-221.
69. Myklebust G, Gran JT. A prospective study of 287 patients with polymyalgia rheumatica and temporal arteritis: clinical and laboratory manifestations at onset of disease and at the time of diagnosis. *Br J Rheumatol.* 1996;35(11):1161-1168.
70. Wadman B, Werner I. Observations on temporal arteritis. *Acta Med Scand.* 1972;192(5):377-383.
71. Whitfield A, Bateman M, Cooke W. Temporal arteritis. *Br J Ophthalmol.* 1963;47:555-566.
72. Miller A, Green M, Robinson D. Simple rule for calculating normal erythrocyte sedimentation rate. *Br Med J (Clin Res Ed).* 1983;286(6361):266.

Prepared by Gerald W. Smetana, MD,
and Robert H. Shmerling, MD
Reviewed by Dan Solomon, MD

CLINICAL SCENARIO

A 72-year-old woman whom you have treated for the past decade comes to see you out of concern for her new-onset headaches. During the past month, her headaches have largely been stable, without progression. She has been tired but volunteers no other associated symptoms. On further questioning, she says that the pain occurs bilaterally in the temporal and occipital areas. She denies jaw claudication but does report scalp tenderness. She has had no vision loss. On physical examination, she has pulseless, nontender temporal arteries.

How do her symptoms influence your decision to pursue a diagnosis of temporal arteritis? Do the physical findings help with your assessment? What is her probability of having temporal arteritis and a positive temporal artery biopsy result? How much will an erythrocyte sedimentation rate (ESR) change the likelihood of temporal arteritis?

UPDATED SUMMARY ON TEMPORAL ARTERITIS

Original Review

Smetana GW, Shmerling RH. Does this patient have temporal arteritis? *JAMA*. 2002;287(1):92-101.

UPDATED LITERATURE SEARCH

We performed an updated MEDLINE literature search from January 2000 to August 2004, using the same search strategy as in our original publication. After reviewing the titles and abstracts, we identified 48 potential new relevant articles. We included studies with an emphasis on commonly available laboratory tests because the interpretation of the results is always tightly coupled to the clinical evaluation for temporal arteritis. We then applied the inclusion and exclusion criteria of our original review. After a detailed review of each retrieved article, 5 articles met the inclusion criteria. Most excluded studies contained no detailed clinical information about historical or physical examination features or failed to require that at least 90% of patients in the temporal arteritis group have a positive temporal artery biopsy result. An additional 10 articles did

not meet our criteria but contained useful background material for our discussion.

NEW FINDINGS

- Combinations of clinical findings are much more powerful at assessing the likelihood of temporal arteritis than individual findings, especially jaw claudication with vision change.
- The diagnostic value of an increased ESR increases with increasing patient age.

Details of the Update

Since publication of our original review, 5 additional studies have provided evidence on the value of the clinical examination in predicting temporal artery biopsy results among patients suspected of having the disease.

Mirroring decision-making in clinical practice, combinations of clinical findings create more clinically important changes in the likelihood of temporal arteritis than individual findings. For example, in a study by Younge et al,¹ the combination of jaw claudication and decreased vision was associated with a positive likelihood ratio (LR+) of 44, whereas either finding alone had much lower LR+s. A multivariate model allows better assessment of the ESR in relation to the patient's age and symptoms. For example, a new headache but normal ESR in a 72-year-old patient is associated with a risk of disease of 12%, but the probability of temporal arteritis increases to 78% when jaw claudication and scalp tenderness occur together. Thus, despite a low likelihood ratio (LR) for a normal ESR (LR, 0.20), a normal ESR can be outweighed by the presence of other factors.

A platelet count greater than $400 \times 10^3/\mu\text{L}$ increases the probability of a positive temporal artery biopsy result among patients suspected of having the disease. In a study in which two-thirds of 91 patients reported vision symptoms, the LR+ was 6.3 (confidence interval [CI], 2.4-17) for platelet count greater than $400 \times 10^3/\mu\text{L}$.² However, a second study, using a lower threshold for the platelet count, found an LR+ of 1.6 (CI, 1.3-1.9). We did not report results of platelet counts in our data abstraction in our original systematic review, because it was not possible to construct LRs from the primary data. However, a multivariate model³ revealed that the platelet count did not

Table 49-7 Summary Likelihood Ratios of Symptoms, Signs, and the Erythrocyte Sedimentation Rate for Temporal Arteritis

Finding (No. of Studies)	LR+ (95% CI)	LR- (95% CI)
Jaw claudication (17)	4.3 (3.0-6.1)	0.72 (0.66-0.79)
Diplopia (5)	3.5 (1.8-6.8)	0.96 (0.93-0.99)
Scalp tenderness (8)	1.7 (1.1-2.4)	0.73 (0.66-0.82)
Any headache (19)	1.7 (1.5-1.9)	0.67 (0.56-0.80)
Any vision symptoms (19)	1.1 (0.94-1.3)	0.97 (0.92-1.0)
"Abnormal" ESR ^a (7)	1.1 (1.0-1.2)	0.2 (0.08-0.51)
ESR > 100 (4)	1.9 (1.1-3.3)	
ESR 50-100 (5)	1.1 (0.87-1.5)	
ESR < 50 (5)	0.55 (0.38-0.80)	

Abbreviations: CI, confidence interval; ESR, erythrocyte sedimentation rate; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

^aAn abnormal ESR was defined by the laboratory analyses of the individual studies. From these data, a normal ESR has a likelihood ratio of 0.2 for temporal arteritis.

add additional information to the other variables. Future studies should reassess the role of the platelet count as a screening test for temporal arteritis among patients with compatible symptoms, especially those with vision complaints.

One retrospective review⁴ assessed the ethnic background among patients with biopsy-proven temporal arteritis. None of the 40 Hispanic patients in the United States referred for temporal artery biopsy had positive results. A study from a tertiary hospital in Spain⁵ showed very few differences between these patients compared to the summary data from the original Rational Clinical Examination article. A smaller study of patients in the United Kingdom⁶ also showed values similar to those in the original Rational Clinical Examination article. The low prevalence in this population should be studied in future case series.

The strict inclusion criteria of our original review required primary data from clinical series, excluding decision analyses from consideration because decision analyses require assumptions about the prevalence of disease and differing clinical features. The published decision analyses preceding our review did not have access to a systematic estimation of these values. However, they provide an alternative strategy about manage-

Box 49-1 Temporal Arteritis Score (for Patients ≥ 50 y)

$$\text{Score} = -240 + 48 \times (\text{headache}) + 108 \times (\text{jaw claudication}) + 56 \times (\text{scalp tenderness}) + 1.0 \times (\text{ESR}) + 70 \times (\text{ischemic optic neuropathy}) + 1.0 \times (\text{age})$$

(If symptom present, substitute 1.0; if negative, substitute 0)

$$\text{Estimated probability} = [\exp^{(\text{score}/50)}]/[1 + \exp^{(\text{score}/50)}]$$

If score less than -110, low risk (<10% chance of positive biopsy result)

If score = -110 to 70, intermediate risk (10%-80% chance of positive biopsy result)

If score > 70, high risk (>80% chance of positive biopsy result)

ment of patients suspected of having temporal arteritis. We identified 3 such studies⁷⁻⁹ through our literature search. Not surprisingly, using different assumptions, these authors developed differing predictive models. Each study modeled empiric treatment strategies, treatment guided by biopsy results, and treatment of all patients irrespective of biopsy results. The model results changed with differing estimated prior probability of disease. None of these studies, however, estimated the influence of particular clinical or laboratory features on the likelihood of positive biopsy results. Therefore, these provide a complementary analysis but do not add to the information in our review or update on The Rational Clinical Examination.

IMPROVEMENTS IN THE DATA PRESENTED IN THE ORIGINAL PUBLICATION

New data allowed us to refine our summary estimates for the LRs of clinical features for temporal arteritis. None of the estimates changed appreciably, although the new data generally led to narrower CIs and, therefore, more confidence in the role of each finding.

CHANGES IN THE REFERENCE STANDARD

The reference standard for the diagnosis of temporal arteritis remains a temporal artery biopsy.

RESULTS OF LITERATURE REVIEW

Table 49-7 details univariate analyses of clinical variables associated with temporal arteritis. As in the original meta-analysis, the presence of jaw claudication or diplopia was associated with the highest LRs. For decreasing the likelihood of temporal arteritis, a normal ESR has the lowest LR.

Multivariate Findings for Temporal Arteritis

Younge et al¹ developed a temporal arteritis score, shown in **Box 49-1**, that estimates the probability of temporal arteritis according to the presence of 6 factors.

The authors derived this score from a large sample of 1113 patients undergoing temporal artery biopsy, all of whom were older than 50 years. This is the largest series in the literature that includes patients undergoing temporal artery biopsy with both positive and negative biopsy results (the entire literature from 1966 to 2000 includes only 2680 patients). We were unable to determine the value of combinations of clinical features in our original review because of the limitations of the meta-analytic design and the lack of individual patient specific data. The temporal arteritis score of Younge et al¹ is an important contribution that assists clinicians in estimating the likelihood of temporal arteritis among patients suspected of having the disease. However, it was derived from a group of patients who were older than 50 years, and its use should be limited to people of similar age. Prospective validation studies are necessary, but the large

Box 49-2 American College of Rheumatology Criteria¹⁰ for Temporal Arteritis

1. Age at disease onset at least 50 y
2. New headache
3. Temporal artery abnormality (tenderness, diminished pulsation unrelated to atherosclerosis of cervical arteries)
4. Increased erythrocyte sedimentation rate (at least 50 mm/h by Westergren method)
5. Abnormal artery biopsy result (vasculitis with mononuclear cell predominance or granulomatous inflammation, usually with multinucleated giant cells)

patient sample provides some reassurance to clinicians who choose to apply the score to their patients.

EVIDENCE FROM GUIDELINES

There are no well-established consensus guidelines for the evaluation, diagnosis, or treatment of patients with suspected or proven temporal arteritis. Clinicians and researchers generally agree on the American College of Rheumatology (ACR) criteria for the classification of giant-cell (temporal) arteritis.¹⁰ These criteria were described as “classification” criteria (rather than “diagnostic”) to make their purpose clear: they are best used among patients with vasculitis to improve standardization and comparability of studies, not necessarily as diagnostic criteria for clinical practice. They are reproduced in **Box 49-2**.

The ACR criteria for all vasculitis syndromes, including temporal arteritis, have been criticized for poor predictive value when applied to individual patients in clinical practice.¹¹ However, other guidelines have not been widely accepted.¹² All of these guidelines use clinical factors presented in the original and updated literature reviews. Because there is no clear consensus about the definition or gold standard for the diagnosis of temporal arteritis beyond a positive temporal artery biopsy result, in our meta-analysis we required at least 90% of individuals considered to have the disease to have histologic “proof.”

CLINICAL SCENARIO—RESOLUTION

Our 72-year-old woman has a new onset of temporal and occipital headache that raises the possibility of temporal arteritis. One should seek the presence of those features that confer a high LR+, including diplopia and jaw claudication. In her case, scalp tenderness is present (LR+, 1.7), but she does not have other historical features that confer a high LR+. On examination, one looks for the presence of beaded, tender, or pulseless temporal arteries. Her pulseless temporal arteries confer an LR+ of 2.7, but the CI around this result is broad (95% CI, 0.55–13).

An ESR measurement would be helpful: a normal ESR confers an LR of 0.2, whereas an elevated ESR greater than 100 mm/h increases the likelihood of disease (LR, 1.9). Intermediate ESR values, that is, values that are elevated but less than 100 mm/h, occur commonly in patients with temporal arteritis and would increase the likelihood to a lesser degree.

The temporal artery score of Younge et al¹ provides an alternate strategy for estimating disease risk by combining the most important clinical features. If we enter the data for our patient into this prediction rule, using hypothetical ESR values of 50 and 100, we obtain the following results.

$$\begin{aligned} \text{For ESR} = 50: \text{Score} &= -240 + 48 \times (\text{headache} = 1) \\ &+ 108 \times (\text{jaw claudication} = 0) + 56 \times (\text{scalp tenderness} = 1) \\ &+ 1.0 \times (\text{ESR} = 50) + 70 \times (\text{ischemic optic neuropathy} = 0) \\ &+ 1.0 \times (\text{age} = 72) \end{aligned}$$

Score = -14. Intermediate risk (probability, 43%)

$$\begin{aligned} \text{For ESR} = 100: \text{Score} &= -240 + 48 \times (\text{headache} = 1) \\ &+ 108 \times (\text{jaw claudication} = 0) + 56 \times (\text{scalp tenderness} = 1) \\ &+ 1.0 \times (\text{ESR} = 100) + 70 \times (\text{ischemic optic neuropathy} = 0) \\ &+ 1.0 \times (\text{age} = 72) \end{aligned}$$

Score = 36. Intermediate risk (probability, 67%)

In this case, using the prediction rule of Young et al,¹ the risk is intermediate according to clinical evaluation. The ESR results do not modify the likelihood of temporal arteritis, as determined by clinical evaluation alone. After this evaluation, temporal arteritis is still a consideration. Previous studies and clinical experience suggest that biopsy should be performed in 7 to 10 days, although the yield of biopsy decreases over time after the initiation of corticosteroid treatment.

TEMPORAL ARTERITIS—MAKE THE DIAGNOSIS

PRIOR PROBABILITY

Temporal arteritis is relatively rare, though the disease may be underdiagnosed.¹³ The prevalence increases with age, and it occurs more commonly among women and whites. One study found that, among white persons 50 years and older, the prevalence of temporal arteritis was 200 cases per 100 000; among persons older than 85 years, the prevalence was 1100 per 100 000.¹⁴ Most published series have been from northern Europe and the northern United States, but the disease has been observed worldwide.

POPULATION FOR WHOM TEMPORAL ARTERITIS DISEASE SHOULD BE CONSIDERED

Temporal arteritis should be considered in all adults aged 50 years and older with appropriate symptoms. Although prevalence varies by sex, race, and geographic locale, no single demographic factor among persons older than 50 years decreases the likelihood enough to exclude the diagnosis.

DETECTING THE LIKELIHOOD OF TEMPORAL ARTERITIS

One can estimate the likelihood of temporal arteritis by using either single features (and applying the summary LRs from our meta-analysis) or by using combinations of features, as established by the prediction rule of Younge et al¹ (see Table 49-8).

REFERENCE STANDARD TESTS

Temporal artery biopsy and histologic evaluation is the reference standard for the diagnosis of temporal arteritis. Other means of diagnosis have been suggested, including positron emission tomography scanning^{15,16} and ultrasonography¹⁷⁻²² for imaging of the temporal artery. Although results of small studies have been promising, studies of these tests have been flawed (primarily by incomplete evaluation against the gold standard, temporal artery biopsy) and are not widely accepted. Although they could at some point prove

Table 49-8 The Single Best Findings or Combinations of Findings Can Be Used to Estimate the Probability of Temporal Arteritis

	LR+ (95% CI)	LR- (95% CI)
Single Best Findings Suggesting the Presence of Temporal Arteritis		
Jaw claudication	4.3 (3.0-6.1)	
Diplopia	3.5 (1.8-6.8)	
Single Best Finding Suggesting the Absence of Temporal Arteritis		
ESR < 50 mm/h (n = 5)		0.55 (0.38-0.80)
Combinations of Findings^a		
Headache + jaw claudication + scalp tenderness at age 60 y		65
Headache + jaw claudication + scalp tenderness at age 80 y		74
Headache + jaw claudication + scalp tenderness at age 60 y, ESR = 50 mm/h		84
Headache + jaw claudication + scalp tenderness at age 80 y, ESR = 50 mm/h		88
No headache + no jaw claudication + no scalp tenderness at age 60 y, ESR = 50 mm/h		7
No headache + no jaw claudication + no scalp tenderness at age 80 y, ESR = 50 mm/h		10

Abbreviations: CI, confidence interval; ESR, erythrocyte sedimentation rate; LR+, positive likelihood ratio; LR-, negative likelihood ratio; OR, odds ratio.

^aThese are examples of various combinations of findings for patients with 3 of 3 symptoms vs 0 of 3 symptoms present at various ages. The addition of age and ESR provides important information when combined with the symptoms.

diagnostically useful in the diagnosis of temporal arteritis, studies to date have not provided sufficient, conclusive evidence confirming the diagnostic value of these tests beyond standard clinical information (including medical history, physical examination, and routine measures of inflammation) and biopsy as alternative reference standards. Magnetic resonance angiography, computed tomography, or standard angiography can be helpful for extracranial disease, including inflammatory involvement of the aorta or its proximal branches.²²

REFERENCES FOR THE UPDATE

- Younge BR, Cook BE Jr, Bartley GB, Hodge DO, Hunder GG. Initiation of glucocorticoid therapy: before or after temporal artery biopsy? *Mayo Clin Proc.* 2004;79(4):483-491.^a
- Foroozan R, Danesh-Meyer H, Savino PJ, Gamble G, Mekari-Sabbagh ON, Sergott RC. Thrombocytosis in patients with biopsy-proven giant cell arteritis. *Ophthalmology.* 2002;109(7):1267-1271.^a
- Gabriel SE, O'Fallon M, Achkar AA, et al. The use of clinical characteristics to predict the results of temporal artery biopsy among patients with suspected giant cell arteritis. *J Rheumatol.* 1995;22(1):93-96.
- Liu NH, LaBree LD, Felson SE, Rao NA. The epidemiology of giant cell arteritis: a 12-year retrospective review. *Ophthalmology.* 2001;108(6):1145-1149.^a
- Gonzalez-Gay MA, Garcia-Porrúa C, Amor-Dorado JC, Llorca J. Influence of age, sex, and place of residence on clinical expression of giant-cell arteritis in Northwest Spain. *J Rheumatol.* 2003;30(7):1548-1551.^a
- Mohamed MS, Bates T. Predictive clinical and laboratory factors in the diagnosis of temporal arteritis. *Ann R Coll Surg.* 2002;84(1):7-9.^a
- Elliot DL, Watts WJ, Reuler JB. Management of suspected temporal arteritis: a decision model. *Med Decis Making.* 1983;3(5):63-68.
- Buchbinder R, Detsky AS. Management of suspected giant cell arteritis: a decision analysis. *J Rheumatol.* 1992;19(8):1220-1228.
- Nadeau SE. Temporal arteritis: a decision-analytic approach to temporal artery biopsy. *Acta Neurol Scand.* 1988;78(2):90-100.
- Hunder GG, Bloch DA, Michel BA, et al. The American College of Rheumatology 1990 criteria for the classification of giant cell arteritis. *Arthritis Rheum.* 1990;33(8):1122-1128.
- Rao JK, Allen NB, Pincus T. Limitations of the 1990 American College of Rheumatology classification criteria in the diagnosis of vasculitis. *Ann Intern Med.* 1998;129(5):345-352.
- Frearson R, Cassidy T, Newton J. Polymyalgia rheumatica and temporal arteritis: evidence and guidelines for diagnosis and management in older people. *Age Ageing.* 2003;32(4):370-374.
- Ostberg G. Temporal arteritis in a large necropsy series. *Ann Rheum Dis.* 1971;30(3):224-235.
- Lawrence RC, Helmick CG, Arnett FC, et al. Estimates of the prevalence of arthritis and selected musculoskeletal disorders in the United States. *Arthritis Rheum.* 1998;41(5):778-799.

15. Brodmann M, Lipp RW, Passath A, Seinost G, Pabst E, Pilger E. The role of 2-18F-fluoro-2-deoxy-D-glucose positron emission tomography in the diagnosis of giant cell arteritis of the temporal arteries. *Rheumatology (Oxford)*. 2004;43(2):241-242.
16. Turlakow A, Yeung HW, Pui J, et al. Fludeoxyglucose positron emission tomography in the diagnosis of giant cell arteritis. *Arch Intern Med*. 2001;161(7):1003-1007.
17. LeSar CJ, Meier GH, DeMasi RJ, et al. The utility of color duplex ultrasonography in the diagnosis of temporal arteritis. *J Vasc Surg*. 2002;36(6):1154-1160.
18. Murgatroyd H, Nimmo M, Evans A, MacEwen C. The use of ultrasound as an aid in the diagnosis of giant cell arteritis: a pilot study comparing histological features with ultrasound findings. *Eye*. 2003;17(3):415-419.
19. Nesher G, Shemesh D, Mates M, Sonnenblick M, Abramowitz HB. The predictive value of the halo sign in color Doppler ultrasonography of the temporal arteries for diagnosing giant cell arteritis. *J Rheumatol*. 2002; 29(6):1224-1226.
20. Pfadenhauer K, Weber H. Duplex sonography of the temporal and occipital artery in the diagnosis of temporal arteritis: a prospective study. *J Rheumatol*. 2003;30(10):2177-2181.
21. Salvarani C, Silingardi M, Ghirarduzzi A, et al. Is duplex ultrasonography useful for the diagnosis of giant-cell arteritis? *Ann Intern Med*. 2002;137(4):232-238.
22. Stanson AW. Imaging findings in extracranial (giant cell) temporal arteritis. *Clin Exp Rheumatol*. 2000;18(4 suppl 20):S43-S48.

^aFor the Evidence to Support the Update for this topic,
see <http://www.JAMAevidence.com>.

This page intentionally left blank

EVIDENCE TO SUPPORT THE UPDATE:

49

Temporal Arteritis

TITLE Thrombocytosis in Patients With Biopsy-Proven Giant Cell Arteritis.

AUTHORS Foroozan R, Danesh-Meyer H, Savino PJ, Gamble G, Mekari-Sabbagh ON, Sergott RC.

CITATION *Ophthalmology*. 2002;109(7):1267-1271.

QUESTION Are the complete blood cell (CBC) count and erythrocyte sedimentation rate (ESR) useful in predicting positive temporal artery biopsy results among patients suspected of having giant-cell arteritis (GCA)?

DESIGN Retrospective, case-control series.

SETTING Specialty eye hospital in Philadelphia, Pennsylvania.

PATIENTS Ninety-one consecutive patients undergoing temporal artery biopsy for suspicion of GCA; biopsy performed within 1 week of presentation. Corticosteroid therapy before biopsy was not allowed; blood tests were conducted within 24 hours of biopsy.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

Diagnostic (gold) standard was temporal artery biopsy; tests included CBC count and Westergren ESR. Definition of elevated platelet levels ($>400 \times 10^3/\mu\text{L}$) was based on reference range greater than 2 SD above the mean; elevated ESR was above age/2 for men and (age + 10)/2 for women. No patients had a clinical course to suggest biopsy-negative GCA.

MAIN OUTCOME MEASURES

Means, sensitivity, specificity, and likelihood ratios (LRs).

MAIN RESULTS

Forty-seven patients had a positive biopsy result; 44 had negative biopsy result.

White blood cell counts were no different between patients with positive and negative biopsy results, although patients

with positive biopsy results were significantly more anemic. Among patients suspected of having temporal arteritis, thrombocytosis significantly predicts the likelihood of a positive temporal artery biopsy result (see Tables 49-9 and 49-10).

CONCLUSION

LEVEL OF EVIDENCE Level 3 (using criteria from original review).

STRENGTHS The investigators asked a unique question regarding the value of laboratory testing to stratify probability of disease.

LIMITATIONS All patients were evaluated at a subspecialty ophthalmology clinic. The sample size was small.

Table 49-9 Comparison of Laboratory Values Between Those With Positive vs Negative Biopsy Results for Giant-Cell Arteritis

Test	Biopsy Result Positive	Biopsy Result Negative	P Value
Mean ESR level, mm/h	82	70	.12
Mean hematocrit level, %	34.8	37	.03
Mean hemoglobin level, g/dL	11.7	12.5	.01
Mean platelet count, $\times 10^3/\mu\text{L}$	433	277	<.001

Abbreviation: ESR, erythrocyte sedimentation rate.

Table 49-10 Likelihood Ratios of Laboratory Values for Giant-Cell Arteritis

Test	Sensitivity	Specificity	LR+ (95% CI)	LR- (95% CI)
ESR	0.79	0.27	1.1 (0.86-1.4)	0.78 (0.37-1.6)
Platelet count $> 400 \times 10^3/\mu\text{L}$	0.57	0.91	6.3 (2.4-17)	0.47 (0.33-0.66)
Combination of ESR and platelet count $> 400 \times 10^3/\mu\text{L}$	0.51	0.91	5.6 (2.1-15)	0.54 (0.40-0.73)

Abbreviations: CI, confidence interval; ESR, erythrocyte sedimentation rate; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

Commentary

This study was performed with high quality, although it was retrospective and selected patients were treated at a specialty eye hospital. Two-thirds of the patients had primarily visual complaints. The results suggest that elevated platelet count may be useful in suggesting the diagnosis of GCA, but LRs may not be helpful enough to preclude biopsy or rule out the need for one. Also, the marginal value of elevated platelet count beyond elements of the medical history, physical examination, and other routine laboratory tests (especially lack of normal ESR) may be small. The authors suggest that platelet count may be better than ESR in predicting results of biopsy, in part because an elevation in ESR is part of what goes into the decision to get a biopsy. However, the definition of elevated ESR (age and sex adjusted) was more restrictive in this study than in many others and may have lessened its predictive power. This study does not examine the value of history-taking or physical examination findings.

Reviewed by Robert H. Shmerling, MD

residence. The only clinical sex-based difference is a higher prevalence of polymyalgia rheumatica in women (see Table 49-11). Women had a statistically significantly lower hemoglobin level than men. No clinical features differed for urban- or rural-dwelling patients. Age of onset at presentation did not significantly influence the clinical presentation. A trend existed toward more polymyalgia rheumatica in younger patients, but this difference was not significant. Hemoglobin levels were minimally lower in younger patients, and more older patients had an increased alkaline phosphatase level.

CONCLUSIONS

LEVEL OF EVIDENCE Level 4 (using criteria from the original review).

STRENGTHS Consistent data set across all patients.

LIMITATIONS The study population consisted only of patients with giant-cell arteritis. The relatively small

TITLE Influence of Age, Sex, and Place of Residence on Clinical Expression of Giant-Cell Arteritis in Northwest Spain.

AUTHORS Gonzalez-Gay MA, Garcia-Porrúa C, Amor-Dorado JC, Llorca J.

CITATION *J Rheumatol.* 2003;30(7):1548-1551.

QUESTION Do age, sex, and urban residence influence the clinical expression of giant-cell arteritis?

DESIGN Retrospective chart review.

SETTING Tertiary referral hospital in northwestern Spain that is the only referral center for a mixed urban and rural area encompassing approximately 250000 people.

PATIENTS All patients with biopsy-proven giant-cell arteritis between 1981 and 2001.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

Clinical and laboratory features of patients with biopsy-proven giant-cell arteritis represented the diagnostic tests. The diagnostic standard was a positive temporal artery biopsy.

MAIN OUTCOME MEASURES

The main outcome measure was sensitivity.

MAIN RESULTS

Few differences exist in the clinical presentation of biopsy-proven giant-cell arteritis according to age, sex, and place of

Table 49-11 Most Presenting Features of Giant-Cell Arteritis Are Similar Between Men vs Women and Patients Younger Than 70 Years vs Older Than 70 Years

Variable	Men (n = 97)	Women (n = 113)	Onset <70 y of Age (n = 42)	Onset >70 y of Age (n = 168)
Men, %			48	46
Age at diagnosis, y	75	75		
Living in urban area, %	27 ^a	46 ^a	31	39
Delay to diagnosis, wk	9.7	11	12	9.9
Headache, %	90	85	88	87
Scalp tenderness, %	34	34	26	36
Constitutional syndrome, %	67	62	76	61
Abnormal temporal artery examination, %	73	78	67	77
Jaw claudication, %	36	45	29	44
Dysphagia, %	3	7	0	7
Polymyalgia rheumatica, %	33 ^a	49 ^a	52	39
Fever, %	8	11	12	9
Visual manifestations, %	26	21	21	24
Permanent visual loss, %	13	12	12	13
Cerebrovascular accident, %	3	1	5	1
Limb claudication of recent onset, %	4	2	7	2
ESR, mean, mm/h	91	95	100	92
Hemoglobin, mean, g/dL	12.2 ^a	11.4 ^a	11.3 ^a	11.9 ^a
Platelet count, mean, $\times 10^3/\mu\text{L}$	407	412	437	402
Increased alkaline phosphatase, %	26	28	48 ^a	22 ^a

Abbreviation: ESR, erythrocyte sedimentation rate.

^aP < .05 for comparison between men and women or between younger and older patients.

number of study subjects limited the power to detect significant differences.

Commentary

This case series provides a detailed summary of clinical and laboratory features among a cohort of patients with biopsy-proven giant-cell arteritis in Spain. The overall prevalence of specific features is similar to that reported in our original review and meta-analysis. Differences include higher incidences of headache and polymyalgia rheumatica and lower incidences of fever and visual manifestations than in our original review. In this study, the authors aimed to identify differences in clinical presentations according to age, sex, and urban location. Remarkably, nearly all features were similar across these patient subsets. The only clinical feature that was statistically significantly different across nearly 60 comparisons was the greater incidence of polymyalgia rheumatica among women compared with men. However, this series may have lacked sufficient statistical power to detect significant differences.

Small differences in hemoglobin and the incidence of elevated alkaline phosphatase level existed in these comparisons, but these are not clinically significant. We have previously shown that anemia does not predict positive biopsy results among patients suspected of having the disease (positive likelihood ratio, 1.5 [95% confidence interval, 0.82-2.9]; negative likelihood ratio, 0.79 [95% confidence interval, 0.6-1.0]). This study suggests that clinical suspicion and the value of particular clinical features of giant-cell arteritis do not differ among these selected patient subsets.

Reviewed by Gerald W. Smetana, MD

TITLE The Epidemiology of Giant Cell Arteritis: A 12-Year Retrospective Review.

AUTHORS Liu NH, LaBree LD, Feldon SE, Rao NA.

CITATION *Ophthalmology*. 2001;108(6):1145-1149.

QUESTION What is the incidence of biopsy-proven giant-cell arteritis among individuals of Hispanic descent?

DESIGN Retrospective chart review.

SETTING Subspecialty academic ophthalmology institute in the United States.

PATIENTS Sequential patients ($n = 121$) undergoing temporal artery biopsy.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

The diagnostic tests were demographic factors including age, sex, and ethnicity. The diagnostic standard was a temporal artery biopsy. The authors explicitly stated the pathologic criteria used to classify a temporal artery biopsy result as positive.

Table 49-12 The Incidence of Giant-Cell Arteritis Differs by Race

Race (No.)	Positive Biopsy Result, %	OR (95% CI)
White (66)	40	22 (3.6-133)
Black (6)	0	0 (0-4.2)
Hispanic (40)	0	0 (0-0.38)
Asian (9)	12	0.61 (0.09-4.0)

Abbreviations: CI, confidence interval; OR, odds ratio.

MAIN OUTCOME MEASURES

Incidence of temporal arteritis among white, Asian, black, and Hispanic patients undergoing temporal artery biopsy. Hispanic patients self-reported whether they considered themselves to be of white or Latino descent.

MAIN RESULTS

Twenty patients (16.5%) had positive temporal artery biopsy results. The mean age of the study population was 70 ± 8.8 years. White patients were older than Asian, black, and Hispanic patients. The mean age for patients with a positive biopsy result was 75 years, whereas that for patients with a negative biopsy result was 69 years. Giant-cell arteritis is rare among a population of Americans of Hispanic ethnicity (Table 49-12).

CONCLUSIONS

LEVEL OF EVIDENCE Level 1 (using criteria from the original review).

STRENGTHS Asked a unique question not previously addressed in the literature.

LIMITATIONS No clinical information was recorded and only demographic and laboratory variables were studied.

Commentary

The original review reconfirmed the observation that temporal arteritis is predominantly a disease of whites. Among all eligible studies in that review, 86% of all patients with positive biopsy results were white. Descriptions of blacks with temporal arteritis have been largely restricted to case reports and small series. The incidence among US Hispanics has not been well studied. In this report, the authors determined the race of all patients undergoing temporal artery biopsy at a referral ophthalmology center in Los Angeles, California. Although Hispanics constituted 33% of all patients referred for biopsy, not a single biopsy result was positive in this group of patients (95% confidence interval, 0%-7.2%).

Reviewed by Gerald W. Smetana, MD

TITLE Predictive Clinical and Laboratory Factors in the Diagnosis of Temporal Arteritis.

AUTHORS Mohamed MS, Bates T.

CITATION *Ann R Coll Surg.* 2002;84(1):7-9.

QUESTION Among patients undergoing temporal artery biopsy, which clinical and laboratory factors predict positive biopsy results?

DESIGN Retrospective chart review.

SETTING Single hospital in the United Kingdom.

PATIENTS All patients ($n = 50$) who underwent temporal artery biopsy between January 1988 and December 1997.

DESCRIPTION OF THE TEST AND DIAGNOSTIC STANDARD

The diagnostic tests were demographic features, presenting clinical features, laboratory investigation, and the duration of corticosteroid therapy before biopsy. The diagnostic standard was a temporal artery biopsy. The authors did not state the criteria used to determine whether a temporal artery biopsy result was positive.

MAIN OUTCOME MEASURES

The main outcome measures were sensitivity and specificity.

MAIN RESULTS

Seventeen patients had temporal arteritis and 33 patients had a normal biopsy result. The mean age was 73 years (range, 60-82 years) for patients with a positive biopsy result and 67 years (range, 49-85 years) for those with a negative biopsy result. The mean durations of steroid therapy for patients with positive and negative biopsy results were 7 and 10 days, respectively. The mean erythrocyte sedimentation rate (ESR) was 56 mm/h for patients with a positive biopsy result and 38 mm/h for those with a negative biopsy result. Seventeen patients (34%) had a positive temporal artery biopsy result (Table 49-13).

Among clinical and laboratory features in a population of 50 patients suspected of having temporal arteritis, an ESR less than 50 mm/h decreased the likelihood of temporal arteritis, whereas an ESR of 50 to 100 mm/h increased the likelihood of temporal arteritis. All other results had a 95% confidence interval that included 1.

CONCLUSIONS

LEVEL OF EVIDENCE Level 1 (using criteria from the original review).

Table 49-13 Likelihood Ratios of Demographic Variables, Symptoms, Signs, and Laboratory Values for Temporal Arteritis (Disease Frequency 17/50)

Feature (No. With Feature)	Sensitivity	Specificity	LR+ (95% CI)	LR- (95% CI)
Jaw pain (6)	0.24	0.94	3.9 (0.79-19)	0.81 (0.61-1.1)
History of fever (4)	0.12	0.94	1.9 (0.29-12)	0.94 (0.77-1.1)
Polymyalgia rheumatica (4)	0.12	0.94	1.9 (0.29-12)	0.94 (0.77-1.1)
Male sex (15)	0.35	0.73	1.5 (0.65-3.4)	0.83 (0.52-1.3)
Neurologic symptoms (21)	0.71	0.42	1.2 (0.79-1.8)	0.69 (0.30-1.6)
Steroid use before biopsy (31)	0.71	0.42	1.2 (0.79-1.8)	0.69 (0.30-1.6)
Headache (44)	0.88	0.12	1.0 (0.81-1.2)	0.97 (0.2-4.8)
Temporal tenderness (36)	0.65	0.24	0.9 (0.60-1.3)	1.5 (0.6-3.5)
Visual symptoms (21)	0.24	0.48	0.5 (0.2-1.2)	1.6 (1.0-2.4)
Ocular signs (8)	0.06	0.79	0.3 (0.04-2.2)	1.2 (0.96-1.5)
ESR				
>100 mm/h (2)			1.9 (0.21-18)	
50-100 mm/h (21)			2.1 (1.1-4.0)	
<50 mm/h (27)			0.44 (0.19-0.86)	

Abbreviations: CI, confidence interval; ESR, erythrocyte sedimentation rate; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

STRENGTHS Standardized data set for all patients.

LIMITATIONS The small sample size resulted in broad confidence intervals for the likelihood ratios.

Commentary

Only the ESR was a significant predictor of disease, but low statistical power limits the conclusions for other findings. The authors studied several factors that proved significant in our original review and meta-analysis but which failed to predict biopsy results. This study illustrates the value of meta-analytic techniques that allow estimates of the operating characteristics of diagnostic tests based on larger samples than available in any individual study.

Reviewed by Gerald W. Smetana, MD

TITLE Initiation of Glucocorticoid Therapy: Before or After Temporal Artery Biopsy.

AUTHORS Younge BR, Cook BE, Bartley GB, Hodge DO, Hunder GG.

CITATION Mayo Clin Proc. 2004;79(4):483-491.

QUESTIONS Do clinical features exist among patients with suspected giant-cell arteritis (GCA) that may help clinicians decide when to initiate glucocorticoid therapy? When is the likelihood of positive biopsy result so high that therapy should begin right away? When is the likelihood low enough to defer treatment until after biopsy results are available?

DESIGN Retrospective, case-control series (cases had positive biopsy results; controls had negative biopsy results).

SETTING Mayo Clinic, Rochester, Minnesota.

PATIENTS One thousand one hundred thirteen sequential patients, identified through the Mayo Surgical Index, undergoing temporal artery biopsy between January 1988 and December 1997. Twenty percent of the patients were receiving oral corticosteroids at the biopsy.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

Diagnostic (gold) standard was temporal artery biopsy; the authors collected multiple clinical features (by medical history, physical examination, and laboratory studies). Standard Mayo Clinic reference ranges for laboratory values were used, including erythrocyte sedimentation rate (ESR) of 0 to 22 mm/h for men and 0 to 29 mm/h for women.

MAIN OUTCOME MEASURES

Sensitivity, specificity, and predictive values of various clinical and laboratory findings with respect to biopsy results were calculated.

MAIN RESULTS

- Three hundred seventy-three patients had positive biopsy results (33.5%); 740 (66.5%) had negative biopsy results.
- The commonly taught combination of headache with ESR had a likelihood ratio (LR) of 2.4 (95% confidence interval [CI], 2.1-2.7) when the ESR was elevated. When neither a headache nor ESR abnormality was present, the LR for temporal arteritis was 0.42 (95% CI, 0.36-0.49).

Clinical findings (LRs and CIs are calculated from data provided in the article) are shown in Table 49-14.

Laboratory findings in patients not receiving oral corticosteroid treatment (LRs and CIs are calculated from data provided in the article) are shown in Table 49-15.

A decision rule was developed from a multivariate model:

$$\text{Temporal arteritis score} = -240 + 48 \times (\text{headache}) + 108 \times (\text{jaw claudication}) + 56 \times (\text{scalp tenderness}) + 1.0 \times (\text{ESR}) + 70 \times (\text{ischemic optic neuropathy}) + 1.0 \times (\text{age})$$

(If symptom present, substitute 1; if negative, 0)

$$\text{Estimated probability} = [\exp^{(\text{score}/50)}] / [1 + \exp^{(\text{score}/50)}]$$

If score < -110, low risk (<10% chance of positive biopsy).

If score = -110 to 70, intermediate risk (10%-80% chance of positive biopsy result).

If score > 70, high risk (>80% chance of positive biopsy result).

The model was validated with prospective data on 289 patients; 86% of the high-risk patients had a positive biopsy result, whereas 9% of the low-risk patients had a positive biopsy result.

Table 49-14 Likelihood Ratios for Single Symptoms and in Combination for Temporal Arteritis

Test/Feature	Sensitivity	Specificity	LR+ (95% CI)	LR- (95% CI)
Single Features				
Jaw claudication	0.40	0.94	6.9 (5.0-9.5)	0.64 (0.59-0.7)
Diplopia	0.04	0.99	3.7 (1.5-9.2)	0.97 (0.95-0.99)
Scalp tenderness	0.33	0.89	3.1 (2.4-4.0)	0.75 (0.70-0.81)
Myalgia/arthalgia	0.46	0.50	2.2 (1.6-3.1)	0.90 (0.86-0.95)
New headache	0.67	0.60	1.7 (1.5-1.9)	0.54 (0.46-0.63)
Decreased vision	0.13	0.92	1.5 (1.0-2.1)	0.95 (0.91-0.99)
Weight loss	0.24	0.81	1.3 (1.0-1.6)	0.93 (0.87-0.99)
Combination of Findings				
Jaw claudication and decreased vision	0.06	1.0	44 (5.9-322)	0.98 (0.97-0.99)
Jaw claudication and diplopia	0.02	10	30 (1.7-519)	0.98 (0.97-0.99)
New headache, jaw claudication, and scalp tenderness	0.15	0.99	19 (8.1-42)	0.86 (0.82-0.90)
Jaw claudication and scalp tenderness	0.17	0.99	18 (8.3-39)	0.84 (0.80-0.88)
New headache and jaw claudication	0.32	0.96	8.7 (5.8-13)	0.71 (0.66-0.76)
New headache and decreased vision	0.06	0.99	6.2 (2.7-14)	0.95 (0.93-0.98)
New headache and scalp tenderness	0.29	0.93	3.9 (2.9-5.3)	0.77 (0.72-0.82)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

Table 49-15 Likelihood Ratios of Laboratory Findings for Temporal Arteritis

Finding	Sensitivity	Specificity	LR+ (95% CI)	LR- (95% CI)
Abnormal platelet count	0.37	0.77	1.6 (1.3-1.9)	0.82 (0.75-0.89)
Abnormal ESR	1.0	0.16	1.2 (1.1-1.2)	0.02 (0-0.14)
Abnormal hemoglobin level	0.80	0.32	1.2 (1.1-1.3)	0.63 (0.50-0.79)

Abbreviations: CI, confidence interval; ESR, erythrocyte sedimentation rate; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

A score derived from clinical features and laboratory testing among patients suspected of having GCA can stratify patients into low, intermediate, and high likelihood of a temporal artery biopsy.

CONCLUSIONS

LEVEL OF EVIDENCE Level 1 (using criteria from the original review).

STRENGTHS The study had a large sample size, standardized data abstraction for all patients, and a temporal biopsy in all patients.

LIMITATIONS Retrospective review.

Commentary

This was a high-quality study, although it was retrospective. The results suggest that several readily available clinical fea-

tures can be combined to establish low, intermediate, and high levels of risk for positive biopsy. Strengths of this study were that the authors separately reported data for patients receiving corticosteroids before biopsy, combined clinical features (as a clinician does in actual practice), and prospectively tested the model derived from the retrospective analysis. An important limitation was the retrospective design.

For identifying patients with temporal arteritis, the data suggest that the findings of headache, jaw claudication, and scalp tenderness have some degree of independence. The independence can be inferred by noticing that multiplying the LR for the presence of each of the findings approximates the LRs when they are assessed in combination. The authors have performed a service for clinical readers by evaluating these variables in a clinical model, confirming that they have independent significance (though jaw claudication is the most important when present), and validating their results by assessing the model prospectively.

Although a normal ESR appeared to rule out disease with a univariate LR of 0.02, the model should be examined for how that finding would work when there is a strong clinical suspicion. For example, a 72-year-old man who has a new headache, but no other signs or symptoms, and an ESR of 20 mm/h would have a score of -100 and should be at low to intermediate risk (probability, 12%). As jaw claudication and scalp tenderness symptoms are added, his risk increases to 78%, even with an ESR of only 20 mm/h. If other investigators validate these data in future research, then age plus clinical findings (headache, scalp tenderness, and jaw claudication in combination) would exceed the importance of the ESR.

Reviewed by Robert H. Shmerling, MD

CHAPTER 50

Does This Patient Have an Acute Thoracic Aortic Dissection?

Michael Klompas, MD

CLINICAL SCENARIOS

CASE 1 A 64-year-old man with a history of hypertension presents to the emergency department after sudden onset of severe, anterior chest pain. On examination, he is alert but uncomfortable. His blood pressure is normal and identical in both arms. His chest is clear, and careful cardiac auscultation fails to reveal a diastolic murmur. A chest radiograph reveals a small pleural effusion but is otherwise unremarkable.

CASE 2 A 59-year-old woman is brought to the emergency department after the sudden onset of tearing chest pain. On examination, she is alert and oriented. Her blood pressure is identical in both arms. Results of her cardiac and pulmonary examinations are normal but she has a dense left-sided motor deficit. A portable chest radiograph raises the question of a widened mediastinum.

WHY IS CLINICAL EXAMINATION IMPORTANT?

A man ... was seized with a pain of the right arm and soon after of the left, ... after these there appeared a tumor on the upper part of the sternum.... He was ordered to think seriously and piously of his departure from this mortal life, which was very near at hand and inevitable.

—J. B. Morgagni, 1761¹

There is no disease more conducive to clinical humility than aneurysm of the aorta.

—Sir William Osler, c 1900²

Acute thoracic aortic dissection, one of the most common and serious diseases of the aorta, carries a high morbidity and mortality rate when it is not recognized and treated promptly. Autopsy series conducted before the era of modern treatment estimated that 40% to 50% of patients with dissection of the proximal aorta died within 48 hours.³ For those fortunate enough to survive the initial 48 hours, the disease was thought to carry a 90% 1-year mortality rate.^{3,4} Since the introduction of modern treatment regimens, the fatality rate has declined dramatically. Patients with proximal ascending dissections who rapidly undergo surgery in experienced tertiary centers have a 30-day survival rate of 80% to 85% and a 10-year survival of 55%.^{4,5} Likewise, patients with dissection of the descending aorta treated with aggressive antihypertensive therapy have a 30-day survival rate greater than 90% and a 10-year survival rate of 56%.^{4,6} Realization of the dramatic benefits of medical intervention depends on rapid establishment of the diagnosis of dissection.

Approximately 4.6 million patients per year present with chest pain to emergency departments in the United States

(8.2% of all emergency department visits).⁷ Although advanced imaging techniques can reliably establish the diagnosis of thoracic aortic dissection in high-risk populations, it is obviously inefficient, uneconomic, and unrealistic to image every patient complaining of chest pain. Indiscriminate use of diagnostic imaging in poorly chosen patients with low pretest probability of having dissection has been predicted to yield up to an 85% rate of false-positive results, depending on the imaging modality chosen.⁸ On the other hand, misdiagnosis of acute thoracic aortic dissection as unstable angina or myocardial infarction can have disastrous iatrogenic consequences should the patient receive anticoagulants or thrombolytic therapy.⁹ Physicians are therefore acutely dependent on the clinical history, examination, and chest radiograph to determine which patients require further study.

Traditionally, clinical diagnosis of thoracic aortic dissection has been inaccurate. Physicians correctly suspect the diagnosis in as few as 15% to 43% of presentations when initially evaluating patients with dissection.^{3,10,11} Diagnostic delay of more than 24 hours after hospitalization occurs in up to 39% of cases.¹² When the diagnosis is made, not infrequently it is an incidental discovery made during an advanced imaging procedure intended to assess for other diagnoses.^{13,14} Autopsies reveal the correct diagnosis is still missed in more than 10% of patients.¹³

The purpose of this review is to offer physicians an evidence-based foundation for using the clinical history, physical examination, and chest radiograph to assess the likelihood of thoracic aortic dissection.

Pathophysiology of Thoracic Aortic Dissection

The aortic wall is composed of 3 contiguous tissue layers in sequence from the vessel lumen proceeding outwards: the intima, media, and adventitia. Weakening of these tissue layers can lead to a tear in the intima, permitting the entry of blood between the intima and adventitia.¹⁵ Passage of blood into this space can extend the tear and create a so-called false lumen. The majority of these tears take place in the ascending aorta, usually in the right lateral wall where the greatest shear force on the artery wall is produced by blood expulsed from the heart under high pressure.³ The tear then extends along the greater curve of the aortic arch and down the descending aorta, though retrograde extension of the tear toward the aortic valve is also possible.¹⁵ Most aortic tears occurring beyond the ascending aorta originate immediately distal to the left subclavian artery.¹⁵ Predisposing factors for the initiation of a thoracic aortic dissection include hypertension,¹⁵ bicuspid aortic valve,¹⁵ coarctation of the aorta,¹⁵ the Marfan syndrome,¹⁶ Ehlers-Danlos syndrome,¹⁷ Turner syndrome,¹⁸ giant cell arteritis,¹⁹ third-trimester pregnancy,²⁰ cocaine abuse,²¹ trauma,²² intra-aortic catheterization,²³ and history of cardiac surgery, particularly aortic valve replacement.²⁴

The clinical features of thoracic aortic dissection are a consequence of the underlying pathophysiologic changes in the aorta. Patients perceive the initial aortic tear as sudden onset

of severe ripping or tearing chest pain. The pain is sometimes described as having a migrating quality, likely corresponding to extension of the tear along the aorta. Depending on the location of the tear and its direction of extension, patients alternately describe the pain as radiating to the neck, back, or abdomen. Occasional presentations of painless dissection have been reported, though these are usually accompanied by other findings.^{25,26}

Retrograde extension of the tear to the aortic valve can result in aortic regurgitation, with its characteristic diastolic murmur. Likewise, if the tear communicates with the pericardial space, patients can present with symptoms of acute pericardial tamponade (hypotension, pulsus paradoxus, jugular venous distention, and muffled heart sounds). Syncope or prolonged unconsciousness can be the initial presentation of patients with pericardial tamponade.

The initial aortic tear and subsequent extension of a false lumen along the aorta can occlude blood flow from the true lumen of the aorta into any of the arteries that originate from the aorta. Depending on which arteries become occluded, patients can present with a variety of corresponding syndromes. These include acute myocardial infarction from occlusion or extension of tear into the coronary arteries (typically the right coronary artery); death, syncope, or hemiplegia after occlusion of one or both carotid arteries; absent peripheral pulses in the major limb vessels secondary to occlusion of the brachiocephalic trunk, left subclavian artery, or distal aorta; anuria from disruption of renal blood flow; and paraplegia or quadriplegia from occlusion of vessels feeding the anterior spinal artery.

Examination for the Signs and Symptoms of Thoracic Aortic Dissection

The classic clinical history for thoracic aortic dissection consists of the sudden onset of severe tearing or ripping chest pain radiating to the interscapular region or low back, occurring in late-middle-aged men with a history of hypertension. Physicians therefore need to inquire of patients about the onset, quality, radiation, and intensity of patients' pain. Inquiry should also be made of history or symptoms suggestive of factors that increase the risk of aortic dissection, including hypertension, Marfan syndrome, bicuspid aortic valve, previous aortic valve replacement, and the other syndromes previously listed.

History-taking from patients with thoracic aortic dissection has tended to be poor; however, there is evidence that a more thorough medical history may increase diagnostic yield. A retrospective chart review of 83 patients with subsequently confirmed thoracic aortic dissection revealed that only 42% of conscious patients were asked all of 3 basic questions about their pain (quality, radiation, intensity at onset).¹⁴ One-quarter of patients were asked 1 or none of these key questions. If all 3 questions were asked, physicians correctly diagnosed thoracic aortic dissection in 30 of 33 patients (91%); if 1 or more of these questions was omitted, then the correct diagnosis was suspected during the initial evaluation in only 22 of 45 (49%) patients ($P < .001$). In

these patients, the diagnosis was made later, usually as an incidental finding during imaging procedures intended to diagnose alternative conditions. Unfortunately, the retrospective design of this study cannot preclude the possibility that physicians were simply more likely to ask about additional classic findings when they already had a strong clinical suspicion of thoracic aortic dissection derived from other data, including physical examination and chest radiograph.

The physical examination should begin with elicitation of vital signs, particularly the blood pressure and pulses on both sides of the body. While checking the blood pressure, the examiner should evaluate for acute pericardial tamponade by assessing for pulsus paradoxus, particularly in a patient with hypotension or jugular venous distention. Frequent allusion is made to the importance of comparing the blood pressure in both arms. Although it is essential to seek evidence of vascular occlusion in the arms, the complete examination should include comparison of all major arteries, including the carotid and femoral pulses, in addition to the radial pulses.

Most of the published series of patients with thoracic aortic dissection comment only on the loss or obvious diminishment of pulses rather than particular blood pressure differentials. Older retrospective autopsy series that do refer to blood pressure differentials arbitrarily designate a difference in systolic pressure between arms of 20 mm Hg or 30 mm Hg as significant.^{3,27} However, a convenience sample of 610 patients without thoracic aortic dissection presenting to an emergency department showed that 53% had interarm differences of greater than 10 mm Hg and 19% had differences greater than 20 mm Hg.²⁸ Nonetheless, a good-quality, prospective, observational study did find that a blood pressure differential of greater than 20 mm Hg was an independent predictor of dissection.²⁹ Hence, a blood pressure differential of at least 20 mm Hg ought to be present to be considered significant.

Cardiac auscultation should focus on detecting the diastolic murmur of aortic regurgitation.³⁰ A rapid neurologic examination directed toward the detection of gross motor and sensory defects such as hemiplegia and paraplegia should ensue.

Rarer clinical findings reported in the literature include pulsatile sternoclavicular joint, hoarseness, dysphagia, superior vena cava syndrome, Horner syndrome, bulbar palsies, acute arterial occlusion, deep vein thrombosis, and bilateral testicular tenderness.³¹⁻³⁷

A chest radiograph should be obtained and examined for abnormalities of the aortic silhouette. This is best accomplished with a standing anteroposterior projection. Unfortunately, the majority of chest radiograph findings associated with thoracic aortic dissection are subjective and not defined. Criteria for radiographic features associated with traumatic thoracic aortic dissection have been proposed but have not been adopted or validated in radiologic studies of nontraumatic dissections.³⁸ Radiographic abnormalities may include wide mediastinum, widening of the aortic knob, difference in diameter between the ascending and descending aorta, and blurring of the aortic margin secondary to local extravasation

of blood.³⁹ The chest radiograph might also reveal unilateral or bilateral pleural effusions. The calcium sign, consisting of the separation of intimal calcification from the outer border of the aortic knob by 1 cm or more, is highly suggestive of dissection but present in a minority of cases.^{37,40} Comparison with previous chest radiographs of the same patient can help the examiner detect suggestive new changes in the aortic contour.

METHODS

Literature Search and Selection

A structured MEDLINE search including 1966 through 2000 was conducted to identify English-language articles examining the accuracy of the clinical history, examination, and chest radiograph in the detection of acute thoracic aortic dissection. Key words used in the search included "physical examination," "medical history taking," "professional competence," "reproducibility of results," "observer variation," "diagnostic tests," "decision support techniques," "Bayes theorem," "sensitivity," "specificity," "thoracic aortic dissection," "aortic aneurysm," and "dissecting aneurysm." Articles focusing only on electrocardiograms (ECGs) were not specifically sought because such analyses document a variety of abnormalities seen with thoracic aneurysm but lack the appropriate clinical information for valid sensitivity and specificity estimates. When studies reported the results of ECGs as part of the overall clinical examination, however, these data were collated. Abstracts were reviewed and the full texts of articles that might meet the inclusion criteria were retrieved. The reference lists of reviewed articles were searched to identify additional sources.

All potential articles were reviewed for explicit inclusion and exclusion criteria. Articles were included if they were original studies describing the clinical findings in a series of 18 or more consecutive patients with confirmed dissection of the thoracic aorta (Table 50-1). Acceptable means of confirmation of diagnosis were surgical exploration, autopsy, aortogram, magnetic resonance imaging, computed tomography, or transesophageal echocardiography. The latter 4 imaging studies were included as acceptable gold-standard investigations according to high sensitivity and specificity.^{41,42} Articles were excluded if more than 15% of their cohorts included trauma patients, patients with chronic thoracic aortic dissection (defined as a dissection presumed to have occurred more than 14 days before presentation), or patients with abdominal aortic aneurysms or if the study selectively included patients with only proximal or distal dissections.

Retrieved studies were graded for quality using criteria similar to that used in previous articles in this series but modified to include only consecutive series. Level 1 studies were defined as prospective, blinded examinations of a large number (>100) of independently selected consecutive patients. Level 2 studies were of identical criteria but included fewer than 100 patients. Level 3 studies were large, prospective investigations but included nonindependently selected patients. Level 4 studies were retrospective reviews of nonindependently selected patients (see Table 1-7).

Table 50-1 Studies Assessing the Accuracy of Clinical Examination for Thoracic Aortic Dissection

Source, y	Clinical Setting, Study Dates	Design	No. of Patient Episodes	Age, y, Mean (Range)	Male, %	Type A, % ^a	Level of Quality ^b
Armstrong et al, ⁴³ 1998	University hospital, 1992-1994	Retrospective review of patients with clinically suspected TAD referred for TEE	75 (34 With TAD)	57 (20-80)	74	91	4
Chan, ⁴⁴ 1991	University hospital, 1987-1989	Prospective evaluation of utility of transesophageal echocardiography in patients with clinically suspected TAD	40 (18 With TAD)	60	60	... ^c	4
Enia et al, ⁴⁵ 1989	Hospital, 1981-1987	Prospective evaluation of transthoracic echocardiography in patients with clinically suspected TAD	46 (35 With TAD)	58 (34-82)	91	66	4
Erb and Tullis, ⁴⁶ 1960	University hospital, 1950-1960	Retrospective chart review	30	56 (36-85)	67	...	4
Hagan et al, ⁵ 2000	12 Tertiary centers in 6 countries, 1996-1998	Multinational prospective international registry; cases identified on admission or review of discharge/surgery/radiology records; 60% of cases referred	464	63	65	62	4
Hume and Porter, ⁴⁷ 1963	University hospital and medical examiner's office, 1950-1962	Retrospective chart review ^d	68	53 (10-79)	79	81	4
Itzchak et al, ⁴⁸ 1975	Hospital, 1960-1973	Retrospective chart review	24	57 (12-86)	75	46	4
Jagannath et al, ⁴⁰ 1986	University hospital, 1965-1977	Retrospective review of radiographs ^e	72 (36 With TAD)	62 (17-85)	Not stated	1/3	4
Levinson et al, ²⁷ 1950	University hospital, 1935-1947	Retrospective chart review of autopsy cases	58	59 (22-90)	72	...	4
Lindsay and Hurst, ⁴⁹ 1967	University hospital, 1949-1966	Retrospective chart review	62	57 (31-83)	65	65	4
Luker et al, ⁵⁰ 1994	Hospital, 1987-1993	Retrospective review of radiologists' initial chest radiograph readings in cases with subsequently confirmed TAD	75	61 (24-77)	49	47	4
Mészáros et al, ¹⁰ 2000	3 Hungarian towns, 1972-1998	Longitudinal, observational, population-based study ^f	86	66 (36-97)	61	86	4
Miller et al, ⁵¹ 1979	University hospital, 1963-1979	Retrospective review of surgically managed cases	73	57 (20-86)	70	73	4
Nielsen, ⁵² 1961	3 Danish hospitals, 1944-1958	Retrospective chart review ^g	40	66 (36-83)	45	...	4
Pate et al, ⁵³ 1976	Memphis, TN, hospitals, dates not given	Retrospective chart review	126	Not reported	79	...	4
Pinet et al, ⁵⁴ 1984	University hospital, 1970-1979	Retrospective chart review	191	58 (19-90)	69	64	4
Slater and DeSanctis, ³⁷ 1976	University hospital, 1963-1973	Retrospective chart review	124	59 (19-81)	73	43	4
Strong et al, ⁵⁵ 1974	University hospital and VA hospital, 1960-1973	Retrospective chart review	59	60 (26-86)	78	46	4
Sullivan et al, ¹¹ 2000	3 University hospital EDs, 1992-1996	Retrospective review of ED patients referred for thoracic imaging	44	65 (36-89)	...	61	4
Viljanen, ¹² 1986	University hospital, 1964-1985	Retrospective review of surgically managed cases	73	51	66	64	4
Von Kodolitsch et al, ²⁹ 2000	University hospital, 1988-1996	Prospective study of patients presenting to ED with history suggestive of TAD	250 (128 With TAD)	53	78	61	3

Abbreviations: ED, emergency department; TAD, thoracic aortic dissection; TEE, transesophageal echocardiography; VA, Veterans Affairs.

^aType A refers to aortic dissections involving the aorta proximal to the subclavian artery.^bSee Table 1-7.^cEllipses indicate information not available.^dTwo cases not confirmed by surgery or autopsy.^eDoes not include data on the frequency of specific radiographic findings but does report interobserver agreement.^fEleven percent of cases were chronic.^gForty cases in which TAD was considered cause of death; also reports additional 18 cases in which TAD was incidental finding on autopsy.

Study Characteristics

A total of 274 studies were identified by the search strategy, of which 21 studies met inclusion criteria (Table 50-1). No level 1 or level 2 studies were located. One study met level 3 criteria; the remaining 20 were level 4. One large series was self-described as prospective in conception and definition of clinical parameters.⁵ An unknown percentage of its patients, however, were identified by physician review of discharge records, echocardiography, and surgical databases. This study was consequently classified conservatively as level 4.⁵ Approximately half the investigations, including the 1 level 3 study, were specifically designed to elucidate the clinical presentation of acute aortic dissection. The remaining reports were either designed to test new imaging modalities or to study the outcomes of medical or surgical management of patients with thoracic aortic dissection. In each case, however, these studies included data on patients' clinical findings at diagnosis. The studies varied considerably in the number and detail of components of the clinical history or examination that were reported. Only the prospective level 3 study explicitly defined the criteria used to establish whether a given clinical finding was present or absent.²⁹

These studies assessed a total of 1848 patients aged 10 to 97 years. The major limitation of all the studies is that patients were selected for inclusion either retrospectively after confirmation of diagnosis by a reference standard study or prospectively according to the presenting clinical picture. Therefore, in all these studies the reference standard and clinical examination were not applied independently of one another. This biases the results of the studies to overestimate the sensitivity of clinical findings because more obvious cases are preferentially included in such series. In addition, physicians performing the reference standard procedure were not blinded to the results of the clinical examination and vice versa. This too could lead to overestimation of sensitivity.

Only 4 studies included control groups.^{29,43-45} Although these investigations can be used to generate data for specificity in addition to sensitivity, their estimations of specificity are heavily influenced by their inclusion biases. The specificities derived from these studies should be interpreted with caution because they reflect only the specificity for a given sign or symptom among patients similar to those included in the studies (ie, those with a full clinical syndrome suggestive of thoracic aortic dissection). These studies likely overestimate sensitivity and underestimate specificity by selecting patients for inclusion because of the presence of the particular sign being considered, thereby creating cohorts with artificially high prevalence of the finding.

Data Analysis

Summary measures for the sensitivity for components of the clinical examination for acute thoracic aortic dissection used published raw data from the reported trials that met criteria. Only 4 studies included specificity data that allowed construction of likelihood ratios (LRs). A random-effects model was used to generate conservative summary measures and confidence intervals (CIs) for the sensitivity and LRs.⁵⁶ For LRs, a

summary measure is reported only when there are more than 2 studies. The uncertainty in these measures is reflected in the broad CIs around the estimates. Interobserver agreement was calculated and interpreted using the κ statistic of Landis and Koch.⁵⁷ Fast Pro version 1.8 software was used for the meta-analysis (Academic Press, San Diego, California).

RESULTS

Accuracy of the Clinical History

Risk Factors

Sixteen studies examining 1553 patients report sensitivities for various components of the clinical history in Table 50-2. Most patients with dissection have a documented history of hypertension (sensitivity, 64%); however, the LR+ of this history is 1.6 (95% CI, 1.2-2.0). The pooled prevalence of the Marfan syndrome in this group of studies was 5% (95% CI, 4%-7%). Given that the Marfan syndrome afflicts only 0.02% to 0.03% of the general population,⁵⁸ the high prevalence of the Marfan syndrome in these series is suggestive of a markedly increased risk associated with this disorder, though the frequency of the Marfan syndrome detected in these series likely reflects the inclusion biases of these studies. The one controlled study that assessed for the Marfan syndrome generated an LR+ of 4.1.²⁹

Symptoms

The majority of patients presented with pain (pooled sensitivity, 90%) of severe intensity (sensitivity, 90%) that occurred suddenly (sensitivity, 84%). All other recorded clinical symptoms were present in a low to moderate proportion of patients (Table 50-2). Patients were most likely to have anterior chest pain (sensitivity, 57%); however, pain was frequently experienced elsewhere, including the posterior chest (32%), back (32%), and abdomen (23%). Likewise, migrating and ripping or tearing pain was present in only 31% and 39% of patients, respectively.

The presence of pain of sudden onset is not diagnostic (LR+, 1.6; 95% CI, 1.0-2.4). The absence of this history, however, substantively decreases the probability of an acute thoracic aortic dissection (LR-, 0.3; 95% CI, 0.2-0.5). Physicians should be cautious about relying too heavily on the absence of sudden pain to exclude aortic dissection because the inclusion biases of these studies likely overestimate the sensitivity.

Pain of a tearing or ripping sensation may also be diagnostically useful. Two studies found almost identical specificities of 94% and 95% for this historical feature.^{29,43} Although the reported specificities were almost identical, the LR+s generated by these 2 studies differed considerably (1.2 vs 11; Table 50-3) reflecting significant heterogeneity in the sensitivity for this history reported by the 2 investigations. The retrospective study found that only 7% of patients had noted tearing or ripping pain.⁴³ By contrast, the better-quality, larger, prospective study, in which physicians were asked to query predefined clinical symptoms of each patient, reported a sensitivity of 62%.²⁹ This figure is more consistent with the other large study with prospectively defined clinical symptoms in this series⁵ and with the pooled sensitivity for this symptom (Table 50-2). Therefore, it seems reasonable to suspect that the higher reported sensitivity

Table 50-2 Sensitivity of the Clinical History in the Diagnosis of Acute Thoracic Aortic Dissection

Source, y	No. of Patients	Sensitivity, %												
		History of Hypertension	Marfan Syndrome	Any Pain	Chest Pain	Anterior Chest Pain	Posterior Chest Pain	Back Pain	Abdominal Pain	Sudden-Onset Pain	Severe Pain	Ripping or Tearing Pain	Migrating Pain	Syncope
Armstrong et al, ⁴³ 1998	34	... ^a	...	94	74	56	27	88	93	7	...	6
Chan, ⁴⁴ 1991	18	56	...	78	78	39	...
Enia et al, ⁴⁵ 1989	35	80
Erb and Tullis, ⁴⁶ 1960	30	53	7	70	40	17
Hagan et al, ⁵ 2000	464	72	5	96	73	61	36	53	30	85	91	51	17	9
Hume and Porter, ⁴⁷ 1963	68	89	4	97	59	59	33	43	49
Levinson et al, ²⁷ 1950	58	59	...	78	47	...	9	36	40	14
Lindsay and Hurst, ⁴⁹ 1967	62	90	...	61	14	13	11
Mészáros et al, ¹⁰ 2000	72	67	...	92	...	64	...	10	10	14
Nielsen, ⁵² 1961	40	18	3	65	...	54	...	8	33	76	16
Pate et al, ⁵³ 1976	126	88	63	...	38	22	...	88	10
Pinet et al, ⁵⁴ 1984	191	53	7	96	63	...	30	89	...	6	...
Slater and DeSanctis, ³⁷ 1976	124	65	5	94	91	43	38	76	4	93	94	...	71	5
Strong et al, ⁵⁵ 1974	59	75	3	32	...	25	27
Sullivan et al, ¹¹ 2000	44	70	0	98	66	34	2
Von Kodolitsch et al, ²⁹ 2000	128	77	7	100 ^b	...	76	...	50 ^c	22	79	86	62	44	10
Summary sensitivity, % (95% CI)	NA	64 (54-72)	5 (4-7)	90 (85-94)	67 (56-77)	57 (48-66)	32 (24-40)	32 (19-47)	23 (16-31)	84 (80-89)	90 (88-92)	39 (14-69)	31 (12-55)	9 (8-12)

Abbreviations: CI, confidence interval; NA, not applicable.

^aEllipses indicate data not available.^bPresence of pain inclusion criterion for study.^cPosterior chest or lower back pain.

Table 50-3 Accuracy of Clinical Findings for Thoracic Aortic Dissection in Consecutive Patients Preselected for High Clinical Suspicion of Dissection Referred for Advanced Imaging

Symptom or Sign	Source, y	LR+ (95% CI)	LR- (95% CI)
History of hypertension	Chan, ⁴⁴ 1991 ^a	1.5 (0.8-3.0)	0.7 (0.4-1.3)
	Enia et al, ⁴⁵ 1989 ^b	1.1 (0.7-1.6)	0.7 (0.4-2.4)
	Von Kodolitsch et al, ²⁹ 2000 ^c	1.8 (1.4-2.3)	0.4 (0.3-0.6)
	Summary	1.6 (1.2-2.0)	0.5 (0.3-0.7)
Sudden chest pain	Chan, ⁴⁴ 1991 ^a	1.0 (0.7-1.4)	0.98 (0.3-3.1)
	Armstrong et al, ⁴³ 1998 ^d	1.5 (1.1-1.9)	0.3 (0.1-0.8)
	Von Kodolitsch et al, ²⁹ 2000 ^c	2.6 (2.0-3.5)	0.3 (0.2-0.4)
	Summary	1.6 (1.0-2.4)	0.3 (0.2-0.5)
"Tearing" or "ripping" pain	Armstrong et al, ⁴³ 1998 ^d	1.2 (0.2-8.1)	0.99 (0.9-1.1)
	Von Kodolitsch et al, ²⁹ 2000 ^c	11 (5.2-22)	0.4 (0.3-0.5)
Migrating pain	Chan, ⁴⁴ 1991 ^a	1.1 (0.5-2.4)	0.97 (0.6-1.6)
	Von Kodolitsch et al, ²⁹ 2000 ^c	7.6 (3.6-16)	0.6 (0.5-0.7)
Pulse deficit	Armstrong et al, ⁴³ 1998 ^d	2.4 (0.5-12)	0.93 (0.8-1.1)
	Enia et al, ⁴⁵ 1989 ^b	2.7 (0.7-9.8)	0.63 (0.4-1.0)
	Von Kodolitsch et al, ²⁹ 2000 ^c	47 (6.6-333)	0.62 (0.5-0.7)
	Summary	5.7 (1.4-23)	0.7 (0.6-0.9)
Focal neurologic deficit	Armstrong et al, ⁴³ 1998 ^d	6.6 (1.6-28)	0.71 (0.6-0.9)
	Von Kodolitsch et al, ²⁹ 2000 ^c	33 (2.0-549)	0.87 (0.8-0.9)
Diastolic murmur	Chan, ⁴⁴ 1991 ^a	4.9 (0.6-40)	0.8 (0.6-1.1)
	Armstrong et al, ⁴³ 1998 ^d	1.2 (0.4-3.8)	0.97 (0.8-1.2)
	Enia et al, ⁴⁵ 1989 ^b	0.9 (0.5-1.7)	1.1 (0.6-1.7)
	Von Kodolitsch et al, ²⁹ 2000 ^c	1.7 (1.1-2.5)	0.79 (0.6-0.9)
	Summary	1.4 (1.0-2.0)	0.9 (0.8-1.0)
Enlarged aorta or wide mediastinum	Chan, ⁴⁴ 1991 ^a	1.6 (1.1-2.3)	0.13 (0.02-1.0)
	Armstrong et al, ⁴³ 1998 ^d	1.6 (1.1-2.2)	0.42 (0.2-0.9)
	Von Kodolitsch et al, ²⁹ 2000 ^c	3.4 (2.4-4.8)	0.31 (0.2-0.4)
	Summary	2.0 (1.4-3.1)	0.3 (0.2-0.4)
Left ventricular hypertrophy on admission electrocardiogram	Chan, ⁴⁴ 1991 ^a	0.2 (0.03-1.9)	1.2 (0.9-1.6)
	Von Kodolitsch et al, ²⁹ 2000 ^c	3.2 (1.5-6.8)	0.84 (0.7-0.9)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

^aA total of 18 (n = 40) patients with thoracic aortic dissection.

^bA total of 35 (n = 46) patients with thoracic aortic dissection.

^cA total of 128 (n = 250) patients with thoracic aortic dissection.

^dA total of 34 (n = 75) patients with thoracic aortic dissection.

and LR are the more accurate data. Migratory pain has performance characteristics that are similar to tearing or ripping pain. The LR+ for the presence of this quality was 7.6 (95% CI, 3.6-16) in one study²⁹ but only 1.1 (95% CI, 0.5-2.4) in the other.⁴⁴ Additional studies of independently selected patients that prospectively ask about the sensation of tearing or ripping and migration of pain are needed to confirm the high LR for these findings. Description of pain as sharp was slightly more prevalent than tearing or ripping; however, this descriptor was elicited in only 2 studies and had an LR+ near unity.^{5,43}

Accuracy of the Physical Examination

Physical examination findings classically associated with thoracic aortic dissection are typically present in less than half of

all cases (Table 50-4). However, when present, signs of thoracic aortic dissection can be helpful. Among the most useful is a pulse differential between carotid, radial, or femoral arteries. Although the pooled sensitivity for this sign is only 31%, a deficit in 1 of these pulses compared with the contralateral side is strongly suggestive of dissection (LR+, 5.7; 95% CI, 1.4-23).^{29,43,45} Focal neurologic deficits, though present in only 17% of cases, may also be helpful. Specificity for this sign is high in the 2 studies in which it has been measured (LR+, 6.6-33; Table 50-3).^{29,43} The absence of a pulse deficit or focal neurologic deficit does not appreciably alter the likelihood of thoracic aortic dissection.

The presence or absence of a diastolic murmur is not helpful. Only one-third of patients with thoracic aortic dissection have a diastolic murmur (sensitivity, 28%). The LR+ and LR-

Table 50-4 Sensitivity of the Physical Examination in the Diagnosis of Acute Thoracic Aortic Dissection

Source, y	No. of Patients	Sensitivity, %							
		Elevated BP	Diastolic Murmur	Pulse Deficit	Pericardial Rub	Congestive Heart Failure	Focal Neurologic Deficit	Shock	New MI on ECG
Armstrong et al, ⁴³ 1998	34	... ^a	15	12	32	26	11
Chan, ⁴⁴ 1991	18	...	22
Enia et al, ⁴⁵ 1989	35	...	49	49
Erb and Tullis, ⁴⁶ 1960	30	...	27	72	0	...	13	...	25
Hagan et al, ⁵ 2000	464	49	32	15	...	7	5	16	3
Hume and Porter, ⁴⁷ 1963	68	68	4	34	10	...
Itzchak et al, ⁴⁸ 1975	24	21	21
Levinson et al, ²⁷ 1950	58	66	28	19	5	...	16	22	32
Lindsay and Hurst, ⁴⁹ 1967	62	29	35	45	23	13	...
Mészáros et al, ¹⁰ 2000	66	44	11	20	2	...	41	36	9
Miller et al, ⁵¹ 1979	73	58	64	29	12
Nielsen, ⁵² 1961	40	30	10
Pate et al, ⁵³ 1976	126	37	21	33	13	21	...
Pinet et al, ⁵⁴ 1984	191	...	35	55	12	38	...
Slater and DeSanctis, ³⁷ 1976	124	36	32	31	19	10	3
Strong et al, ⁵⁵ 1974	59	66	20	34	5	...
Sullivan et al, ¹¹ 2000	44	12	14	...	2
Viljanen, ¹² 1986	73	...	29	37	22	30	...
Von Kodolitsch et al, ²⁹ 2000	128	41	40	38	13	12	2
Summary sensitivity (95% CI)	NA	49 (41-57)	28 (21-36)	31 (24-39)	6 (3-13)	15 (4-33)	17 (12-23)	19 (15-26)	7 (4-14)

Abbreviations: BP, blood pressure; CI, confidence interval; ECG, electrocardiogram; MI, myocardial infarction; NA, not applicable.

^aEllipses indicate data not available.

(LR+, 1.4; 95% CI, 1.0-2.0; LR-, 0.9; 95% CI, 0.8-1.0) are close to 1, suggesting that the presence or absence of a diastolic murmur should not be considered helpful.^{29,43-45} Unfortunately, these studies do not comment on whether the diastolic murmurs identified were known to be new or old. It is possible that if a diastolic murmur was known to be new that it had greater diagnostic utility.

Patients' blood pressure on presentation is not helpful. Although approximately half of patients present with elevated blood pressure (pooled sensitivity, 49%; 95% CI, 41%-57%), an equal proportion are either hypotensive or normotensive. Only 1 study permitted calculation of an LR for hypertension; however, this study confirmed its low diagnostic yield (LR+, 1.3 for systolic blood pressure >150 mm Hg).²⁹ Pericardial rub is rarely present (pooled sensitivity, 6%; 95% CI, 3%-13%). Assessment for pulsus paradoxus and jugular venous distension is not enumerated in any of the studies.

Electrocardiographic findings consistent with acute myocardial infarction do not rule out aortic dissection. New Q waves or ST-segment elevation were observed in 7% of admission ECGs (Table 50-4). Similarly, normal ECG results were documented in 8% to 31% (mean, 22%) of patients.^{5,10,11,37,46,52} The remaining ECGs had a variety of other abnormalities, including left ventricular hypertrophy, atrial fibrillation, and nonspecific ST-segment changes. As part of the clinical evaluation, ECGs have not been studied well but seem to have little utility for detecting or ruling out thoracic aortic dissection.

Accuracy of the Chest Radiograph

Pooling of 13 studies permitted analysis of 1337 radiographs. Only 3 studies commented on the proportion of portable vs conventional radiographs. The proportions of portable radiographs reported in these investigations were 24%, 61%, and 80%.^{29,43,50} Radiographic findings classically associated with thoracic aortic dissection are not reliably present (Table 50-5). However, most patients with thoracic aortic dissection do tend to have abnormal findings on chest radiographs (sensitivity, 90%) so that a completely normal radiograph result helps to decrease the likelihood of the diagnosis. In particular, absence of wide mediastinum and abnormal aortic contour decreases the probability of disease (LR-, 0.3; 95% CI, 0.2-0.4; Table 50-5).

Interobserver and intraobserver agreement for physician assessment of radiographs has been reported in 2 studies, both using radiologists as participants. Agreement was generally found to be fair ($\kappa = 0.25$ for intraobserver agreement on suspicion for aortic dissection⁵⁰; $\kappa = 0.23-0.33$ for interobserver agreement on presence of wide mediastinum, irregularities of the aortic contour, and pleural effusion⁴⁰). These low rates of interobserver agreement underscore the lack of validated standards for defining the radiographic features of aortic dissection.

Accuracy of Combinations of Findings

Most clinical findings associated with thoracic aortic dissection are insensitive when considered in isolation. Com-

Table 50-5 Sensitivity of the Chest Radiograph in the Diagnosis of Acute Thoracic Aortic Dissection

Source, y	No. of Patients	Sensitivity, % ^a				
		Abnormal Aortic Contour	Pleural Effusion	Displaced Intimal Calcification	Wide Mediastinum	Abnormal Chest Radiograph Findings
Armstrong et al, ⁴³ 1998	34	86	100
Chan, ⁴⁴ 1991	18	94	...
Earnest et al, ³⁹ 1979	74	66	27	7	11	93
Hagan et al, ⁵ 2000	427	50	19	14	62	88
Itzchak et al, ⁴⁸ 1975	24	88	17	4	83	...
Luker et al, ⁴⁸ 1994	75	76	...	8	...	85
Pate et al, ⁵³ 1976	87	...	10	...	70	90
Pinet et al, ⁵⁴ 1984	191	56	...
Slater and DeSanctis, ³⁷ 1976	116	96	9	9	...	96
Strong et al, ⁵⁵ 1974	59	54	...	2	34	95
Sullivan et al, ¹¹ 2000	31	42	84
Viljanen, ¹² 1986	73	75	...
Von Kodolitsch et al, ²⁹ 2000	128	76 ^b	13
Summary sensitivity (95% CI)	NA	71 (56-84)	16 (12-21)	9 (6-13)	64 (44-80)	90 (87-92)

Abbreviations: CI, confidence interval; NA, not applicable.

^aEllipses indicate data not available.

^bMedastinal or aortic widening.

binations of findings, though not often found, markedly increase the accuracy of clinical assessment for thoracic aortic dissection. The single level 3 study described increasing accuracy of progressive combinations of findings (Table 50-6).²⁹ For example, aortic pain alone (pain of sudden onset, tearing, or ripping in character or both) has an LR+ of 2.6; the presence of both aortic pain and pulse or blood pressure differentials increases the LR+ to 10 (95% CI, 1.4-80). Further addition of mediastinal or aortic widening on chest radiograph clinches the diagnosis with an LR+ of 66 (95% CI, 4.1-1062). Unfortunately, this diagnostically valuable triad was present in only 27% of patients. Conversely, patients without any findings from the triad (aortic pain, pulse of blood pressure differential, and mediastinal widening) are unlikely to have a thoracic aortic dissection, given an LR- of 0.07 (95% CI, 0.03-0.17). However, 4% of patients in this category, without any of the above signs, were nonetheless ultimately diagnosed with aortic dissection. Given the high morbidity of a missed diagnosis, even such a pronounced LR- is insufficient to defer diagnostic imaging if thoracic aortic dissection is still clinically suspected.

The improved accuracy of combinations of clinical findings may further be inferred from a holistic view of the 4 studies that selected patients for inclusion on the basis of an overall clinical picture suggestive of thoracic aortic dissection. Despite the relative rarity of thoracic aortic dissection compared with other acute causes of pain, approximately half the patients selected for these studies turned out to have thoracic aortic dissection (pooled sensitivity, 52%). By comparison, only 0.003% of patients presenting to an emergency department with acute back, chest, or abdominal pain are eventually diagnosed with dissection.²⁹ This implies that a full clinical history, examination, and radiograph substan-

tially select for patients with acute dissection. Furthermore, among patients referred for aortic imaging who turn out not to have an acute dissection, approximately half to three-quarters are diagnosed with alternative serious diseases that can potentially be identified by imaging intended to confirm the diagnosis of thoracic aortic dissection (Table 50-7).^{29,33,43-45,59} The clinical syndrome suspicious for thoracic aortic dissection, although far from pathognomonic for acute dissection, does detect patients with serious disease that merit advanced diagnostic imaging.

THE BOTTOM LINE

Despite the large number of case series describing patients with thoracic aortic dissection, the clinical examination for thoracic aortic dissection has yet to be prospectively scrutinized in an independent, blinded study. The extant data permit estimation of the sensitivity of clinical history, phys-

Table 50-6 Positive Likelihood Ratio of Aortic Dissection in Patients With Combinations of Findings^a

No. of Findings	LR+ (95% CI)
3	66 (4.1-1062)
2	5.3 (3.0-9.4)
1	0.5 (0.3-0.8)
0	0.1 (0.0-0.2)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio.

^aData from Von Kodolitsch et al.²⁹ Findings include aortic pain (severe, sudden-onset tearing pain), blood pressure or pulse differential between arms, or wide mediastinum on chest radiograph.

Table 50-7 Final Diagnoses in Patients With Clinical Syndromes Suggestive of Thoracic Aortic Dissection but Without Thoracic Aortic Dissection on Further Study

Diagnosis	No. (%) of Patients ^a			
	Von Kodolitsch et al, ²⁹ 2000 (n = 122)	Enia et al, ⁴⁵ 1989 (n = 11)	Armstrong et al, ⁴³ 1998 (n = 41) ^b	Eagle et al, ³³ 1986 (n = 51) ^c
Acute coronary syndrome	18 (15)	2 (18)	8 (20)	12 (24)
Chest wall syndrome	18 (15)
Mediastinal cyst or tumor	4 (8)
Neuroradicular syndrome	1 (0.8)
Pulmonary disease	1 (0.8)
Hypertensive crisis	11 (9)
Gastrointestinal disease (esophagitis, PUD, gastritis, pancreatitis)	12 (9.8)	2 (4)
Pneumothorax	2 (1.6)
Pulmonary embolism	6 (4.9)	1 (9)	...	1 (2)
Pleuritis	5 (4.0)	1 (2)
Pericarditis	7 (5.7)	4 (36)	3 (7)	3 (6)
Nondissecting aneurysm	...	1 (9)	13 (32)	4 (8)
Aortic plaque rupture or intramural hemorrhage	9 (22)	...
Valvular pathology	4 (10)	5 (10)
Arteriosclerotic emboli	1 (2)
No definitive diagnosis	4 (3.3)	3 (27)	14 (34)	14 (28)

Abbreviation: PUD, peptic ulcer disease.

^aEllipses indicate data not available.

^bSome patients without thoracic aortic dissection were given multiple diagnoses.

^cIncluded 55 patients with suspected thoracic aortic dissections but negative aortogram results; 4 patients were false-negative cases and later demonstrated to have thoracic aortic dissection.

ical examination, and chest radiography but likely overestimate the accuracy of the clinical examination by selectively including more obvious cases. A small number of studies have included control populations and may therefore estimate the specificity of components of the clinical examination; however, the accuracy of these data is again limited by the lack of independence between the selection of patients for study and clinical findings.

Given the high, rapid mortality associated with undiagnosed thoracic aortic dissection, prospective, independent studies of the clinical examination are needed to aid physicians in determining which aspects of the clinical examination ought to be relied on to refer patients rationally for further diagnostic studies. Until then, the current literature permits the following limited conclusions about the clinical examination:

- Most patients with thoracic aortic dissection have severe pain of abrupt onset. The absence of pain of sudden onset substantially decreases the probability of dissection (LR-, 0.3; 95% CI, 0.2-0.5); however, the study design of the reports included in this article precludes accurate assessment of the sensitivity and specificity of these features. The presence of tearing or ripping pain (LR+, 1.2-11) or pain that migrates (LR+, 1.1-7.6) may prove useful, but additional data are required to know whether they are reliable features of the clinical history.

- Physical findings associated with thoracic aortic dissection tend to be present in a third or fewer cases; however, pulse deficits (LR+, 5.7; 95% CI, 1.4-23) or focal neurologic deficits (LR+, 6.6-33) greatly increase the likelihood of thoracic aortic dissection in the appropriate clinical setting. The presence or absence of a diastolic murmur is not useful (LR+, 1.4; LR-, 0.9).
- A normal aorta and mediastinum on chest radiograph helps exclude the diagnosis (LR-, 0.3; 95% CI, 0.2-0.4), but no particular radiographic abnormality is dependably present.
- The presence of the above findings in combination increases the LR+ for dissection, but even the absence of multiple findings does not definitively exclude the diagnosis. Clinical history, examination, and radiography can help rule in aortic dissection but are not sufficiently accurate to rule out the disease.

CLINICAL SCENARIOS—RESOLUTIONS

CASE 1 The patient's clinical history of sudden onset of severe chest pain is worrisome. His history of hypertension slightly increases his risk of a thoracic aortic dissection. The absence of a diastolic murmur, blood pressure differential, neurologic deficit, and widened mediastinum does not reliably exclude the diagnosis of thoracic aortic dissection. Given the high mortality of untreated or mistreated thoracic aortic dissection, this patient merits further advanced imaging.

CASE 2 The presence of a neurologic deficit in a patient with a clinical history consistent with thoracic aortic dissection is a specific finding. This patient has a high likelihood of having an acute thoracic aortic dissection and ought to undergo urgent diagnostic imaging to locate and delineate the suspected lesion.

Author Affiliation at the Time of the Original Publication

Department of Medicine, Brigham and Women's Hospital, Boston, Massachusetts.

Acknowledgments

I thank Donald Glower, MD, Manesh Patel, MD, and David Isbell, MD, for thoughtful comments on an earlier version of this work.

REFERENCES

- Morgagni JB. De sedibus et causis morborum, 1761. In: Rosman HS, Patel S, Borzak S, Paone G, Retter K. Quality of history taking in patients with aortic dissection. *Chest*. 1998;114(3):793-795.
- Bean RB, Bean WB. *Sir William Osler: Aphorisms From His Bedside Teachings and Writings*. Springfield, IL: Charles C Thomas; 1961:138.
- Hirst AE Jr, Johns VJ Jr, Kime SW Jr. Dissecting aneurysm of the aorta: a review of 505 cases. *Medicine (Baltimore)*. 1958;37(3):217-279.
- Nienaber CA, von Kodolitsch Y. Meta-analysis of changing mortality pattern in thoracic aortic dissection. *Herz*. 1992;17(6):398-416.
- Hagan PG, Nienaber CA, Isselbacher EM, et al. The International Registry of Acute Aortic Dissection (IRAD): new insights into an old disease. *JAMA*. 2000;283(7):897-903.
- Masuda Y, Yamada Z, Morooka N, et al. Prognosis of patients with medically treated aortic dissections. *Circulation*. 1991;84(suppl 5):III7-III13.
- Burt CW. Summary statistics for acute cardiac ischemia and chest pain visits to United States EDs, 1995-1996. *Am J Emerg Med*. 1999;17(6):552-559.
- Barbant SD, Eisenberg MJ, Schiller NB. The diagnostic value of imaging techniques for aortic dissection. *Am Heart J*. 1992;124(2):541-543.
- Marian AJ, Harris SL, Pickett JD, et al. Inadvertent administration of rtPA to a patient with type 1 aortic dissection and subsequent cardiac tamponade. *Am J Emerg Med*. 1993;11(6):613-615.
- Mészáros I, Morocz J, Szlávi J, et al. Epidemiology and clinicopathology of aortic dissection. *Chest*. 2000;117(5):1271-1278.
- Sullivan PR, Wolfson AB, Leckey RD, Burke JL. Diagnosis of acute thoracic aortic dissection in the emergency department. *Am J Emerg Med*. 2000;18(1):46-50.
- Viljanen T. Diagnostic difficulties in aortic dissection: retrospective study of 89 surgically treated patients. *Ann Chir Gynaecol*. 1986;75(6):328-332.
- Spittell PC, Spittell JA Jr, Joyce JW, et al. Clinical features and differential diagnosis of aortic dissection: experience with 236 cases (1980 through 1990). *Mayo Clin Proc*. 1993;68(7):642-651.
- Rosman HS, Patel S, Borzak S, et al. Quality of history taking in patients with aortic dissection. *Chest*. 1998;114(3):793-795.
- Larson EW, Edwards WD. Risk factors for aortic dissection: a necropsy study of 161 cases. *Am J Cardiol*. 1984;53(6):849-855.
- Murdoch JL, Walker BA, Halperin BL, et al. Life expectancy and causes of death in the Marfan syndrome. *N Engl J Med*. 1972;286(15):804-808.
- Mattar SG, Kumar AG, Lumsden AB. Vascular complications in Ehlers-Danlos syndrome. *Am Surg*. 1994;60(11):827-831.
- Sybert VP. Cardiovascular malformations and complications in Turner syndrome. *Pediatrics*. 1998;101(1):E11.
- Soderbergh J, Malmvall BE, Andersson R, Bengtsson BA. Giant cell arteritis as a cause of death: report of 9 cases. *JAMA*. 1986;255(4):493-496.
- Pumphrey CW, Fay T, Weir I. Aortic dissection during pregnancy. *Br Heart J*. 1986;55(1):106-108.
- Rashid J, Eisenberg MJ, Topol EJ. Cocaine-induced aortic dissection. *Am Heart J*. 1996;132(6):1301-1304.
- Rogers FB, Osler TM, Shackford SR. Aortic dissection after trauma: case report and review of the literature. *J Trauma*. 1996;41(5):906-908.
- Ohmoto Y, Ikari Y, Hara K. Aortic dissection during directional coronary atherectomy. *Int J Cardiol*. 1996;55(3):289-291.
- Von Kodolitsch Y, Loose R, Ostermeyer J, et al. Proximal aortic dissection late after aortic valve surgery: 119 cases of a distinct clinical entity. *Thorac Cardiovasc Surg*. 2000;48(6):342-346.
- Greenwood WR, Robinson MD. Painless dissection of the thoracic aorta. *Am J Emerg Med*. 1986;4(4):330-333.
- Gerber O, Heyer EJ, Vieux U. Painless dissections of the aorta presenting as acute neurologic syndromes. *Stroke*. 1986;17(4):644-647.
- Levinson DC, Edmeades DT, Griffith GC. Dissecting aneurysm of the aorta: its clinical, electrocardiographic and laboratory features. *Circulation*. 1950;1(3):360-387.
- Singer AJ, Hollander JE. Blood pressure: assessment of interarm differences. *Arch Intern Med*. 1996;156(17):2005-2008.
- Von Kodolitsch Y, Schwartz AG, Nienaber CA. Clinical prediction of acute aortic dissection. *Arch Intern Med*. 2000;160(19):2977-2982.
- Choudhry NK, Etchells EE. The rational clinical examination: does this patient have aortic regurgitation? *JAMA*. 1999;281(23):2231-2238.
- Leonard JC, Hasleton PS. Dissecting aortic aneurysms: a clinicopathological study. *Q J Med*. 1979;48(189):55-63.
- Chan-Tack KM. Aortic dissection presenting as bilateral testicular pain. *N Engl J Med*. 2000;343(16):1199.
- Eagle KA, Quertermous T, Kritzer GA, et al. Spectrum of conditions initially suggesting acute aortic dissection but with negative aortograms. *Am J Cardiol*. 1986;57(4):322-326.
- Gleeson H, Hughes T, Northridge D, Prendergast BD. Bulbar palsy and chest pain. *Lancet*. 2000;356(9232):826.
- Robertson GS, Macpherson DS. Aortic aneurysm presenting as deep venous thrombosis. *Lancet*. 1988;1(8590):877-878.
- Pacifico L, Spodick D. ILEAD—ischemia of the lower extremities due to aortic dissection: the isolated presentation. *Clin Cardiol*. 1999;22(5):353-356.
- Slater EE, DeSanctis RW. The clinical recognition of dissecting aortic aneurysm. *Am J Med*. 1976;60(5):625-633.
- Fultz PJ, Melville D, Ekanej A, et al. Nontraumatic rupture of the thoracic aorta: chest radiographic features of an often unrecognized condition. *AJR Am J Roentgenol*. 1998;171(2):351-357.
- Earnest F IV, Muham JR, Sheedy PF II. Roentgenographic findings in thoracic aortic dissection. *Mayo Clin Proc*. 1979;54(1):43-50.
- Jagannath AS, Sos TA, Lockhart SH, et al. Aortic dissection: a statistical analysis of the usefulness of plain chest radiographic findings. *AJR Am J Roentgenol*. 1986;147(6):1123-1126.
- Nienaber CA, Von Kodolitsch Y, Nicolas V, et al. The diagnosis of thoracic aortic dissection by noninvasive imaging procedures. *N Engl J Med*. 1993;328(1):1-9.
- Sarasim FP, Louis-Simonet M, Gaspoz JM, Junod AF. Detecting acute thoracic aortic dissection in the emergency department: time constraints and choice of the optimal diagnostic test. *Ann Emerg Med*. 1996;28(3):278-288.
- Armstrong WF, Bach DS, Carey LM, et al. Clinical and echocardiographic findings in patients with suspected acute aortic dissection. *Am Heart J*. 1998;136(6):1051-1060.
- Chan KL. Usefulness of transesophageal echocardiography in the diagnosis of conditions mimicking aortic dissection. *Am Heart J*. 1991;122(2):495-504.
- Enia F, Ledda G, Lo Mauro R, et al. Utility of echocardiography in the diagnosis of aortic dissection involving the ascending aorta. *Chest*. 1989;95(1):124-129.
- Erb BD, Tullis IF. Dissecting aneurysm of the aorta: the clinical features of thirty autopsied cases. *Circulation*. 1960;22:315-325.
- Hume DM, Porter RR. Acute dissecting aortic aneurysms. *Surgery*. 1963;53:122-154.
- Itzhak Y, Rosenthal T, Adar R, et al. Dissecting aneurysm of thoracic aorta: reappraisal of radiologic diagnosis. *Am J Roentgenol Radium Ther Nucl Med*. 1975;125(3):559-570.
- Lindsay J, Hurst JW. Clinical features and prognosis in dissecting aneurysm of the aorta: a re-appraisal. *Circulation*. 1967;35(5):880-888.
- Luker GD, Glazer HS, Eagar G, et al. Aortic dissection: effect of prospective chest radiographic diagnosis on delay to definitive diagnosis. *Radiology*. 1994;193(3):813-819.

51. Miller DC, Stinson EB, Oyer PE, et al. Operative treatment of aortic dissections: experience with 125 patients over a sixteen-year period. *J Thorac Cardiovasc Surg.* 1979;78(3):365-382.
52. Nielsen NC. Dissecting aneurysm of the aorta. *Acta Med Scand.* 1961;170:117-127.
53. Pate JW, Richardson RL, Eastridge CE. Acute aortic dissections. *Am Surg.* 1976;42(6):395-404.
54. Pinet F, Froment JC, Guillot M, et al. Prognostic factors and indications for surgical treatment of acute aortic dissections: a report based on 191 observations. *Cardiovasc Intervent Radiol.* 1984;7(6):257-266.
55. Strong WW, Moggio RA, Stansel HC Jr. Acute aortic dissection: twelve-year medical and surgical experience. *J Thorac Cardiovasc Surg.* 1974;68(5):815-821.
56. Eddy DM, Hasselblad V, Shachter RD. *Meta-Analysis by the Confidence Profile Method: The Statistical Synthesis of Evidence.* San Diego, CA: Academic Press; 1992.
57. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics.* 1977;33(1):159-174.
58. Pyeritz RE. The Marfan syndrome. *Annu Rev Med.* 2000;51:481-510.
59. Oliver TB, Murchison JT, Reid JH. Spiral CT in acute non-cardiac chest pain. *Clin Radiol.* 1999;54(1):38-45.

Prepared by Michael Klompaas, MD
Reviewed by Frank Lederle, MD

CLINICAL SCENARIO

A 64-year-old man with history of hypertension is treated in the emergency department for a chief complaint of severe chest pain of recent onset, radiating to the abdomen. His heart examination is remarkable for the presence of an S4 but no murmurs. His electrocardiogram (ECG) has changes consistent with acute inferior myocardial infarction. You have drawn up a syringe full of tissue plasminogen activator that you are about to inject when a thought suddenly occurs to you: Could this patient have an acute thoracic aortic dissection?

UPDATED SUMMARY ON THORACIC AORTIC DISSECTION

Original Review

Klompaas M. Does this patient have an acute thoracic aortic dissection? JAMA. 2002;287(17):2262-2272.

UPDATED LITERATURE SEARCH

Additional aortic dissection studies were sought with the same parent search criteria used for The Rational Clinical Examination series, combined with the terms, “dissecting aneurysm,” “aortic rupture,” “aortic aneurysm, thoracic,” “aneurysm, dissecting,” “aortic diseases/diagnosis,” and the text word, “thoracic aortic dissection.” The search was conducted for studies published between 2000 and August 2004. In addition, articles citing the original Rational Clinical Examination articles were reviewed. The search strategy resulted in 468 articles. Titles and abstracts were reviewed with the same limitation criteria as in the original article to find large, consecutive series of patients suspected to have aortic dissection, whose diagnosis was confirmed with a reference standard investigation (computed tomography [CT] angiography, magnetic resonance imaging [MRI], transesophageal echocardiography [TEE], aortogram, surgical exploration, or autopsy). As before, studies limited to proximal or distal aortic dissection or abdominal aortic dissection were excluded. One new study was identified.

NEW FINDINGS

- Younger patients with thoracic dissection (<40 years old) are more likely to have abrupt chest pain and Marfan syndrome but less likely to have systolic hypertension compared with older patients.¹

Details of the Update

The only new investigation identified was an update² of the International Registry of Acute Aortic Dissection (IRAD) database report that figured prominently in the original review.³ This article, primarily directed at reporting the frequency with which different diagnostic modalities were used to make the diagnosis of aortic dissection, included a table of the clinical features of 628 registry patients (vs 464 reported in the original IRAD article). As a registry of patients with thoracic aortic dissection, the data can be used to estimate the sensitivity.

The registry also reports the results of imaging. Because many patients had multiple imaging studies (66%), we can use the results to estimate the sensitivity of the tests used as a reference standard. There was no statistical difference in the sensitivity for TEE, CT, MRI, or aortography (though relatively few patients had the latter 2 studies). Overall, these studies had a sensitivity of 0.91 (95% confidence interval [CI], 0.87-0.94).

IMPROVEMENTS IN THE DATA PRESENTED IN THE ORIGINAL PUBLICATION

The additional patients in the IRAD database confirmed the frequency of the clinical features of acute aortic dissection already described in The Rational Clinical Examination article. There were no substantial changes in the sensitivity from the original cohort. The reported frequency of clinical features varied by no more than a few percentage points between the first and second IRAD articles. These data allow us to refine our sensitivity estimates with narrower CIs. Additionally, the IRAD report shows the difference in sensitivity for patients younger than 40 years vs aged 40 years or older.

CHANGES IN THE REFERENCE STANDARD

None.

RESULTS OF LITERATURE REVIEW

The abrupt onset of chest pain is the most sensitive finding for a thoracic aortic dissection (Table 50-8).

Table 50-8 Sensitivity of Findings for Thoracic Aortic Dissection

Finding (No. of Studies)	Summary Sensitivity (95% CI)
History	
Hypertension (14)	0.65 (0.57-0.73)
Hypertension, age < 40 y (1)	0.34 (0.23-0.46)
Marfan syndrome (10)	0.04 (0.03-0.06)
Marfan syndrome, age < 40 y (1)	0.50 (0.38-0.62)
Symptoms	
Abrupt onset (8)	0.84 (0.81-0.86)
Abrupt onset, age < 40 y (1)	0.96 (0.89-0.99)
Chest pain (10)	0.71 (0.58-0.83)
Chest pain, age < 40 y (1)	1.0 (0.96-1.0)
Back pain (11)	0.30 (0.18-0.44)
Signs	
Pulse deficit (17)	0.32 (0.26-0.39)
Murmur of aortic insufficiency (17)	0.28 (0.22-0.35)
Chest Radiograph	
Widened mediastinum (10)	0.63 (0.44-0.80)

Abbreviation: CI, confidence interval.

EVIDENCE FROM GUIDELINES

The American College of Radiology has published appropriateness criteria to guide the choice of imaging modality for diagnosing acute thoracic aortic dissection.⁴ The guidelines advocate that all patients suspected of having an aortic dissection have a chest radiograph. Ironically, although the guidelines recommend a chest radiograph for all patients, much of the discussion of the radiograph observes significant limitations, including its lack of specificity, the subjectivity of interpretation, and imperfect sensitivity. Experts, however, recommend chest radiographs as a means of ruling out other

pathology (especially when a baseline comparison radiograph is available).

The guidelines also discuss the appropriateness of reference standard imaging modalities, including aortography, CT, MRI, and TEE. All 4 formats are highly sensitive and specific. Computed tomography with contrast injection is believed to be most appropriate, however, because it is safer and less invasive than angiography or TEE, as well as being faster, cheaper, and more readily available than all 3 other modalities. Transesophageal echocardiography requires an experienced physician available at short notice for providing additional data for operative planning.

CLINICAL SCENARIO—RESOLUTION

This clinical scenario underscores some of the particular problems in the diagnosis and immediate management of severe chest pain in the emergency department. In this scenario, the clinician is faced with 2 realistic diagnostic possibilities that are life threatening and yet have contradictory treatments (thrombolysis can be deadly in a patient with aortic dissection). Unfortunately, clinical evaluation to distinguish between aortic dissection and acute myocardial infarction is limited in the setting of a patient with clear ECG changes yet with symptoms consistent with aortic pain. No single aspect of clinical history, physical examination, ECG, or chest radiography is completely sensitive in the diagnosis of aortic dissection.

Nonetheless, some evidence-based options do exist to aid the rapid treatment of this patient. A chest radiograph ought to be obtained and compared, if possible, against a previous study of the same patient. A completely normal radiograph result would substantially decrease the probability of aortic dissection, whereas the detection of a widened mediastinum, change in the aortic contour, or displacement of intimal calcification can be highly suggestive of the diagnosis.

Ultimately, however, this patient needs a reference standard study to exclude aortic dissection. The most favorable options would be CT with contrast injection or TEE. The advantage of the former is its rapid diagnostic yield and ready availability. In this patient, where the possibility exists that he has a proximal aortic dissection causing an acute myocardial infarction, a TEE might be particularly advantageous.

THORACIC AORTIC DISSECTION—MAKE THE DIAGNOSIS

PRIOR PROBABILITY

Although no current studies address the prior probability of an acute aortic dissection, a recent population-based epidemiologic study allows us to infer a 2% thoracic aortic dissection prevalence among patients with chest pain.⁵

POPULATION FOR WHOM A THORACIC AORTIC DISSECTION MIGHT BE CONSIDERED

- Patients with acute chest pain, especially those with hypertension or a Marfanoid habitus

DETECTING THE LIKELIHOOD OF A THORACIC AORTIC DISSECTION

Although clinical history, physical examination, and chest radiography can be suggestive of aortic dissection, none of these elements alone is sufficiently sensitive or specific to independently rule in or rule out this high-stakes diagnosis. Nonetheless, certain findings on the clinical evaluation can be helpful in suggesting the diagnosis and the need to perform a reference standard investigation such as CT angiography or TEE (Table 50-9). Almost all patients have severe pain (pooled sensitivity, 90%) of sudden onset (pooled sensitivity, 84%). The presence of a pulse or blood pressure differential from one side of the body to the other in a patient with severe chest pain is not often found in patients with dissection (sensitivity, 31%), but the finding increases the likelihood of aortic dissection when discovered (positive likelihood ratio [LR], 5.7). Similarly, a new focal neurologic deficit occurs infrequently (sensitivity, 17%) but also increases the likelihood of an aortic dissection when it is present (positive LR, 6.6–33.0). A widened mediastinum on chest radiograph is neither reliably present (pooled sensitivity, 64%) nor diagnostic of aortic dissection (positive LR, 2.0). However, almost all chest radiographs from patients with dissection will have some abnormality (pooled sensitivity, 90%), so a completely normal chest radiograph result decreases the probability of dissection being present (LR, 0.3).

New data suggest that the presenting features in young patients with aortic dissection may differ from those of older patients, but the accuracy of those findings has not been studied. Despite the lack of data quantifying the accuracy, young patients (<40 years old) with acute chest discomfort and Marfanoid features may have a greatly increased LR for aortic dissection compared with all other patients with chest discomfort.

Table 50-9 Accuracy of Clinical Findings for Thoracic Aortic Dissection in Consecutive Patients Preselected for High Clinical Suspicion of Dissection Referred for Advanced Imaging^a

Symptom or Sign (Total No. of Patients)	LR+ (95% CI)	LR- (95% CI)
Focal neurologic deficit (325) ^{6,7}	6.6–33	0.7–0.9
Pulse deficit (371) ^{6,8}	5.7 (1.4–23)	0.7 (0.6–0.9)
Enlarged aorta or wide mediastinum (365) ^{6,7,9}	2.0 (1.4–3.1)	0.3 (0.2–0.4)
History of hypertension (336) ^{6,8,9}	1.6 (1.2–2.0)	0.5 (0.3–0.7)
Sudden chest pain (365) ^{6,7,9}	1.6 (1.0–2.4)	0.3 (0.2–0.5)
“Tearing” or “ripping” quality (325) ^{6,7}	1.2–11	0.4–1.0
Diastolic murmur (411) ^{6,9}	1.4 (1.0–2.0)	0.9 (0.8–1.0)
Migrating pain (290) ^{6,9}	1.1–7.6	0.6–1.0

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

^a95% CI reported when 3 or more studies combined; otherwise, the LRs reported represent the range.

Table 50-10 Positive Likelihood Ratio of Aortic Dissection in Patients With Combinations of Findings⁷

No. of Findings ^a	LR (95% CI)
3	66 (4.1–1062)
2	5.3 (3.0–9.4)
1	0.5 (0.3–0.8)
0	0.1 (0–0.2)

Abbreviations: CI, confidence interval; LR, likelihood ratio.

^aFindings include aortic pain (sudden-onset severe tearing pain), blood pressure or pulse differential between arms, and wide mediastinum on chest radiograph.

Consideration of combinations of findings can substantially alter the posttest probability of the diagnosis (Table 50-10). The absence of aortic pain (new, severe, “tearing” pain), blood pressure differential, or widened mediastinum substantially decreases the probability of aortic dissection (LR, 0.1), whereas the presence of all 3 of these findings is highly suggestive of the diagnosis (LR, 66).

REFERENCE STANDARDS

Computed tomography, aortography, MRI, or TEE.

REFERENCES FOR THE UPDATE

- Januzzi JL, Isselbacher EM, Fattori R, et al; for the International Registry of Aortic Dissection (IRAD). Characterizing the young patient with aortic dissection: results from the International Registry of Aortic Dissection (IRAD). *J Am Coll Cardiol.* 2004;43(4):665–669.^a
- Moore AG, Eagle KA, Bruckman D, et al. Choice of computed tomography, transesophageal echocardiography, magnetic resonance imaging, and aortography in acute aortic dissection: International Registry of Acute Aortic Dissection (IRAD). *Am J Cardiol.* 2002;89(10):1235–1238.^a
- Hagan PG, Nienaber CA, Isselbacher EM, et al. The International Registry of Acute Aortic Dissection (IRAD): new insights into an old disease. *JAMA.* 2000;283(7):897–903.
- American College of Radiology. Acute chest pain, suspected aortic dissection, ACR appropriateness criteria, 1999. http://www.acr.org/SecondaryMainMenuCategories/quality_safety/app_criteria/pdf/ExpertPanelonCardiovascularImaging/AcuteChestPainSuspectedAorticDissectionDocZ.aspx. Published 1995. Updated 2005. Accessed June 14, 2008.
- Clouse WD, Hallett JW Jr, Schaff HV, et al. Acute aortic dissection: population-based incidence compared with degenerative aortic aneurysm rupture. *Mayo Clin Proc.* 2004;79(2):176–180.

6. von Kodolitsch Y, Schwartz AG, Nienaber CA. Clinical prediction of acute aortic dissection. *Arch Intern Med.* 2000;160(19):2977-2982.
7. Armstrong WF, Bach DS, Carey LM, Froehlich J, Lowell M, Kazerooni EA. Clinical and echocardiographic findings in patients with suspected acute aortic dissection. *Am Heart J.* 1998;136(6):1051-1060.
8. Enia F, Ledda G, Lo Mauro R, Matassa C, Raspanti G, Stabile A. Utility of echocardiography in the diagnosis of aortic dissection involving the ascending aorta. *Chest.* 1989;95(1):124-129.
9. Chan KL. Usefulness of transesophageal echocardiography in the diagnosis of conditions mimicking aortic dissection. *Am Heart J.* 1991; 122(2):495-504.

^aFor the Evidence to Support the Update for this topic,
see <http://www.JAMAevidence.com>.

EVIDENCE TO SUPPORT THE UPDATE:

50

Thoracic Aortic Dissection

TITLE Characterizing the Young Patient With Aortic Dissection: Results From the International Registry of Aortic Dissection (IRAD).

AUTHORS Januzzi JL, Isselbacher EM, Fattori R, et al; International Registry of Aortic Dissection (IRAD).

CITATION *J Am Coll Cardiol.* 2004;43(4):665-669.

QUESTION How do the presentation and prognosis of aortic dissection differ for younger vs older patients?

DESIGN Retrospective case-control study using international data registry.

SETTING Five US hospitals and 8 non-US hospitals (Europe, Israel, Japan).

PATIENTS Nine hundred fifty-one patients enrolled from January 1996 to November 2001.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

Patients were retrospectively identified at each center. Physician reviewers used a common form to collect data from patients' medical records. The diagnostic standard was the unified conclusion from the combination of medical history, imaging studies, surgical visualization, or postmortem examination. Data collected included demographic information, details of the clinical presentation, and the results of imaging studies. Results were stratified by age into 2 cohorts, younger than 40 years or aged 40 years or older.

MAIN OUTCOME MEASURES

Risk factors, clinical presentation, results of imaging studies, and mortality for aortic dissection patients younger than 40 years compared with those aged 40 years or older.

MAIN RESULTS

Sixty-eight patients younger than 40 years were compared with 883 patients aged 40 years or older. Younger patients were less likely to have a history of hypertension (34% vs 72%) but were more likely to have Marfan syndrome (50% vs 2%), bicuspid aortic valve (9% vs 1%), or previous aortic valve replacement surgery (12% vs 5%).

Young patients were even more likely to complain of pain of abrupt onset (96% vs 82%), but other symptoms of aortic dissection were similar between the 2 groups. Younger patients were less likely to be hypertensive (25% vs 45%). Mortality rates did not differ between the 2 groups (22% vs 24%).

CONCLUSIONS

LEVEL OF EVIDENCE Level 4.

STRENGTHS Multicenter, multinational study with a large number of patients.

LIMITATIONS Retrospectively collected data without any attempt to capture data not recorded at original patient presentation. Selection and evaluation of patients were done without blinding to the ultimate clinical diagnosis or the results of previous studies. The cohort of younger patients was small relative to the sample size of older patients analyzed. This is a descriptive rather than an interventional study.

Younger patients with aortic dissection are substantially more likely to have Marfan syndrome or bicuspid aortic valve as their predisposing factors and less likely to have hypertension. Otherwise, the clinical presentation and prognosis of younger patients are similar to those of older patients.

Reviewed by Michael Klompas, MD

TITLE Choice of Computed Tomography, Transesophageal Echocardiography, Magnetic Resonance Imaging, and Aortography in Acute Aortic Dissection: International Registry of Acute Aortic Dissection (IRAD).

AUTHORS Moore AG, Eagle KA, Bruckman D, et al.

CITATION *Am J Cardiol.* 2002;89(10):1235-1238.

QUESTIONS Which imaging modalities are currently being used to diagnose acute thoracic aortic dissection and what is their sensitivity?

DESIGN International data registry.

SETTING Five US hospitals and 8 non-US hospitals (Europe, Israel, Japan).

PATIENTS Six hundred twenty-eight patients enrolled from January 1996 to December 1999.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

Physician reviewers used a common form to collect data retrospectively from medical records. The diagnostic standard was the unified conclusion from the combination of medical history, imaging studies, surgical visualization, or postmortem examination. Data collected included demographic information, details of the clinical presentation, imaging modalities used and the order in which they were performed, and the sensitivity of each imaging modality.

MAIN OUTCOME MEASURES

First and second choice of imaging modalities chosen for each patient. Sensitivity of each imaging modality.

MAIN RESULTS

The study report includes 618 patients who had imaging. The most commonly used imaging modality was computed tomography (CT), used for 75% of patients; however, an almost identical number (72%) received transesophageal echocardiography (TEE). Two-thirds of patients (66%) had 2 imaging studies done. CT was performed first in 63% (n = 379); TEE, in 32% (n = 193). Aortography (n = 24) and magnetic resonance imaging (MRI) (n = 9) were infrequently used as

Table 50-11 Sensitivity of Imaging Modalities to Diagnose Thoracic Aortic Dissection^a

Imaging Procedure	Sensitivity
Computed tomography	0.93 (0.90-0.95) ^b
Transesophageal echocardiography	0.88 (0.83-0.92)
Magnetic resonance imaging	1.0 (0.7-1.0)
Aortography	0.88 (0.69-0.96)

^aThe data represent the sensitivity for the initial study only.

^bResults are statistically identical for proximal vs distal dissection, though transesophageal echocardiography appears to have a higher sensitivity for proximal than distal dissections (0.90 vs 0.80; P = .06).

the initial study. The sensitivity of each imaging modality is reported in Table 50-11.

CONCLUSIONS

LEVEL OF EVIDENCE Level 4.

STRENGTHS Multicenter, multinational study describing actual clinical practice and the performance of CT and TEE under routine clinical conditions.

LIMITATIONS Retrospective case series rather than a prospective evaluation of the sensitivity and specificity of each imaging modality. The series consequently reflects the biases of each center in choosing various radiographic techniques in accordance with local clinical culture and variable equipment and operator availability. Likewise, the interpretation of the radiographic studies was not necessarily done by blinded, expert reviewers and hence might misestimate the true sensitivity of the various tests. The sample size was small for patients imaged with MRI and aortography. The data on clinical presentation are particularly limited because the data were collected retrospectively from medical records, without any attempt to ascertain missing information. In addition, the data were not abstracted by blinded clinicians, because they had access to the patients' final diagnoses.

Computed tomography is the most commonly used modality to diagnose aortic dissection. Two-thirds of patients, however, receive more than 1 imaging test. CT, TEE, MRI, and aortography all have high sensitivity; however, with the possible exception of MRI, they can all yield false-negative results.

Reviewed by Michael Klompas, MD

CHAPTER 51

Does This Woman Have an Acute Uncomplicated Urinary Tract Infection?

Stephen Bent, MD

Brahmajee K. Nallamothu, MD, MPH

David L. Simel, MD, MHS

Stephan D. Fihn, MD, MPH

Sanjay Saint, MD, MPH

CLINICAL SCENARIOS

CASE 1 A 24-year-old healthy woman calls her primary care physician, complaining of a burning pain when urinating and increased urinary frequency for several hours. She has had 2 previous urinary tract infections (UTIs), and this episode seems “just like the other ones.” She is sexually active with 1 partner and uses a condom with spermicide. She denies fever, back pain, nausea, vomiting, vaginal discharge, and hematuria.

CASE 2 A 20-year-old woman presents to your office, complaining of urinary frequency, burning on urination, and vaginal discharge. She has had occasional fevers and chills but denies nausea, vomiting, and back pain. She is sexually active with 1 partner, takes oral contraceptive pills, and intermittently they use condoms. Physical examination shows her to be in mild discomfort and febrile but without tenderness in her costovertebral areas. Pelvic examination demonstrates minimal white vaginal discharge, no vaginal lesions or rashes, and no cervicitis. Her dipstick urinalysis result is negative for leukocyte esterase, nitrite, and blood.

WHY IS THIS AN IMPORTANT QUESTION TO ANSWER WITH A CLINICAL EXAMINATION?

Acute uncomplicated UTIs are common in women, accounting for more than 7 million office visits annually in the United States¹ and affecting half of all women at least once during their lifetimes.² A recent study of sexually active young women found the incidence of cystitis to be 0.5% to 0.7% per year.³ In aggregate, the direct costs of these infections have been estimated to be \$1.6 billion annually in the United States.⁴

One might anticipate that the management of acute uncomplicated UTI would be relatively uniform because the causative agents and in vitro susceptibilities are known, and therapeutic responses to antimicrobials have been studied carefully.^{2,5-7} Unfortunately, the evaluation and treatment of acute uncomplicated UTI in women vary substantially among physicians,⁸ likely reflecting the limitations of routine diagnostic assessments. When done correctly, however, the history taking and physical examination can be used in the initial evaluation of patients suspected of having an acute uncomplicated UTI and can guide the selection of additional diagnostic and therapeutic strategies.^{2,7}

Definitions

Several types of UTI are described by their location: urethritis, cystitis, pyelonephritis, and perinephric abscess. The

usual reference standard for diagnosing UTI is the presence of “significant” bacteria in a clean-catch or catheterized urine specimen, most commonly defined as the isolation of at least 10^5 colony-forming units (CFU) per milliliter of a single uropathogen.² In women who present with symptoms of cystitis or urethritis (lower UTI), it has been suggested that the best diagnostic criterion for clean-catch urine is the isolation of uropathogens in concentrations as low as at least 10^2 CFU/mL.⁹

Uncomplicated UTIs occur in individuals who have a normal urinary tract system. A UTI in an individual with a functional or anatomic abnormality of the urinary tract (including a history of polycystic renal disease, nephrolithiasis, neurogenic bladder, diabetes mellitus, immunosuppression, pregnancy, indwelling urinary catheter, or recent urinary tract instrumentation) is considered complicated and may have a higher risk of treatment failure.¹⁰ Differentiating between these types of UTIs is important because uncomplicated infections are usually cured with simple antimicrobial regimens.¹⁰

The prevalence of asymptomatic bacteriuria (significant bacteriuria without symptoms of UTI) in women of reproductive age is approximately 5%.^{11,12} This value represents the pretest probability of disease (the probability of UTI before any diagnostic tests are applied). Several historical features, symptoms, and signs associated with acute UTI may be useful for screening, allowing the clinician to estimate the probability of UTI in a patient after taking a medical history and performing a physical examination. Historical features such as a history of UTI, recent sexual activity, or contraceptive use identify individuals at greater risk of developing a UTI. Symptoms of an acute infection include burning or pain on urination (dysuria), frequent voiding of small volumes of urine (frequency), the urge to void immediately (urgency), and the presence of blood in the urine (hematuria). Discomfort in the lower abdominal area is also consistent with a UTI. In contrast, patients who report vaginal discharge or irritation are less likely to have a UTI and more likely to have vaginitis or cervicitis. The presence of fever and suprapubic or costovertebral angle tenderness may indicate infection of the upper urinary tract.

Differential Diagnoses

Vaginal infections (eg, *Gardnerella*, *Candida albicans*, *Trichomonas*), sexually transmitted diseases that may lead to pelvic inflammatory disease (eg, *Chlamydia trachomatis*, *Neisseria gonorrhoeae*), and other sexually transmitted diseases (eg, herpes simplex virus) that mimic symptoms of UTI are considered separate from UTIs. These infections are caused by different microbes; limited to female genital structures, with a unique set of complications if untreated; and require different forms of treatment.¹³ Differentiating between vaginal infections, sexually transmitted diseases, and UTIs can be difficult because symptoms and signs commonly overlap.¹³

METHODS

We searched the English-language medical literature to determine the accuracy and precision of the clinical examination in women suspected of having an acute UTI. We searched MEDLINE for articles from 1966 through September 2001, with a search strategy similar to that used by other authors in this series.¹⁴ Search terms included “urinary tract infection,” “diagnostic tests,” “physical examination,” and “sensitivity and specificity.” This computerized search was supplemented with a manual review of the bibliographies of all identified articles, additional “core” articles (identified a priori as articles used to develop a recent guideline for treating acute uncomplicated UTI in women), 3 commonly used clinical skills textbooks,¹⁵⁻¹⁷ and contact with experts in the field. One of the authors (B.K.N.) initially screened the titles and abstracts of the search results. Two of the authors (S.B. and B.K.N.) then independently reviewed and abstracted data from articles identified as relevant.

We included studies in our review if they contained original data on the accuracy or precision of the symptoms or signs in diagnosing acute uncomplicated UTI in healthy women. Articles were excluded if they evaluated infants, children or adolescents, pregnant women, nursing home patients, or patients with complicated UTI or contained insufficient or incomplete data to allow calculation of likelihood ratios (LRs) for signs or symptoms of acute UTI.

We also chose to include articles on the dipstick test in this analysis because it is commonly used in the clinical setting and provides an immediate result that can be incorporated with other elements of the initial clinical assessment. During our search, we discovered that a previous systematic review evaluated the diagnostic accuracy of the dipstick test.¹⁸ Because this was a high-quality review (meeting all 6 criteria of a previously published guideline for evaluating systematic reviews),¹⁹ we chose to use the information about the accuracy of the dipstick test synthesized in that article.

Quality Assessment of Included Articles

The methodological quality of the included articles was assessed independently by 2 authors (S.B. and B.K.N.), using criteria adapted from other authors in this series.^{14,20} Disagreements were resolved by a third author (S.S.). Level 1 studies included those with an independent blind comparison of signs or symptoms with a gold standard among a large number (≥ 50) of consecutive patients suspected of having a UTI. Level 2 studies were similar to those in level 1 but involved a smaller number of patients (< 50). The remaining levels are described in Table 1-7.

Data Analysis

We used published raw data from the studies that met our criteria to calculate summary measures for the LRs for components of the clinical examination for UTI. LRs are related to sensitivity and specificity (positive likelihood ratio [LR+] = sensitivity/[1 – specificity]; negative likeli-

hood ratio [LR-] = [1 – sensitivity]/specificity) but are more clinically useful because they can be used to generate posttest probabilities.²¹ A random-effects model was used to generate conservative summary measures and confidence intervals (CIs) for the LRs and estimates of disease prevalence.^{22,23} Uncertainty in these measures is reflected in the broad CIs around the estimates. When a summary LR included studies of lower quality, we conducted sensitivity analyses to examine the influence of excluding lower-quality studies on the summary LR and the effectiveness score, a measure of the discriminatory power of a diagnostic test.²⁴

RESULTS

Study Characteristics

We found 9 studies of the 464 identified by the search that satisfied all inclusion criteria (Table 51-1). Six studies²⁵⁻³⁰ reported the accuracy of 1 or more symptoms in the diagnosis of UTI, 2 studies^{31,32} reported the accuracy of symptoms and physical examination signs, and 1 study reported the accuracy of self-diagnosis.³³

The studies were published between 1965 and 2001 and generally involved patients with 1 or more symptoms of a UTI who presented to outpatient clinics. The summary prevalence of UTI in the 5 studies that included only symptomatic patients and used an appropriate gold standard was 48% (95% CI, 41%-55%),^{25-28,30} indicating a high probability of disease for women who met the studies' inclusion criteria. In all of the included studies, UTI was defined by the presence of at least 10000 or 100000 CFU/mL of a single uropathogen, except for the most recent study, which used a cutoff of at least 100 CFU/mL.³³

Five^{25-28,30} of the 8 studies describing the accuracy of symptoms were of high quality (level 1). Both studies^{31,32} describing the accuracy of the physical examination were of lower quality (levels 3 and 4), as was the study examining self-diagnosis (level 5).³³ Reasons for quality scores lower than level 1 are shown in Table 51-1. Two of the lower-quality studies^{29,31} included patients with vaginal discharge but without symptoms of UTI and therefore did not specifically address the diagnostic accuracy of signs and symptoms exclusively in women suspected of having a UTI.

Table 51-1 Studies Used to Determine the Accuracy of Clinical History and Physical Examination in Women Suspected of Having Urinary Tract Infection

Source, y	Methodologic Quality ^a	Inclusion Criteria	No. of Patients	Mean Age, y ^b	Incidence of UTI, %	Setting and Country
Symptoms						
Gallagher et al, ²⁵ 1965	Level 1	Women with symptoms of UTI	130	...	59	Urban clinics in New Zealand
Mond et al, ²⁶ 1965	Level 1	Women with symptoms of UTI	83	...	45	General practice in the United Kingdom
Lawson et al, ²⁷ 1973	Level 1	Women aged 15-55 y with symptoms of UTI	343	...	47	Two general practices in the United Kingdom
Dans and Klaus, ²⁸ 1976	Level 1	Women reporting dysuria	84	26	46	US adult walk-in clinic
Komaroff et al, ²⁹ 1978	Level 4 (including women without symptoms suggestive of UTI)	Women with symptoms suggestive of urinary or vaginal infection	821	24	12	US ambulatory care facility
Nazareth and King, ³⁰ 1993	Level 1	Women aged 16-45 y presenting with frequency or dysuria	54	29	28	Two general practices in suburban London
Self-diagnosis						
Gupta et al, ³³ 2001	Level 5 (no urine culture in women without symptoms)	Women >18 y with a history of recurrent UTI	172	23	NA	US university-based clinic
Symptoms and Physical Examination Findings						
Wong et al, ³¹ 1984	Level 4 (including patients without symptoms suggestive of UTI)	Women with symptoms of UTI or with both UTI and vaginal complaints and random selection of women with vaginitis or STD	53 Cases, 139 controls	...	NA	US STD clinic
Wigton et al, ³² 1985	Level 3 (retrospective chart review)	Retrospective review of patients who had urine culture in emergency department	216 In training set, 236 in validation set	...	NA	US emergency department

Abbreviations: NA, indicates not applicable; STD, sexually transmitted disease; UTI, urinary tract infection.

^aMethodologic quality criteria are described in the "Methods" section (see also Table 1-7). Reasons for methodologic quality scores lower than level 1 are shown in parentheses.

^bEllipses indicate not mentioned.

Precision

The precision of a symptom or sign refers to the degree to which different examiners report the same finding (eg, dysuria present or absent) when interviewing or examining the same patient. None of the identified studies described the precision of the medical history or physical examination in the diagnosis of UTI, possibly because the questions and examination procedures were considered to be unambiguous. For example, most of the historical items consist of asking yes or no questions such as, Are you having burning or pain with urination? Variations in interview style and the phrasing of questions may affect results, but there is no information from the identified studies to suggest particular wording of questions or specific ways to examine patients for the 2 relevant physical examination signs (costovertebral angle tenderness and vaginal discharge).

Accuracy

Symptoms

Eight studies²⁵⁻³² examined the accuracy of 9 symptoms for predicting the presence of UTI. These symptoms and the corresponding LR+ and LR- from each study are shown in Table 51-2. Three of the symptoms (flank pain, abdominal pain, fever) had both summary LR+ and summary LR- with CIs overlapping 1.0 and are therefore not useful as diagnostic tests.

Four symptoms significantly increased the probability of UTI: dysuria, frequency, hematuria, and back pain. Four symptoms significantly decreased the probability of UTI: absence of dysuria, absence of back pain, a history of vaginal discharge, and a history of vaginal irritation. The symptoms with the greatest diagnostic power were a history of vaginal discharge (LR, 0.34) and a history of vaginal irritation (LR, 0.24); both of these symptoms substantially reduced the probability of UTI.

Self-diagnosis

One study examined the accuracy of self-diagnosis and included 172 women in a university-based practice with recurrent UTI (more than 2 UTIs in the past year).³³ During the study period, 88 of the women reported 172 episodes of self-diagnosed UTI; 144 of these episodes (84%; 95% CI, 77%-90%) were found to have positive urine culture results. Additionally, 64 women reported mild symptoms that they did not self-diagnose as UTI and another 20 women never had symptoms. In this population of patients, the positive predictive value of self-diagnosis was high (84%). LRs for self-diagnosis can be calculated assuming that the women with mild symptoms or no symptoms correctly self-diagnosed with no infection (these women did not have a urine culture, but all symptoms resolved spontaneously). If this assumption is true, the LR for a positive self-diagnosis is 4.0, whereas the LR for a negative self-diagnosis is 0 (Table 51-2).

Combinations of Symptoms

One study²⁹ provided information to calculate the LRs for combinations of symptoms in the diagnosis of UTI (Table 51-3). In this study, the presence of dysuria and frequency

without vaginal discharge or irritation was associated with a high LR (25). Conversely, the LR for the combination of vaginal discharge or irritation without dysuria was low (0.3). Although the LRs from this study must be interpreted with caution because of the study's low quality score (level 4), the observed LRs were similar to those calculated by combining the individual summary LRs from the other studies (Table 51-3).

Physical Examination

Two studies^{31,32} reported the accuracy of 2 physical examination signs for the presence of UTI. Both studies were of relatively low quality, and therefore the summary data do not represent strong evidence of the true accuracy of these signs (Table 51-2). The presence of costovertebral angle tenderness increases the likelihood of infection, but the LR is only weakly predictive and similar in magnitude to the related symptom of back pain. The presence of vaginal discharge on examination decreases the likelihood of UTI (LR, 0.69), although it is less powerful than the LR for the symptom of vaginal discharge reported by the patient (0.34).

Dipstick Urinalysis

Because a high-quality systematic review examining the accuracy of the dipstick urinalysis for the prediction of UTI exists, we used the data synthesized in the report by Hurlbut and Littenberg.¹⁸ Those authors identified and summarized 51 studies and generated summary receiver operating characteristic (ROC) curves for combinations of the nitrite and leukocyte esterase dipstick tests. They found that the nitrite-positive or leukocyte-esterase-positive combination had the greatest area under the ROC curve. The point on the summary ROC curve with the best accuracy represents a sensitivity of 75% and a specificity of 82%. With these values, the LR+ for a urinalysis is 4.2 and the LR- is 0.3 (Table 51-2). A range of similar points on the ROC curve that was supported by the largest number of studies was also examined, and the resulting LRs were similar in magnitude. Although other combinations of the nitrite and leukocyte esterase test will increase either sensitivity or specificity (eg, requiring both to be positive will decrease sensitivity and increase specificity), the nitrite- or leukocyte-esterase-positive combination was the most accurate test.¹⁸

Sensitivity Analysis

Because the largest study to examine the accuracy of symptoms was also of lower quality,²⁹ we performed a sensitivity analysis to determine the effect of this study on the summary LRs. Inclusion of this study always made the symptoms (dysuria, frequency, vaginal irritation, and vaginal discharge) appear to be more powerful diagnostic tests. However, in no case did inclusion of this study improve a test with marginal discriminatory power into the highly effective range (effectiveness score ≥ 3.0).²⁴ The LR+ and LR- for dysuria and frequency excluded 1.0, whether or not the study was included, with one exception. The LR+ for increased urinary frequency was 1.8 (95% CI, 1.1-3.0) when all studies were included vs 1.4 (95% CI, 1.0-1.9) when the study was excluded. That study²⁹ has a larger effect on the diagnostic value of vaginal symptoms

Table 51-2 Clinical Signs and Symptoms in the Prediction of Urinary Tract Infection^a

Study	LR+ (95% CI)	LR- (95% CI)
Dysuria		
Gallagher et al ²⁵	1.3 (1.1-1.6)	0.28 (0.12-0.67)
Mond et al ²⁶	1.4 (1.1-1.8)	0.22 (0.07-0.70)
Lawson et al ²⁷	1.2 (1.0-1.5)	0.77 (0.60-0.99)
Nazareth and King ³⁰	1.1 (0.87-1.5)	0.58 (0.14-2.4)
Komaroff et al ²⁹	3.2 (2.7-3.7)	0.16 (0.09-0.27)
Wong et al ³¹	3.0 (2.0-4.6)	0.53 (0.39-0.73)
Wigton et al ³² (training set)	1.4 (1.1-1.8)	0.69 (0.52-0.92)
Wigton et al ³² (validation set)	1.1 (0.81-1.4)	0.94 (0.72-1.2)
Summary	1.5 (1.2-2.0)	0.48 (0.31-0.74)
Frequency		
Gallagher et al ²⁵	0.96 (0.87-1.1)	1.6 (0.44-6.0)
Mond et al ²⁶	0.99 (0.90-1.1)	1.2 (0.17-8.0)
Lawson et al ²⁷	1.1 (1.0-1.3)	0.65 (0.43-0.97)
Dans and Klaus ²⁸	1.4 (1.0-2.1)	0.63 (0.37-1.1)
Nazareth and King ³⁰	1.0 (0.80-1.3)	0.87 (0.20-3.8)
Komaroff et al ²⁹	10 (7.8-13)	0.07 (0.04-0.16)
Wong et al ³¹	5.2 (3.1-8.7)	0.45 (0.32-0.63)
Wigton et al ³² (training set)	1.8 (1.0-3.5)	0.87 (0.75-1.0)
Wigton et al ³² (validation set)	1.3 (0.80-2.0)	0.93 (0.80-1.1)
Summary	1.8 (1.1-3.0)	0.59 (0.35-1.0)
Hematuria		
Gallagher et al ²⁵	1.8 (0.80-3.9)	0.88 (0.75-1.0)
Mond et al ²⁶	2.9 (1.0-8.6)	0.81 (0.66-1.0)
Nazareth and King ³⁰	6.5 (1.4-30)	0.70 (0.49-1.0)
Wigton et al ³² (training set)	1.6 (0.82-3.3)	0.92 (0.82-1.0)
Wigton et al ³² (validation set)	1.4 (0.60-3.4)	0.96 (0.88-1.1)
Summary	2.0 (1.3-2.9)	0.92 (0.86-0.98)
Fever		
Gallagher et al ²⁵	2.4 (1.2-4.9)	0.75 (0.61-0.92)
Mond et al ²⁶	2.8 (0.77-9.9)	0.87 (0.73-1.0)
Lawson et al ²⁷	0.65 (0.32-1.3)	1.0 (0.97-1.1)
Nazareth and King ³⁰	0 (0-175)	0.92 (0.78-1.1)
Wigton et al ³² (training set)	1.5 (0.74-3.0)	0.94 (0.84-1.0)
Wigton et al ³² (validation set)	2.1 (1.0-4.6)	0.89 (0.80-0.99)
Summary	1.6 (1.0-2.6)	0.9 (0.9-1.0)
Flank Pain		
Gallagher et al ²⁵	1.1 (0.64-1.7)	0.98 (0.77-1.2)
Mond et al ²⁶	1.1 (0.54-2.2)	0.97 (0.74-1.3)
Lawson et al ²⁷	1.1 (0.87-1.4)	0.92 (0.77-1.1)
Summary	1.1 (0.90-1.4)	0.84 (0.82-1.1)
Lower Abdominal Pain		
Gallagher et al ²⁵	0.99 (0.76-1.3)	1.0 (0.63-1.6)
Mond et al ²⁶	1.2 (0.67-2.1)	0.91 (0.65-1.3)
Wong et al ³¹	1.5 (0.90-2.4)	0.87 (0.71-1.1)
Summary	1.1 (0.90-1.4)	0.89 (0.75-1.0)
Vaginal Discharge		
Dans and Klaus ²⁸	0.80 (0.53-1.2)	1.3 (0.82-2.0)
Komaroff et al ²⁹	0.11 (0.06-0.19)	12 (8.9-16)

(continued)

Table 51-2 Clinical Signs and Symptoms in the Prediction of Urinary Tract Infection^a (*Continued*)

Study	LR+ (95% CI)	LR- (95% CI)
Vaginal Discharge		
Wong et al ³¹	0.43 (0.27-0.69)	1.9 (1.4-2.5)
Summary	0.34 (0.14-0.86)	3.1 (1.0-9.3)
Vaginal Irritation		
Komaroff et al ²⁹	0.09 (0.05-0.18)	6.2 (5.0-7.6)
Wong et al ³¹	0.63 (0.37-1.1)	1.2 (1.0-1.5)
Summary	0.24 (0.06-0.93)	2.7 (0.88-8.5)
Back Pain		
Wigton et al ³² (training set)	1.7 (1.1-2.6)	0.80 (0.67-0.96)
Wigton et al ³² (validation set)	1.6 (1.1-2.5)	0.81 (0.68-0.97)
Nazareth and King ³⁰	0.78 (0.25-2.4)	1.1 (0.79-1.5)
Summary	1.6 (1.2-2.1)	0.83 (0.74-0.94)
Self-diagnosis		
Gupta et al ³³	4.0 (2.9-5.5)	0 (0-0.08)
Vaginal Discharge on Physical Examination		
Wong et al ³¹	0.81 (0.66-0.99)	1.9 (1.1-3.3)
Wigton et al ³² (training set)	0.32 (0.12-0.89)	1.1 (1.0-1.2)
Wigton et al ³² (validation set)	0.44 (0.19-1.0)	1.1 (1.0-1.2)
Summary	0.69 (0.50-0.94)	1.1 (1.0-1.2)
Costovertebral Angle Tenderness on Physical Examination		
Wigton et al ³² (training set)	2.0 (1.2-3.4)	0.82 (0.71-0.95)
Wigton et al ³² (validation set)	1.4 (0.8-2.4)	0.91 (0.79-1.0)
Summary	1.7 (1.1-2.5)	0.86 (0.78-0.96)
Dipstick Urinalysis ^b		
Hurlbut and Littenberg ¹⁸	4.2	0.3

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

^aThe study by Wigton et al³² included 2 separate sets of patients evaluated by retrospective chart review: a training set and a validation set. Likelihood ratios in bold are significant.

^bA positive result was defined as leukocyte esterase positive or nitrite positive; a negative result was defined as both negative. The values were taken from a receiver operating characteristic curve, so no CI could be calculated.

because fewer studies were involved. The absence of vaginal discharge, a feature reported in only 3 studies, makes a UTI more likely whether or not this study²⁹ is included (LR, 3.1 [95% CI, 1.0-9.3] for all studies vs LR, 1.7 [95% CI, 1.3-2.2] when excluded). The presence of vaginal discharge still decreases the likelihood of a UTI whether or not the study by Komaroff et al²⁹ is included (LR, 0.34 [95% CI, 0.14-0.86] for all studies vs LR, 0.60 [95% CI, 0.39-0.91] when the study is excluded).

COMMENT

Symptoms suggestive of UTI are common complaints of young women seeking urgent medical care. Although textbooks of clinical medicine¹⁵⁻¹⁷ routinely mention many of the symptoms and signs of UTI, the overall accuracy of these symptoms and signs has not previously been critically and systematically evaluated. A clear understanding of the value of each of these diagnostic tests may enable physicians to make more informed decisions about the choice of specific tests and management options.

Rule Out Complicated Urinary Tract Infection

The initial step is to be certain that the patient does not have a complicated UTI as defined by the factors listed earlier (see “Definitions” section). The probability of UTI in patients with risk factors for a complicated infection is not known because these patients were not included in the studies identified by our search. Such patients may be at greater risk of treatment failure,¹⁰ and clinicians may want to consider early urine culture and empirical treatment as shown at the top of the proposed algorithm (Figure 51-1).

Pretest Probability and the Diagnostic Value of Presenting to a Clinician

With a standard evidence-based technique,²¹ a clinical encounter begins with an estimation of the pretest probability of disease, followed by the application of 1 or more diagnostic tests to determine the posttest probability of disease. We consider the pretest probability of UTI to be equal to the prevalence observed in studies of asymptomatic bacteriuria, or approximately 5%.^{11,12} In this review, 5 studies reported

the prevalence of UTI in patients presenting with 1 or more symptoms of acute UTI, and the summary prevalence was 48% (95% CI, 41%–55%).

The probability of UTI changes substantially when a patient presents to a clinician, increasing from 5% (in historical controls without symptoms) to approximately 50% (in patients in the included studies who presented with 1 or more symptoms). This change in probability corresponds to an LR of 19, representing a powerful “diagnostic test.” Clinically, it is useful to know that patients who present with 1 or more symptoms of UTI have a high probability of infection. Because all of the studies included in this review evaluated the diagnostic value of symptoms and signs after patients presented to a clinician, the relevant pretest probability for these tests is 50%.

Although the pretest probability of UTI in the average patient who presents with 1 or more symptoms is approximately 50%, this varies considerably according to the individual’s risk profile. There are 3 well-established risk factors for acute UTI in young women: recent sexual intercourse,^{3,34–38} use of spermicide (on condoms or with diaphragms) during sexual intercourse,^{3,34–36,39,40} and history of UTI.^{3,36} Other risk factors, including a maternal history of UTI,³⁴ a history of childhood onset of UTI,³⁴ and the presence of bacterial vaginosis,⁴¹ also have been found to be associated with UTI. The presence of any of these risk factors increases the pretest probability of UTI and should be considered when evaluating patients. Unfortunately, the diagnostic power of these risk factors (sensitivity, specificity, or LRs) is not known, because the majority of studies assessing these risk

Table 51-3 Likelihood Ratios for Combinations of Symptoms

Symptom Combinations	Overall LR Using Combinations of Individual Symptoms ^a	Based on Data From Komaroff et al ²⁹
	Posttest Probability of UTI, % ^b	Summary LR ^c
Dysuria present	1.5	77
Frequency present	1.8	
Vaginal discharge absent	3.1	
Vaginal irritation absent	2.7	
Overall ^d	23	25
Dysuria absent	0.5	4
Vaginal discharge or irritation present	0.3 or 0.2	
Overall	0.1–0.2	0.3
Dysuria or frequency present	1.5 or 1.8	9
Vaginal discharge or irritation present	0.3 or 0.2	
Overall	0.3–0.5	0.7

Abbreviations: LR, likelihood ratio; UTI, urinary tract infection.

^aThe overall LR was calculated by multiplying the summary LRs from Table 2 for each of the findings in each set of symptom combinations. LRs < 1 are rounded off to make computation easier when combining findings.

^bThe pretest probability of UTI in the study by Komaroff et al²⁹ was 12% (the prevalence of UTI in the study).

^cLikelihood ratios were calculated from the observed change in the pretest and posttest probability of UTI; confidence intervals cannot be calculated because the raw data were not available.

^dValues are rounded to nearest integer.

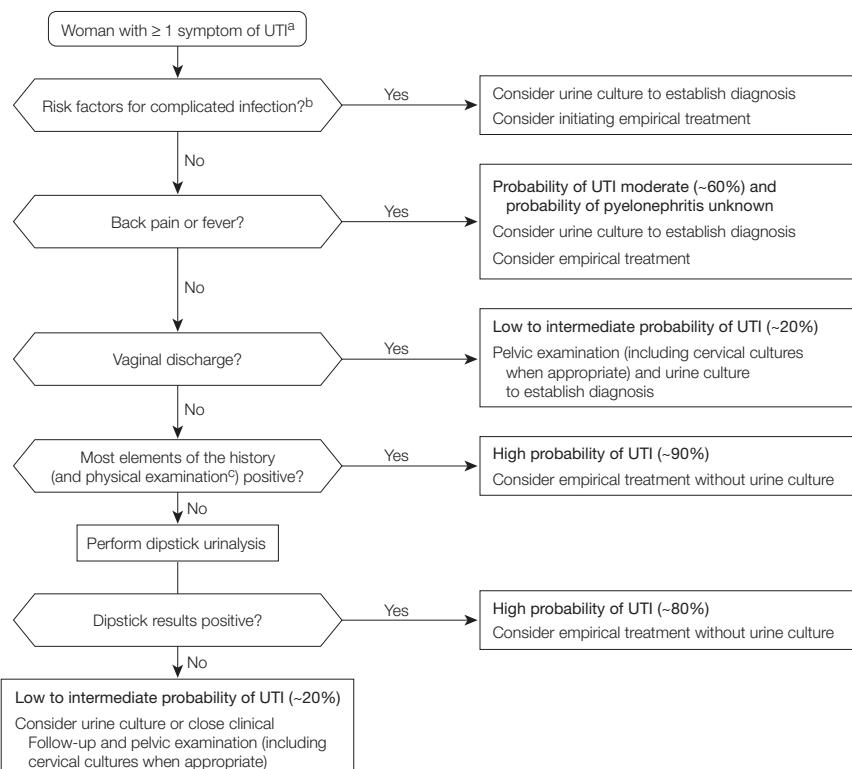


Figure 51-1 Proposed Algorithm for Evaluating Women With Symptoms of Acute Urinary Tract Infection

Abbreviation: UTI, urinary tract infection.

^aIn women who have risk factors for sexually transmitted diseases, consider testing for chlamydia. The US Preventive Services Task Force recommends screening for chlamydia for all women aged 25 years or younger and women of any age with more than 1 sexual partner, a history of sexually transmitted disease, or inconsistent use of condoms.⁵²

^bFor a definition of complicated UTI, see the “Definitions” section of the text.

^cThe only physical examination finding that increases the likelihood of UTI is costovertebral angle tenderness, and clinicians may consider not performing this test in patients with typical symptoms of acute uncomplicated UTI (as in telephone management).

factors used a case-control design or did not present sufficient data to calculate LRs.^{3,4,35-39,42} Further research is needed to determine the diagnostic power of these risk factors so that the information can be used during the clinical encounter to estimate the pretest probability of disease.

Refining Probability With the Medical History and Physical Examination

In the included studies, all diagnostic tests were evaluated by their ability to change the already high (50%) probability of UTI in the study population. Because these patients initially presented with at least 1 symptom, some of the power of each symptom was already “used up” by the time the patient presented to a clinician (and the probability of UTI increased from 5% to 50%). In a sense, the diagnostic power of the symptom is being “used” twice. Initially, the presenting symptom (most commonly dysuria or frequency) caused the patient to present to a clinician and was at least partially responsible for raising the probability of UTI from 5% to 50%. Subsequently, the value of the presenting symptom and all other potentially relevant symptoms was assessed after presentation to a clinician.

It is therefore not surprising that most of the individual symptoms and signs have LRs relatively close to 1.0 and therefore do not have great additional diagnostic power after presentation. The main exception to this finding is the history of vaginal discharge or vaginal irritation, which reduces the probability of UTI.

One study found that back pain and costovertebral angle tenderness were useful for predicting the presence of UTI.³² This study was a retrospective chart review of patients who had a urine culture in an emergency department, and it is possible that back pain and costovertebral angle tenderness were predictive of upper UTI (pyelonephritis). However, because none of the included studies performed a gold standard test for upper UTI, we were unable to determine whether individual symptoms and signs were more predictive of upper vs lower UTI. Most patients with symptoms suggestive of UTI and features classically associated with upper UTI (back pain, fever) are evaluated and treated for presumed pyelonephritis (Figure 51-1), even though the diagnostic accuracy of these signs and symptoms for predicting upper UTI is not known. Because most patients in the included studies did not have back pain and fever, we believe that the other symptoms evaluated in our review are most useful for predicting lower UTI (cystitis).

In contrast to the value of individual tests, certain combinations of symptoms result in large changes in the probability of UTI and represent powerful diagnostic tests. The combination of dysuria and frequency without vaginal discharge or irritation corresponds to an LR of 25. Although the combined LRs were generated from only 1 study of lower quality,²⁹ these LRs were similar to those found when multiplying the summary LRs for the individual symptoms, suggesting that they are reasonable estimates of the true diagnostic power of these combinations. In addition, another study⁴³ that was excluded from our analysis (because it

included an unknown number of asymptomatic patients) used the same combinations of symptoms and found similar positive predictive values and LRs.

Although evaluated in only 1 study,³³ self-diagnosis appears to be a useful diagnostic test (LR, 4.0) in women with recurrent UTI. Because this study did not perform urine cultures for women with mild or no symptoms, there is some uncertainty in the LR estimates. Similarly, the study population consisted of mostly highly educated single white women, and it is not clear whether the results apply to other groups of women. Nonetheless, these findings suggest that women learn to recognize the symptoms of UTI and are able to accurately diagnose a new infection, a finding that deserves further study and may have important implications for treatment of this large group of patients.

Refining Probability Using Dipstick Urinalysis

Dipstick urinalysis alone is a moderately powerful diagnostic test (Table 51-2). If the dipstick is used alone, the posttest probabilities for women with symptoms of a UTI are 81% (positive result) and 23% (negative result).

A Diagnostic Algorithm for Evaluating Patients With Symptoms of Urinary Tract Infection

Figure 51-1 shows a proposed algorithm for evaluating patients with symptoms of UTI. Although the algorithm itself has not been prospectively studied, the recommendations are based on the posttest probabilities of UTI generated from the summary LRs in the current analysis (Table 51-2). In women with risk factors for a complicated UTI or with back pain, fever, or malaise (suggesting possible pyelonephritis), a urine culture with initial empirical treatment is recommended. If a woman reports a history of vaginal discharge, the posttest probability of UTI from this single historical item is reduced to 23%, and a pelvic examination to rule out a vaginal infection should be considered in addition to a dipstick urinalysis and urine culture.

The algorithm highlights the finding that the medical history and physical examination alone can substantially increase the posttest probability of UTI, effectively “ruling in” the diagnosis. Because the only physical examination finding that increases the probability of UTI is costovertebral angle tenderness, the physical examination may be omitted without a substantial loss of diagnostic power in patients without a history of vaginal discharge or irritation. With individual summary LRs, a patient with dysuria, frequency, and hematuria (but no back pain at this point in the algorithm) has a posttest probability of UTI of 81%; with the combined LR estimate of dysuria and frequency without vaginal discharge (LR, 25), the posttest probability of UTI is 96%. Given these high probabilities of UTI, clinicians should consider empirical treatment without urine culture or dipstick urinalysis.

Conversely, even mostly negative history responses, physical examination findings, and dipstick urinalysis results cannot reliably rule out the diagnosis of UTI in women without a history of vaginal discharge or irritation. For example, to generate

the lowest possible posttest probability of disease, a woman must still present with at least 1 symptom. If she presents with frequency (LR, 1.8) with no dysuria (LR, 0.5) and no back pain (LR, 0.8) (the only 2 negative symptoms other than vaginal symptoms), a negative dipstick result (LR, 0.3), and no other positive symptoms, her posttest probability of disease is still 18%, which is considerably higher than the prevalence of asymptomatic bacteriuria in the population (5%). Although we do not address the optimum treatment of such patients, we believe that the relatively high probability of UTI (~20%) warrants a urine culture (Figure 51-1), an approach that has been supported by others.¹⁰ Clinicians may also want to consider performing a pelvic examination, especially in patients at high risk for sexually transmitted disease or if the urine culture result is negative and symptoms persist. As noted, it is theoretically possible to rule out UTI in women who present with vaginal discharge, in which the lowest possible posttest probability of disease is 6% (if they also have no dysuria, no back pain, a negative dipstick result, and no other positive symptoms). We recommend that clinicians consider obtaining a urine culture in patients with at least 1 urinary symptom and vaginal discharge because the posttest probability of disease will only rarely reach this lowest possible 6%.

If the medical history and physical examination are neither strongly positive nor negative, a positive dipstick result still results in a high posttest probability of disease (approximately 80%), and empirical therapy should again be considered without urine culture. In all of the scenarios in the algorithm, urine culture may be indicated, without regard to the posttest probabilities, if the patient has experienced recurrent infection and antibiotic resistance is suspected.

Older guidelines for the evaluation of patients with suspected UTI recommend urine culture in all patients, even in those found to have a high probability of UTI after the medical history and physical examination.^{29,44} More recent reviews and management strategies suggest that a diagnosis of UTI can be established in women who present with typical symptoms and are found to have a positive dipstick or urinalysis result (without obtaining a urine culture).^{10,45-48}

Unlike these treatment recommendations, our proposed algorithm (Figure 51-1) suggests that, in selected patients with mostly positive symptoms, the probability of UTI is so high (~90%) that empirical treatment may be considered without dipstick testing or urinalysis. A similar strategy was recently evaluated in a randomized trial comparing management via telephone with office evaluation in 72 women with suspected UTI.⁴⁹ The investigators found no difference in symptom scores or patient satisfaction with the 2 strategies. Previous studies examining the effect of symptom-based treatment of patients with suspected UTI (after a telephone call or office visit to a health care provider) have shown that empirical therapy decreases costs without increasing adverse outcomes.^{50,51} However, the main purposes of the current algorithm are to define the posttest probabilities of disease from specific clinical scenarios and to allow clinicians to make informed testing and treatment decisions based on their clinical judgment. Further research is needed to determine clinical

outcomes, costs, and patient satisfaction associated with different testing and treatment strategies for treating patients who present with specific constellations of symptoms of UTI.

CLINICAL SCENARIOS—RESOLUTIONS

In the first case, the woman has 2 symptoms of UTI (dysuria and frequency), has no vaginal discharge, and believes that her current symptoms are similar to those of previous episodes. These features all increase her probability of UTI, which is greater than 90%. Her sexual history does not suggest that she is at high risk for a sexually transmitted disease. With the algorithmic approach, the patient should be asked about risk factors for complicated infection, as well as symptoms classically associated with pyelonephritis (fever, back pain, nausea, vomiting). As has been shown, telephone evaluation and treatment of similar patients may be an appropriate strategy.^{49,50} In this patient, a positive dipstick urinalysis result would further increase the probability of UTI, whereas a negative result would not rule out infection.

In the second case, the woman has 2 symptoms of UTI (dysuria and frequency), as well as vaginal discharge (which decreases the probability of UTI and increases the probability of vaginal infection). A pelvic examination does not suggest a specific diagnosis and the dipstick urinalysis result is negative. The posttest probability of UTI is approximately 20%, illustrating that even a negative physical examination result and dipstick test result are insufficient to rule out UTI in a patient with 1 or more symptoms. A urine culture will help determine the need for treatment, and cervical cultures are indicated to rule out chlamydia and gonorrhea and help determine the cause of her symptoms.

THE BOTTOM LINE

In a woman who presents with 1 or more symptoms of UTI, the probability of infection is high (approximately 50%). Four symptoms (dysuria, frequency, hematuria, and back pain) and 1 sign (costovertebral angle tenderness) increase the probability of UTI when present. Combinations of symptoms can substantially increase the likelihood of UTI, effectively ruling in the disease according to the medical history alone. Patients with recurrent infection may be able to accurately self-diagnose UTI.

In contrast, the medical history and physical examination cannot reliably rule out UTI in women who present with urinary symptoms. Although 4 symptoms (absence of dysuria, absence of back pain, and a history of vaginal discharge or vaginal irritation) and 1 sign (vaginal discharge) decrease the probability of UTI, even combinations of symptoms, signs, and a negative dipstick result rarely decrease the probability of UTI below 20%. A urine culture and pelvic examination should be considered in patients who present with some symptoms of UTI but with mostly negative history responses and physical examination findings.

Dipstick urinalysis, which is a simple and inexpensive test, is moderately powerful and should be considered in women with appropriate urinary tract symptoms. If the dipstick result is positive, the probability of UTI is high, especially when combined with other positive findings from the medical history and physical examination. If the dipstick result is negative, the probability of disease is still relatively high (23%) and a urine culture should be considered to rule out infection.

Care should be taken to identify women with vaginal discharge or vaginal symptoms. If either is present, a pelvic examination and cervical culture are indicated to rule out infection caused by chlamydia⁵² or gonorrhea, as well as other vaginal infections that require definitive therapy. Similarly, in women with back pain, fever, or significant malaise, an office examination, combined with dipstick urinalysis and urine culture, may aid in the diagnosis of pyelonephritis, although the accuracy of individual tests for establishing upper UTI is not known.

Knowledge of the LRs for specific symptoms, signs, and diagnostic tests used to evaluate patients with suspected UTI may improve the ability of clinicians to more accurately predict the probability of infection in individual patients. It seems reasonable to offer empirical treatment when the probability of infection is high and to pursue additional diagnostic testing (eg, urine culture, pelvic examination, and cervical cultures) when the probability of UTI is low or intermediate. However, the actual cost-effectiveness of specific testing and treatment strategies is not clearly established, and prospective studies examining clinical benefits, adverse effects, costs, and patient satisfaction with specific approaches are needed.

Author Affiliations at the Time of the Original Publication

University of California, San Francisco, and General Internal Medicine Section, San Francisco Veterans Affairs Medical Center, San Francisco (Dr Bent); Department of Medicine, University of Michigan Medical School, Ann Arbor (Dr Nallamothu); Department of Medicine, Durham Veterans Affairs Medical Center and Duke University School of Medicine, Durham, North Carolina (Dr Simel); Health Services Research and Development Center of Excellence, Veterans Affairs Puget Sound Health Care System, Seattle, Washington (Dr Fihn); and Department of Medicine, Ann Arbor Veterans Affairs Medical Center and Patient Safety Enhancement Program, University of Michigan Health System, Ann Arbor (Dr Saint).

Acknowledgments

We thank Lori Bastian, MD, PhD, Michael Hayden, MD, Kathleen Klink, MD, and Joanne Piscitelli, MD, for their thoughtful comments on an earlier version of this work.

REFERENCES

1. Schappert SM. *National Ambulatory Medical Care Survey: 1992 Summary*. Hyattsville, MD: National Center for Health Statistics; 1994:94. Advanced data from Vital and Health Statistics No. 253.
2. Kunin C. *Urinary Tract Infections: Detection, Prevention, and Management*. 5th ed. Baltimore, MD: Williams & Wilkins; 1997.
3. Hooton TM, Scholes D, Hughes JP, et al. A prospective study of risk factors for symptomatic urinary tract infection in young women. *N Engl J Med*. 1996;335(7):468-474.
4. Foxman B, Barlow R, D'Arcy H, Gillespie B, Sobel JD. Urinary tract infection: self-reported incidence and associated costs. *Ann Epidemiol*. 2000;10(8):509-515.
5. Stamm WE, Hooton TM. Management of urinary tract infections in adults. *N Engl J Med*. 1993;329(18):1328-1334.
6. Johnson JR, Stamm WE. Urinary tract infections in women: diagnosis and treatment. *Ann Intern Med*. 1989;111(11):906-917.
7. Johnson JR, Stamm WE. Diagnosis and treatment of acute urinary tract infections. *Infect Dis Clin North Am*. 1987;1(4):773-791.
8. Berg AO. Variations among family physicians' management strategies for lower urinary tract infection in women: a report from the Washington Family Physicians Collaborative Research Network. *J Am Board Fam Pract*. 1991;4(5):327-330.
9. Stamm WE, Counts GW, Running KR, Fihn S, Turck M, Holmes KK. Diagnosis of coliform infection in acutely dysuric women. *N Engl J Med*. 1982;307(8):463-468.
10. Hooton TM, Stamm WE. Management of acute uncomplicated urinary tract infection in adults. *Med Clin North Am*. 1991;75(2):339-357.
11. Bengtsson C, Bengtsson U, Björkelund C, Lincoln K, Sigurdsson JA. Bacteriuria in a population sample of women: 24-year follow-up study: results from the prospective population-based study of women in Gothenburg, Sweden. *Scand J Urol Nephrol*. 1998;32(4):284-289.
12. Hooton T, Scholes D, Stapleton AE, et al. A prospective study of asymptomatic bacteriuria in sexually active young women. *N Engl J Med*. 2000;343(14):992-997.
13. Komaroff AL. Acute dysuria in women. *N Engl J Med*. 1984;310(6):368-375.
14. Metlay JP, Kapoor WN, Fine MJ. Does this patient have community-acquired pneumonia? diagnosing pneumonia by history and physical examination. *JAMA*. 1997;278(17):1440-1445.
15. Bickley L. *Bates' Guide to Physical Examination and History Taking*. Philadelphia, PA: Lippincott Williams & Wilkins; 1999.
16. DeGowin R. *DeGowin and DeGowin's Diagnostic Examination*. 6th ed. New York, NY: McGraw-Hill; 1994.
17. Orient J. *Sapira's Art and Science of Bedside Diagnosis*. 2nd ed. Philadelphia, PA: Lippincott Williams & Wilkins; 2000.
18. Hurlbut T, Littenberg B. The diagnostic accuracy of rapid dipstick tests to predict urinary tract infection. *Am J Clin Pathol*. 1991;96(5):582-588.
19. Oxman AD, Cook DJ, Guyatt GH. Users' guides to the medical literature, VI: how to use an overview. *JAMA*. 1994;272(17):1367-1371.
20. Holleman DR Jr, Simel DL. Does the clinical examination predict airflow limitation? *JAMA*. 1995;273(4):313-319.
21. Friedland D, Go A, Davoren J, et al. *Evidence-Based Medicine: A Framework for Clinical Practice*. Stamford, CT: Appleton & Lange; 1998.
22. Eddy DM, Hasselblad V. *FastPro v1.8: Software for Meta-analysis by the Confidence Profile Method*. San Diego, CA: Academic Press; 1992.
23. Eddy DM, Hasselblad V, Shacter RD. *Meta-Analysis by the Confidence Profile Method: The Statistical Synthesis of Evidence*. San Diego, CA: Academic Press; 1992.
24. Hasselblad V, Hedges LV. Meta-analysis of screening and diagnostic tests. *Psychol Bull*. 1995;117(1):167-178.
25. Gallagher D, Montgomerie J, North J. Acute infections of the urinary tract and the urethral syndrome in general practice. *BMJ*. 1965;1(5435):622-626.
26. Mond N, Percival A, Williams J, Brumfitt W. Presentation, diagnosis, and treatment of urinary-tract infections in general practice. *Lancet*. 1965;1(7384):514-516.
27. Lawson DH, Clarke A, McFarlane DB, McAllister TA, Linton AL. Urinary tract symptomatology in general practice. *J R Coll Gen Pract*. 1973;23(133):548-555.
28. Dans PE, Klaus B. Dysuria in women. *Johns Hopkins Med J*. 1976;138(1):13-18.
29. Komaroff AL, Pass TM, McCue JD, Cohen AB, Hendricks TM, Friedland G. Management strategies for urinary and vaginal infections. *Arch Intern Med*. 1978;138(7):1069-1073.
30. Nazareth I, King M. Decision making by general practitioners in diagnosis and management of lower urinary tract symptoms in women. *BMJ*. 1993;306(6885):1103-1106.

31. Wong ES, Fennell CL, Stamm WE. Urinary tract infection among women attending a clinic for sexually transmitted diseases. *Sex Transm Dis.* 1984; 11(1):18-23.
32. Wigton RS, Hoellerich VL, Ornato JP, Leu V, Mazzotta LA, Cheng IH. Use of clinical findings in the diagnosis of urinary tract infection in women. *Arch Intern Med.* 1985;145(12):2222-2227.
33. Gupta K, Hooton TM, Roberts PL, Stamm WE. Patient-initiated treatment of uncomplicated recurrent urinary tract infections in young women. *Ann Intern Med.* 2001;135(1):9-16.
34. Scholes D, Hooton TM, Roberts PL, Stapleton AE, Gupta K, Stamm WE. Risk factors for recurrent urinary tract infection in young women. *J Infect Dis.* 2000;182(4):1177-1182.
35. Remis RS, Gurwith MJ, Gurwith D, Hargrett-Bean NT, Layde PM. Risk factors for urinary tract infection. *Am J Epidemiol.* 1987;126(4):685-694.
36. Strom BL, Collins M, West SL, Kreisberg J, Weller S. Sexual activity, contraceptive use, and other risk factors for symptomatic and asymptomatic bacteruria: a case-control study. *Ann Intern Med.* 1987;107(6):816-823.
37. Foxman B, Geiger AM, Palin K, Gillespie B, Koopman JS. First-time urinary tract infection and sexual behavior. *Epidemiology.* 1995;6(2):162-168.
38. Foxman B, Gillespie B, Koopman J, et al. Risk factors for second urinary tract infection among college women. *Am J Epidemiol.* 2000;151(12):1194-1205.
39. Fihn SD, Boyko EJ, Normand EH, et al. Association between use of spermicide-coated condoms and *Escherichia coli* urinary tract infection in young women. *Am J Epidemiol.* 1996;144(5):512-520.
40. Fihn SD, Latham RH, Roberts P, Running K, Stamm WE. Association between diaphragm use and urinary tract infection. *JAMA.* 1985;254(2): 240-245.
41. Harmanli OH, Cheng GY, Nyirjesy P, Chatwani A, Gaughan JP. Urinary tract infections in women with bacterial vaginosis. *Obstet Gynecol.* 2000;95(5):710-712.
42. Foxman B, Marsh J, Gillespie B, Rubin N, Koopman JS, Spear S. Condom use and first-time urinary tract infection. *Epidemiology.* 1997;8(6): 637-641.
43. Berg AO, Heidrich FE, Fihn SD, et al. Establishing the cause of genitourinary symptoms in women in a family practice: comparison of clinical examination and comprehensive microbiology. *JAMA.* 1984;251(5):620-625.
44. Greenfield S, Friedland G, Scifers S, Rhodes A, Black WL, Komaroff AL. Protocol management of dysuria, urinary frequency, and vaginal discharge. *Ann Intern Med.* 1974;81(4):452-457.
45. Bacheller CD, Bernstein JM. Urinary tract infections. *Med Clin North Am.* 1997;81(3):719-730.
46. Orenstein R, Wong ES. Urinary tract infections in adults. *Am Fam Physician.* 1999;59(5):1225-1234, 1237.
47. Hooton TM. A simplified approach to urinary tract infection. *Hosp Pract (Off Ed).* 1995;30(2):23-30.
48. Hooton TM. Practice guidelines for urinary tract infection in the era of managed care. *Int J Antimicrob Agents.* 1999;11(3-4):241-245.
49. Barry HC, Hickner J, Ebelle MH, Ettenhofer T. A randomized controlled trial of telephone management of suspected urinary tract infections in women. *J Fam Pract.* 2001;50(7):589-594.
50. Saint S, Scholes D, Fihn SD, Farrell RG, Stamm WE. The effectiveness of a clinical practice guideline for the management of presumed uncomplicated urinary tract infection in women. *Am J Med.* 1999; 106(6): 636-641.
51. O'Connor PJ, Solberg LI, Christianson J, Amundson G, Mosser G. Mechanism of action and impact of a cystitis clinical practice guideline on outcomes and costs of care in an HMO. *Jt Comm J Qual Improv.* 1996;22(10):673-682.
52. Berg AO. Screening for chlamydial infection: recommendations and rationale. *Am J Prev Med.* 2001;20(3 suppl):90-94.

This page intentionally left blank

Prepared by Stephen Bent, MD
Reviewed by Joanne T. Piscitelli, MD,
and Marisa D'Silva, MD

CLINICAL SCENARIO

A 20-year-old healthy woman calls her student health clinic to report 1 day of dysuria with increased frequency. There has been no vaginal discharge or irritation, fever, back pain, nausea, or vomiting. She is sexually active with 1 partner and they use condoms. The symptoms seem similar to those of a previous urinary tract infection (UTI). What is the patient's probability of UTI based solely on the information from the medical history? Should she come in for a physical examination or a dipstick urinalysis to provide additional evidence that she has a UTI? Can UTI be ruled out without a urine culture?

- One new study reports that the probability of UTI after a negative dipstick result is approximately 20%, agreeing with our original estimate and indicating that it is difficult to rule out UTI with the clinical examination and dipstick urinalysis testing.
- "Telephone diagnosis" may be a reasonable option for patients without risk factors for complicated UTI who call with dysuria or urinary frequency, although current studies lack power to determine whether telephone diagnosis leads to an increase in pyelonephritis or sexually transmitted diseases.

Details of the Update

In women who present with 1 or more urinary tract symptoms compatible with UTI, the pretest probability is estimated to be 48%. A study of the diagnostic accuracy of dipstick urinalysis assessed 277 consecutive women presenting with symptoms suggestive of UTI.¹ In this study, all women received a urine culture, and the culture result was positive in 168 patients (incidence of UTI, 168/277 = 61%). For a positive dipstick urinalysis result, the LR is 1.5 (95% confidence interval [CI], 1.3-1.8), whereas the likelihood ratio (LR) for a normal dipstick result (negative LR) was 0.19 (95% CI, 0.10-0.36). In this study, a positive dipstick result was a less powerful predictor of UTI than the summary estimate from a previously published systematic review of 51 studies.² However, these findings agree with our original assessment that a normal dipstick urinalysis result does not lower the probability of UTI enough to rule out infection.

Two articles^{3,4} located by our search were previously discussed in a letter to the editor following the original Rational Clinical Examination article. Both articles involved prospective recruitment of patients with symptoms suggestive of UTI, and both examined the diagnostic accuracy of signs and symptoms for predicting UTI. However, neither article used an acceptable gold standard in all patients. One article³ tested all patients with a dipstick urinalysis and sent cultures only when the dipstick result was positive. The other article⁴ did not state how the decision to perform cultures for patients was made, and only 63% of patients received a urine culture. Because both of these studies were subject to verification bias (gold standard applied only when a preliminary test result is

UPDATED SUMMARY ON URINARY TRACT INFECTION IN ADULT WOMEN

Original Review

Bent S, Nallamothu BK, Simel DL, Fihn SD, Saint S. Does this woman have an acute uncomplicated urinary tract infection? *JAMA*. 2002;287(20):2701-2710.

UPDATED LITERATURE SEARCH

We searched MEDLINE from September 2001 through July 2004, using the same strategy as in our original publication. Search terms included "urinary tract infection," "diagnostic tests," "physical examination," and "sensitivity and specificity." We also manually reviewed the bibliographies of all identified articles and contacted experts in the field to identify other relevant articles. The search identified 35 titles that were reviewed by 2 investigators. Four articles were deemed potentially relevant, although none addressed the clinical examination.

NEW FINDINGS

- There are no new data from high-quality studies that change our previous estimates of the diagnostic accuracy of signs and symptoms for predicting UTI.

positive), we chose not to add the results to the summary estimates generated in our original review that came from 5 level 1 studies (prospective, independent blind comparison of signs or symptoms to a gold standard among a large number [>50] of consecutive patients suspected of having a UTI). Because the reference standard test was not applied to all patients, the prevalence of UTI among women presenting with symptoms in these studies (25%³ and 36%⁴) may underestimate the true prevalence.

Two studies^{5,6} examined the use of telephone diagnosis and management for selected patients who present with symptoms of UTI but who are at low risk for complicated UTI (ie, no diabetes, pregnancy, immunosuppression, or known renal disease). These studies evaluated the treatment of patients after a presumed diagnosis was made according to the symptoms elicited from the patient during a telephone call. The first study⁵ was a population-based, before-and-after study, with concurrent control groups of women calling to report their symptoms of dysuria or urinary frequency. Among 3889 patients with presumed acute, uncomplicated UTI, use of the telephone guideline decreased office visits by 33% and led to a nearly 3-fold increase in the use of a guideline-recommended antibiotic. The authors found a nonsignificant increase in return visits for evaluation of a possible sexually transmitted disease after guideline implementation but cautioned that their study was not adequately powered to detect small increases in outcomes such as pyelonephritis or sexually transmitted diseases. A second study⁶ randomly assigned a similar population of 72 women without risk factors for complicated UTI to either a telephone management protocol or an office visit. All women received a urine culture and all were contacted at 3 and 7 days to determine symptom severity. The authors found that 64% of enrolled patients had positive urine culture results. All patients were treated with antibiotics in the telephone group, whereas 32 of 36 patients were treated in the clinic-visit group. There was no difference in the change in symptom scores or the rate of treatment failure between groups, likely because almost all patients received antibiotic treatment. These authors also observed that the sample size was inadequate to detect differences in adverse events between groups.

IMPROVEMENTS IN THE DATA PRESENTED IN THE ORIGINAL PUBLICATION

We revised Table 51-2 so that it now shows a similar number of significant digits for LRs < 1 and 1 to 10. We identified no data to suggest changes in our original estimates of the diagnostic accuracy for signs, symptoms, or dipstick urinalysis for predicting UTI in women. Although we believe that the best estimate of the prevalence of UTI among patients with suggestive symptoms comes from the level 1 studies in our original report (48%; 95% CI, 41%-55%), we believe that there may be significant variability in this estimate according to differences in clinical setting, patient characteristics, or geographic location.

CHANGES IN THE REFERENCE STANDARD

The reference standard remains an appropriately obtained urine specimen for culture.

RESULTS OF LITERATURE REVIEW

Individual findings do not have great diagnostic power to change the high pretest probability of UTI in women (~50%). One study from our original review suggests that multiplying LRs from individual symptoms generates a multivariate LR that is a reasonable estimate of the diagnostic accuracy of combined symptoms.⁷

EVIDENCE FROM GUIDELINES

The US Preventive Services Task Force⁸ recommends against screening for asymptomatic bacteriuria other than during pregnancy. No US federal or Canadian guidelines address the evaluation of women primary care patients who have symptoms compatible with UTIs.

Many experts previously recommended urine culture in all patients with suspected UTI, even in those found to have a high probability of UTI after the medical history and physical examination.^{7,9} More recent reviews and management strategies suggest that a diagnosis of UTI can be established in women who present with typical symptoms and are found to have a positive dipstick or urinalysis result (without obtaining urine culture).¹⁰⁻¹⁴

CLINICAL SCENARIO—RESOLUTION

Although the pretest probability of UTI in the average patient who presents with symptoms is approximately 50%, this patient also has dysuria, frequency, and no vaginal discharge or irritation. Her posttest probability of UTI is greater than 90%. The history-taking should include questions about risk factors for complicated UTI (diabetes, immunosuppression, pregnancy, known renal disease). In patients without these risk factors who have a high probability of UTI, 2 studies^{5,6} suggest that telephone diagnosis and management may be appropriate, although it is not clear whether such strategies increase the risk of adverse events because of untreated pyelonephritis or sexually transmitted disease.

In a patient who presents with an isolated symptom of UTI (such as dysuria), an office visit with a negative dipstick result decreases the probability of UTI to approximately 20%. Because many clinicians will think that this probability is still too high, they might choose a strategy of urine culture or close clinical follow-up and consider performing a pelvic examination to assess for other conditions. All patients who have risk factors for complicated UTI, as well as a report of back pain, fever, or vaginal discharge, require further evaluation.

URINARY TRACT INFECTION, WOMEN—MAKE THE DIAGNOSIS

PRIOR PROBABILITY

The pretest probability of UTI among women with compatible symptoms is 48% (95% CI, 41%-55%).

POPULATION FOR WHOM URINARY TRACT INFECTION SHOULD BE CONSIDERED

Urinary tract infection should be considered in all adult women who present with 1 or more suggestive symptoms (frequency, dysuria, hematuria, fever, flank, or abdominal pain). Women with complicated UTI from a functional or anatomical abnormality of the urinary tract may present differently.

DETECTING URINARY TRACT INFECTION IN ADULT WOMEN

Combinations of symptoms (Table 51-4) can substantially increase the probability of UTI, effectively ruling in the diagnosis according to the medical history alone. In contrast, the history and physical examination cannot reliably exclude the diagnosis of UTI in women who present with urinary symptoms. A urine culture and pelvic examination should be considered in patients who present with some symptoms of UTI but otherwise a mostly negative history for UTI, a normal physical examination result, and a normal dipstick urinalysis result.

Table 51-4 Univariate Findings and Multivariate Approach for Diagnosing Urinary Tract Infection in Adult Women

	Univariate Findings	
	Present	Absent
Dysuria	1.5 (1.2-2.0)	0.5 (0.3-0.7)
Frequency	1.8 (1.1-3.0)	0.5 (0.4-1.0)
Vaginal discharge	0.3 (0.1-0.9)	3.1 (1.0-9.3)
Vaginal irritation	0.2 (0.1-0.9)	2.7 (0.9-8.5)
	Abnormal	Normal
Dipstick result ^b	4.2	0.3

Multivariate Approach

Multiply the above individual LRs for combinations of findings (eg, dysuria present and vaginal discharge absent yields a combined LR = 4.7; dysuria absent and vaginal discharge present yields a combined LR = 0.15).

Abbreviations: CI, confidence interval; LR, likelihood ratio.

^aLRs < 1 are rounded off to make computation easier when combining findings.

^bThe dipstick values were selected from visual inspection of a summary receiver operating characteristic curve to maximize the accuracy, so CIs could not be determined.²

REFERENCE STANDARD TESTS

The reference standard remains an appropriately obtained urine specimen for culture.

REFERENCES FOR THE UPDATE

- Verest LF, van Esch WM, van Ree JW, Stobberingh EE. Management of acute uncomplicated urinary tract infections in general practice in the south of The Netherlands. *Br J Gen Pract*. 2000;50(453):309-310.
- Hurlbut TA 3rd, Littenberg B. The diagnostic accuracy of rapid dipstick tests to predict urinary tract infection. *Am J Clin Pathol*. 1991;96(5):582-588.
- Dobbs FF, Fleming DM. A simple scoring system for evaluating symptoms, history and urine dipstick testing in the diagnosis of urinary tract infection. *J R Coll Gen Pract*. 1987;37(296):100-104.
- Fahey T, Webb E, Montgomery AA, Heyderman RS. Clinical management of urinary tract infection in women: a prospective cohort study. *Fam Pract*. 2003;20(1):1-6.
- Saint S, Scholes D, Fihn SD, Farrell RG, Stamm WE. The effectiveness of a clinical practice guideline for the management of presumed uncomplicated urinary tract infection in women. *Am J Med*. 1999;106(6):636-641.
- Barry HC, Hickner J, Ebell MH, Ettenhofer T. A randomized controlled trial of telephone management of suspected urinary tract infections in women. *J Fam Pract*. 2001;50(7):589-594.
- Komaroff AL, Pass TM, McCue JD, Cohen AB, Hendricks TM, Friedland G. Management strategies for urinary and vaginal infections. *Arch Intern Med*. 1978;138(7):1069-1073.
- Screening for Asymptomatic Bacteriuria: Recommendation Statement. Rockville, MD: Agency for Healthcare Research and Quality; 2004. <http://www.ahrq.gov/clinic/uspstf/uspsbact.htm>. Accessed June 16, 2008.
- Greenfield S, Friedland G, Scifers S, Rhodes A, Black WL, Komaroff AL. Protocol management of dysuria, urinary frequency, and vaginal discharge. *Ann Intern Med*. 1974;81(4):452-457.
- Hooton TM, Stam WE. Management of acute uncomplicated urinary tract infection in adults. *Med Clin North Am*. 1991;75(2):339-357.
- Bacheller CD, Bernstein JM. Urinary tract infections. *Med Clin North Am*. 1997;81(3):719-730.
- Orenstein R, Wong ES. Urinary tract infections in adults. *Am Fam Physician*. 1999;59(5):1225-1234, 1237.
- Hooton TM. A simplified approach to urinary tract infection. *Hosp Pract (Off Ed)*. 1995;30(2):23-30.
- Hooton TM. Practice guidelines for urinary tract infection in the era of managed care. *Int J Antimicrob Agents*. 1999;11(3-4):241-245; discussion 261-264.

This page intentionally left blank

CHAPTER 52

What Is Causing This Patient's Vaginal Symptoms?

Matthew R. Anderson, MD, MS

Kathleen Klink, MD

Andreas Cohnssen, MD

CLINICAL SCENARIOS

CASE 1 An otherwise healthy 33-year-old woman presents with a complaint of foul-smelling vaginal discharge. She is sexually active with 1 male partner. This is the first time she has had this symptom and is worried that it may represent a serious health problem. What diagnostic maneuvers—medical history, physical examination, and office laboratory tests—will allow the clinician to determine the cause of her symptoms?

CASE 2 A 35-year-old woman with 2 sexual partners in the last year complains of an itchy, smelly discharge. The pelvic examination reveals no vulvar or vaginal inflammation; a foamy, thin discharge with a pH of 5.0; and some bleeding at the cervix. The wet preparation reveals 2 clue cells per high-power field and, after thorough review of the slide, no motile organisms are seen. What is the chance that this patient has vaginal candidiasis, bacterial vaginosis, or vaginal trichomoniasis?

WHY IS THE CLINICAL EXAMINATION IMPORTANT?

Vaginal complaints are common in primary care. They are the most common reason for gynecologic consultation and account for approximately 10 million office visits annually.¹ Current recommendations for the diagnosis of vaginal complaints in premenopausal women involve a vaginal examination and microscopy. The evaluation has traditionally been oriented toward the detection of vaginal candidiasis, bacterial vaginosis, and trichomoniasis, which are the 3 most common causes of vaginitis in this age group.²⁻⁴

Prevalence of these 3 conditions will vary, depending on the clinical setting. National figures show that 40% to 50% of patients with vaginal symptoms have bacterial vaginosis; 20% to 25% have vaginal candidiasis; and 15% to 20% have trichomoniasis.⁵ In the studies surveyed for this review, which involved symptomatic women presenting in primary care, the prevalence of vaginal candidiasis ranged from 17% to 39%;^{6,7} bacterial vaginosis, 22% to 50%;^{8,9} and trichomoniasis, 4% to 35%.^{10,11} The number of undiagnosed patients ranged from 7% to 72%.^{6,12}

Women who present with vaginal complaints often receive tests for gonorrhea or chlamydia, though the association between gonorrhea, chlamydia, and vaginal discharge is not confirmed.^{13,14} It would be prudent, however, to test for gonorrhea and chlamydia in sexually active patients who are younger than 25 years and in all patients who have fever, lower abdominal pain, a symptomatic sexual partner, a new sexual partner, or more than 1 sexual partner.¹⁴ Additional less common causes of vulvovaginal symptoms are infection with herpes

simplex¹⁵; allergic reactions to chemical irritants, latex,¹⁶ or semen¹⁷; mechanical irritation caused by lack of lubrication; and atrophic vaginitis in postmenopausal women.¹⁸

About 30% of women with vaginal complaints go without a diagnosis even after a complete evaluation using techniques more comprehensive than those usually available.^{8,19,20} Perhaps this explains why many clinicians appear to treat patients without performing a pH examination of the discharge or microscopy.²¹ In actual clinical practice, diagnoses of vaginal complaints do not show good agreement with diagnoses based on cultures.²² These concerns led us to evaluate the role of the clinical examination in the diagnosis of vaginal complaints.

Point-of-care testing for vaginal complaints is a new and rapidly evolving field. A number of commercially available office kits use a vaginal discharge sample to diagnose bacterial vaginosis,²³ trichomoniasis,²³ and vaginal candidiasis.⁶ A systematic review of these diagnostic kits is, however, beyond the scope of this article.

How to Elicit Symptoms and Signs

Elicitation of Symptoms

Patients who have vaginitis generally complain of some combination of discharge, odor, irritation, or itch. Discharges are characterized by color (clear, white, green, gray, yellow), consistency (thin, thick, curdlike), and amount (more or less than usual). We could locate no scale that allows the patient to quantify precisely the amount of her discharge.

Signs

Patients may have irritation manifested as erythema, excoriation, or discharge on the perineum or introitus. The discharge is sampled during a speculum examination with a

swab from the posterior fornix or picked up on the speculum. Some clinicians ask patients to provide a self-collected sample of their vaginal discharge.²⁴

The sample can be tested for pH with phenaphthazine paper. When gel is used on the speculum, care must be taken not to contaminate the sample because the pH may become altered. In addition, semen, douches, and intravaginal medication can all make the vaginal pH more basic.

Characteristic findings on the wet mount are shown in **Figure 52-1**. Microscopy is performed by placing a drop of vaginal fluid on 2 slides. A drop of saline is mixed with the discharge on one slide, whereas a drop of 10% potassium hydroxide is placed on the second slide. The examiner then “whiffs” the potassium hydroxide slide to determine the presence of the characteristic fishy (amine) odor of bacterial vaginosis. The potassium hydroxide slide is set aside or put on a warmer. The other vaginal sample is examined under $\times 400$ power for trichomonads, clue cells, yeasts, presence or absence of lactobacilli (long rods²⁵), and the presence of leukocytes. Clue cells are epithelial cells with a finely granulated cytoplasm and indistinct borders,²⁶ which appear to have been coated with sand. The potassium hydroxide slide is examined for yeast. Yeast may be seen on the saline preparation, obviating the need to perform the potassium hydroxide microscopic examination.

Two excellent resources exist for learning how to perform the wet mount examination and whiff test. The Seattle STD/HIV Prevention Training Center has produced a short, downloadable instructional video.²⁷ The video illustrates the technique of the wet mount examination and includes clips of common findings such as yeast, clue cells, and motile trichomonads. For those more comfortable with

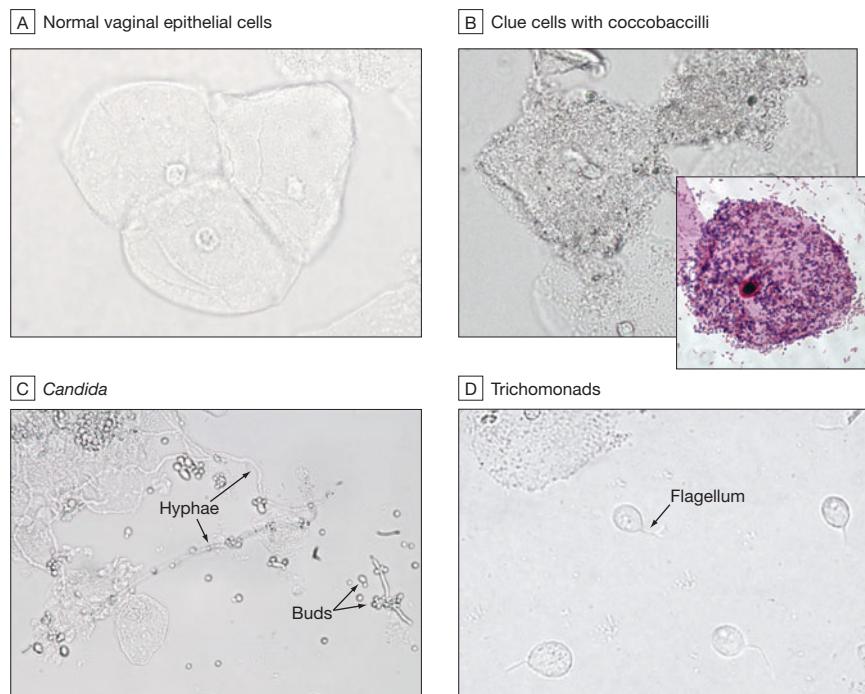


Figure 52-1 Microscopic Examination of Vaginal Samples

A, Normal saline wet mount showing a clump of 3 normal vaginal epithelial cells (original magnification, $\times 600$). Reproduced with permission from William L. Thelmo, MD. B, Normal saline wet mount showing 2 clue cells (original magnification, $\times 400$). Inset, Gram stain demonstrating how coccobacilli on the surface of vaginal epithelial cells create the characteristic granular appearance and indistinct borders of clue cells (original magnification, $\times 1000$). Reproduced with permission from Lorna Rabe, Magee-Womens Research Institute, Pittsburgh, Pennsylvania. C, Normal saline wet mount showing numerous *Candida* hyphae and buds (original magnification, $\times 400$). Reproduced with permission from Lorna Rabe. D, Normal saline wet mount showing 4 trichomonads. Trichomonads can often be identified easily because of their characteristic jerky motility (original magnification, $\times 600$). Reproduced with permission from the Medical Laboratory Evaluation proficiency testing program of the American College of Physicians Services Inc.

paper materials, the Association of Professors of Gynecology and Obstetrics' pamphlet on the diagnosis of vaginitis²⁸ contains photographs of the methods and findings of the wet mount examination.

Under the Clinical Laboratory Improvement Act, the wet mount examination is considered a moderately complex test, and the practitioner's laboratory must obtain a Certificate of Provider-Performed Microscopy Procedures from the local state health department.²⁹

METHODS

Search Strategy

We undertook a MEDLINE review of the literature from 1966 through April 2003, combining the term "diagnosis" with the terms "vaginitis," "vaginal discharge," "candidiasis," "bacterial vaginosis," and "trichomoniasis." We reviewed more than 500 abstracts and obtained a copy of articles (>100) that appeared likely to meet our review criteria. We also examined all articles mentioned in the most recent American College of Obstetricians and Gynecologists *Technical Bulletin*.³ Each article was reviewed by at least 1 author and in ambiguous cases by all 3. Included articles and review articles were culled for further references. We attempted to contact the authors of all articles included in this review and to request additional references. We received replies from 7 authors, but no additional references were produced.

Inclusion and Exclusion Criteria

Articles were included if they (1) involved original research performed on symptomatic patients in a primary care setting (including sexually transmitted disease clinics), (2) compared a diagnostic test with a recognized criterion standard, (3) allowed the calculation of sensitivity or specificity, and (4) discussed tests that would provide diagnostic information during the course of the office visit. We excluded articles that reported on women treated in specialty or referral settings, those with recurrent or treatment-refractory vaginitis, or asymptomatic patients (for example, women treated for routine pelvic examination).

Evaluation of Methods

Eighteen articles met our inclusion and exclusion criteria and are listed in Table 52-1.^{6-12,23,30-39} We graded the articles' diagnostic methodologic quality on a 3-point scale (highest to lowest quality). The grading and criteria are listed in Box 52-1. A different quality score from other Rational Clinical Examination articles (see Table 1-7) was required, because the focus of our study involved 3 different types of vaginitis, each of which have different laboratory criterion standards.

Evaluation of Criterion Standards

The diagnostic criterion standard for vaginal candidiasis is a positive culture result or identification of yeast by

microscopy. Because many asymptomatic women have vaginal yeast colonization, it is not clear whether a positive culture result or microscopy alone confirms *Candida* as the cause of symptoms, yet this is the current diagnostic criterion standard. We accepted studies that used microscopy only as a criterion standard but considered these of lower quality.

We used the Amsel criteria⁴⁰ as the criterion standard for the diagnosis of bacterial vaginosis. Bacterial vaginosis is diagnosed when 3 of 4 findings are present: (1) a thin, homogeneous vaginal discharge; (2) clue cells; (3) positive whiff test; and (4) vaginal pH level higher than 4.5.⁴⁰ Several articles used either Gram stain or a positive culture for *Gardnerella vaginalis* as criterion standards, which we also accepted, although we did not consider this optimal.

The criterion standard applied to the diagnosis of trichomoniasis is a positive culture result. Immunofluorescence and polymerase chain reaction are probably equivalent to culture. We accepted studies that included identification of trichomonads by direct microscopy or Papanicolaou tests, although these were considered of lesser quality.

Data Extraction

Sensitivity, specificity, and likelihood ratios (LRs) were either taken directly from the article or calculated from data provided in the article. All of the authors extracted the data and computed sensitivity and specificity from each article independently. Disagreements were resolved by consensus. All data and any calculations were sent to the primary authors for their review. One author of an article¹² we included provided additional data that have been incorporated into this review. A fourth person independently verified all data points. The absence of standard definitions for a variety of symptoms and signs, along with ambiguous phrasing of terms, made it impossible to combine results across studies.

Statistical Analysis

Statistical analysis was performed using SPSS (version 10.0; SPSS Inc, Chicago, Illinois) and Stata (version 8; StataCorp, College Station, Texas) statistical software. When there were no patients in one of the 4 cells of a 2 × 2 table (true positive, false positive, false negative, true negative), the value 0.5 was added to each cell of the 2 × 2 table for calculating the LRs.

Results

Precision

Precision refers to the degree to which independent observers will find the same result when applying the same test. No study reported the precision of the tests reviewed in this article.

Accuracy of Symptoms

Tables 52-2 and 52-3 present the sensitivity, specificity, and LRs for all symptoms. The reviewed articles tested the following symptoms for their usefulness in the diagnosis of vaginal complaints: (1) characteristics of the discharge (quantity,

Table 52-1 Included Studies of Diagnostic Strategies for Vaginal Symptoms

Source, y	No. of Patients	Setting	Symptoms	Vaginal Candidiasis, No. (%)	Bacterial Vaginosis, No. (%)	Vaginal Trichomoniasis, No. (%)	Quality Score ^a	Criterion Standard
Abbott, ¹² 1995 ^b	71	Urban ED or walk-in clinic; Denver, CO	Vaginal itching, discharge, or pain	23 (32)	29 (41)	5 (7)	2	Candidiasis: culture only
Abu Shaqra, ³⁰ 2001	301	Private gynecologists; Zarka, Jordan	Vaginal discharge	78 (26)	90 (30)	9 (3)	2	Bacterial vaginosis: Nugent criteria ^c
Bennett et al, ¹¹ 1989	157	Urban ED; Kansas City, MO	Vaginal discharge	NA	NA	55 (35)	2	Trichomoniasis: culture, microscopy, immunofluorescence
Bleeker et al, ³¹ 1989 ^d	97	Urban general hospital gynecology clinic; Amsterdam, The Netherlands	Vaginal discharge	24 (25)	37 (38)	13 (13)	3	Bacterial vaginosis: Spiegel criteria ^e ; trichomoniasis: microscopy; candidiasis: microscopy
Borchardt et al, ³² 1992	69	3 Clinics (1 STD clinic); San Jose, Costa Rica	Not indicated	NA	NA	10 (15)	2	Trichomoniasis: culture
Briselden and Hillier, ²³ 1994	176	STD clinic; Seattle, WA	Genital complaints	NA	79 (45)	19 (11)	2	Bacterial vaginosis: clinical criteria; trichomoniasis: culture, microscopy
Bro, ⁷ 1989	361	General practices (n = 29); Aarhus, Denmark	Increased vaginal discharge, malodor, or pruritus	141 (39)	NA	NA	2	Candidiasis: culture, microscopy
Carlson et al, ⁶ 2000 ^f	124	Gynecology outpatient clinic; Helsinki, Finland	Suspected vaginitis	21 (17)	NA	NA	2	Candidiasis: culture
Chandeying et al, ¹⁰ 1998	240	University gynecology outpatient clinic; Songklia, Thailand	Vaginal discharge	53 (22)	91 (38)	10 (4)	3	Bacterial vaginosis: Amsel criteria ^g ; candidiasis: microscopy; trichomoniasis: microscopy
Eckert et al, ³³ 1998	774	STD clinic; Washington state	"A new problem"	186 (24)	294 (38)	116 (15)	2	Candidiasis: culture
Fule et al, ³⁴ 1990	200	Hospital gynecology clinic; Solapur, India	Abnormal vaginal discharge	NA	34 (17)	NA	2	Bacterial vaginosis: culture and exclusion of other causes
Holst et al, ³⁵ 1987	101	Community health center; Lund, Sweden	Genital malodor or abnormal vaginal discharge	23 (23)	34 (34)	9 (9)	2	Bacterial vaginosis: Amsel criteria ^g
Krieger et al, ³⁶ 1988	600	STD clinic; Seattle, WA	"New problems"	NA	NA	90 (15)	2	Trichomoniasis: culture
Livengood et al, ³⁷ 1990	67	2 Hospital gynecology clinics	NA	NA	67 (100)	NA	2	Bacterial vaginosis: Amsel criteria ^g
O'Dowd and West, ⁹ 1987 ^h	162	Department of General Practice; Nottingham, England	Vaginal symptoms	NA	81 (50)	NA	3	Bacterial vaginosis: culture only
Ryu et al, ³⁸ 1999	177	University obstetrics/gynecology clinic; Seoul, Korea	Vaginal discharge	NA	NA	18 (10)	2	Trichomoniasis: culture
Schaaf et al, ⁸ 1990 ⁱ	123	County hospital family planning clinic or community-based women's health center; San Francisco, CA	Evaluation for vaginitis	32 (26)	27 (22)	9 (7)	2	Bacterial vaginosis: Amsel criteria ^g ; trichomoniasis: culture; candidiasis: culture
Wathne et al, ³⁹ 1994	101	Swedish community health center; Lund, Sweden	Vaginal discharge or malodor	23 (23)	34 (34)	9 (9)	2	Bacterial vaginosis: Amsel criteria ^g ; trichomoniasis: culture; candidiasis: culture

Abbreviations: ED, emergency department; NA, information not reported; STD, sexually transmitted disease.

^aSee Box 52-1 for criteria for quality scoring.

^bAdditional unpublished data from this study were included in this review.

^cDetermined using criteria from Nugent et al.²⁵

^dTwenty-two patients were not diagnosed.

^eDetermined using criteria from Spiegel et al.⁵⁰

^fSeventy-four patients were not diagnosed.

^gDetermined using criteria from Amsel et al.⁴⁰

^hNineteen patients were not diagnosed.

ⁱFifty-one patients were not diagnosed. Women with herpes or urinary tract infections were excluded.

^jData appear to be same as in Holst et al.³⁵ Data on bacterial vaginosis were reported differently in this article and have been excluded from our analysis.

color, consistency), (2) presence or absence of itching, (3) irritative symptoms (redness, pain/burning, swelling), (4) odor (present, fishy, or foul), (5) patient's self-diagnosis, (6) urinary tract symptoms, (7) bleeding, and (8) dyspareunia.

Discharge Characteristics

Patients' descriptions of their discharge do not appear useful diagnostically with 1 exception. A "cheesy" discharge increases the likelihood of candidiasis (LR, 2.4; 95% confidence interval [CI], 1.4-4.2), whereas a watery discharge makes it less likely (LR, 0.12; 95% CI, 0.02-0.82).

Itching

Several studies confirm that 70% to 90% of patients with vaginal candidiasis complain of itching (range of LRs, 1.4 to 3.3). Similarly, these studies show LRs ranging from 0.18 to 0.79 for women who do not have itching; thus, lack of itching decreases the likelihood of candidal infection. Itching symptoms are not useful for assessing the likelihood of bacterial vaginosis or trichomoniasis.

Irritative Symptoms

The limited data suggest that irritative symptoms are slightly useful in the diagnosis of candidiasis. Erythema increases the likelihood of candidiasis slightly (LR, 2.0; 95% CI, 1.5-2.8); its absence decreases its likelihood (LR, 0.84; 95% CI, 0.76-0.92).

Odor

The presence of an odor perceived by the patient decreases the likelihood of candidiasis (range of LRs, 0.35 to 0.48), whereas the absence of an odor increases its likelihood (range

Box 52-1 Criteria for Quality Scoring

LEVEL 1

Explicit inclusion and exclusion criteria.

More than 95% of patients received specified diagnostic evaluation including criterion standard.

More than 2 persons performed the diagnostic test, and a measure was made of interobserver variability.

Sensible normal range defined for continuous variables (when applicable) and criterion standards were used (Amsel⁴⁰ criteria for bacterial vaginosis, culture for vaginal trichomoniasis, and culture for vaginal candidiasis).

(No studies met all level 1 criteria.)

LEVEL 2

Level 2 studies failed 1 or more level 1 criteria or used the following criterion standards: for bacterial vaginosis, Amsel⁴⁰ modification, Spiegel,⁵⁰ Nugent,²⁵ culture and exclusion of other causes; for vaginal trichomoniasis, polymerase chain reaction, immunofluorescence; and for vaginal candidiasis, culture.

(Fifteen studies met level 2 criteria.)

LEVEL 3

Level 3 studies failed 1 or more level 1 criteria or used the following criterion standards: for bacterial vaginosis, *Gardnerella* culture; for vaginal trichomoniasis, microscopy or Papanicolaou test; and for vaginal candidiasis, microscopy.

(Three studies met level 3 criteria.)

Table 52-2 Accuracy of Symptoms for Diagnosis of Vaginal Candidiasis or Bacterial Vaginosis

Symptom	Diagnosis	No. of Patients With Diagnosis	Sensitivity, %	Specificity, %	LR+ (95% CI)	LR- (95% CI)	Reference
Type of discharge described by patient							
Any	VC	32 ^a	72 (NS)	... ^b	8
	BV	27 ^a	59 (NS)	8
	BV	67	91	38
Cheesy	VC	23	65	73	2.4 (1.4-4.2)	0.48 (0.27-0.86)	12
Increased	VC	186	NS	34
	BV	34	59	67	1.8 (1.2-2.8)	0.61 (0.40-0.95)	36
Watery	VC	23	4	63	0.12 (0.02-0.82)	1.5 (1.2-1.9)	12
White	VC	32 ^a	41 (NS)	8
	VC	186	NS	34
Yellow	VC	32 ^a	19 (NS)	8
	VC	186	NS	34
	BV	27 ^a	26 (NS)	8
Malodor or odor	VC	23	26	46	0.48 (0.23-1.0)	1.6 (1.1-2.4)	12
	VC	32 ^a	16 (NS)	8
	VC	23	21	37	0.35 (0.16-0.77)	2.1 (1.5-3.0)	40
	BV	34	97	40	1.6 (1.3-2.0)	0.07 (0.01-0.51)	36
	BV	67	73	38
	BV	27 ^a	41 (NS)	8
	BV	34	53	40

(continued)

Table 52-2 Accuracy of Symptoms for Diagnosis of Vaginal Candidiasis or Bacterial Vaginosis (*Continued*)

Symptom	Diagnosis	No. of Patients With Diagnosis	Sensitivity, %	Specificity, %	LR+ (95% CI)	LR- (95% CI)	Reference
Itching	VC	23	87	50	1.7 (1.3-2.4)	0.26 (0.09-0.78)	12
	VC	140	79	58	1.8 (1.6-2.2)	0.38 (0.27-0.53)	7
	VC	32 ^a	69 (NS)	8
	VC	23	91	47	1.7 (1.4-2.2)	0.18 (0.05-0.70)	40
	VC ^b	186	50	64	1.4 (1.2-1.7)	0.78 (0.67-0.91)	34
	BV	34	41	37	0.66 (0.42-1.0)	1.6 (1.0-2.4)	36
Chief complaint	BV	27 ^a	67 (NS)	8
	VC	186	27	92	3.3 (2.4-4.8)	0.79 (0.72-0.87)	34
Irritation	BV	67	45	38
	BV	27 ^a	48 (NS)	8
Pain or burning ^c	VC	32 ^a	69 (NS)	8
Redness ^c	VC	186	20	88	34
	VC	186	28	86	2.0 (1.5-2.8)	0.84 (0.76-0.92)	34
Swelling ^c	VC	186	24	92	1.4 (1.2-1.7)	0.78 (0.67-0.91)	34
Urinary tract							
Increased frequency of urination	VC	32 ^a	16 (NS)	8
Dysuria	VC	32 ^a	13 (NS)	8
	BV	27 ^a	11 (NS)	8
	BV	34	32	40
External dysuria	VC	186	33	85	2.2 (1.6-2.9)	0.79 (0.71-0.88)	34
Other							
"Another" yeast infection	VC	23	35	90	3.3 (1.2-9.1)	0.72 (0.53-1.0)	12
Abnormal bleeding	BV	67	4	38

Abbreviations: BV, bacterial vaginosis; CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio; NS, reported by author to be not significantly associated with diagnosis; VC, vaginal candidiasis.

^aPatient may have had more than 1 diagnosis.

^bEllipses indicate data not reported.

^cElicited by clinician.

Table 52-3 Accuracy of Symptoms for the Diagnosis of Vaginal Trichomoniasis

Symptom	No. of Patients With Diagnosis	Sensitivity, %	Specificity, %	LR+ (95% CI)	LR- (95% CI)	Reference
Type of discharge described by patient						
Any	8 ^a	75 (NS)	8
	17	65	29	0.90 (0.63-1.3)	1.2 (0.62-2.5)	39
White	8 ^a	13 (NS)	8
Yellow	8 ^a	50 (NS)	8
Malodor or odor						
Any	8 ^a	50 (NS)	8
"Fishy"	13	46	45	0.84 (0.45-1.6)	1.2 (0.68-2.1)	32
Itching	17	35	76	1.5 (0.74-3.0)	0.85 (0.59-1.2)	39
	8 ^a	75 (NS)	8
Irritation	8 ^a	63 (NS)	8
Urinary tract						
Increased frequency of urination	8 ^a	38 (NS)	8
Dysuria	8 ^a	38 (NS)	8
	17	0	97	0.64 (0.04-10)	1.0 (0.85-1.3)	39
Postcoital bleeding	17	0	97	0.9 (0.06-13)	1.0 (0.75-1.4)	39
Dyspareunia	17	6	96	1.4 (0.18-11)	0.98 (0.87-1.1)	39

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio; NS, reported by author to be not significantly associated with diagnosis.

^aPatient may have had more than 1 diagnosis.

^bEllipses indicate data not reported.

Table 52-4 Accuracy of Signs for the Diagnosis of Vaginal Candidiasis

Sign	No. of Patients With Diagnosis	Sensitivity, %	Specificity, %	LR+ (95% CI)	LR- (95% CI)	Reference
Type of discharge noted by clinician						
Any	32 ^a	87 (NS)	... ^b	8
Yellow	32 ^a	16 (NS)	8
White	32 ^a	63 (NS)	8
Curdy	140	16	97	6.1 (2.5-14)	0.86 (0.80-0.93)	7
Flocculent	23	43	84	2.7 (1.3-5.5)	0.67 (0.46-0.98)	40
Consistency of discharge						
Thick	32 ^a	52	8
Curdy	186	18	99	15 (6.4-36)	0.83 (0.78-0.89)	34
Curdy	53	72	100	130 (19-960)	0.28 (0.19-0.44)	10
Thin	32 ^a	48	8
Inflammation						
Any	140	46	78	2.1 (1.5-2.8)	0.69 (0.58-0.82)	7
Perineal edema or erythema	23	57	77	2.5 (1.3-4.6)	0.56 (0.35-0.92)	12
Vulvar edema	186	17	98	7.8 (4.2-15)	0.85 (0.79-0.91)	34
Erythema or edema	23	91	40
Vulvar erythema	186	54	79	2.5 (2.1-3.1)	0.58 (0.49-0.68)	34
Vaginal erythema	186	18	94	2.9 (1.9-4.5)	0.88 (0.82-0.94)	34
Vulvar excoriations	186	4	99	8.4 (2.3-31)	0.96 (0.93-0.99)	34
Vulvar fissures	186	17	96	4.6 (2.7-7.7)	0.86 (0.80-0.92)	34
Vaginal wall	32 ^a	23	8
Vulvar	53	40	95	8.2 (4.0-16)	0.63 (0.51-0.79)	10
Cervical mucopus	186	21	72	0.75 (0.55-1.0)	1.1 (1.0-1.2)	34
Odor noted by clinician						
Any	32 ^a	6	8
"Fishy"	24	0	28	0.03 (0-0.47)	2.9 (2.4-5.0)	32
Combined signs						
Curdy discharge or vulvar inflammation	53	81	95	17 (8.8-32)	0.20 (0.11-0.35)	10
Curdy discharge in presence of itching	53	77	100	150 (20-1000)	0.23 (0.14-0.37)	10

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio; NS, reported by author to be not significantly associated with diagnosis.

^aPatient may have had more than 1 diagnosis.

^bEllipses indicate data not reported.

of LRs, 1.6 to 2.1). Complaints of malodor (or odor) are so strongly associated with bacterial vaginosis that absence of malodor virtually ruled out the condition in 1 study (LR, 0.07; 95% CI, 0.01-0.51).³⁵ A fishy odor noticed by the patient is not helpful in diagnosing trichomoniasis.

Self-Diagnosis

Women who complain of having "another yeast infection" are more likely to have candidiasis (LR, 3.3; 95% CI, 1.2-9.1).

Urinary tract symptoms were not found to be associated with any of the 3 diagnoses in 1 study,⁸ whereas Eckert et al³³ found "external" dysuria associated with candidiasis.

Bleeding

In one study of 17 patients with trichomoniasis, no patient complained of postcoital bleeding.³⁸ Of 67 patients with bacterial vaginosis in the study by Livengood et al,³⁷ only 4% complained of abnormal bleeding.

Dyspareunia

Only 1 of 17 patients with trichomoniasis complained of dyspareunia, which is a nonsignificant association.³⁸

Accuracy of Signs

Tables 52-4 and 52-5 present the sensitivity, specificity, and LRs for all signs. We evaluated (1) characteristics of the discharge (amount, color, consistency), (2) inflammatory findings (edema, erythema, excoriations, tenderness, mucopus), and (3) odor.

Discharge

The finding of a discharge on examination does not distinguish between the 3 conditions. More than 60% of patients with these diagnoses have a discharge. A thick, curdy, or flocculent white discharge is strongly predictive of candidiasis (range of LRs, 2.7 to 130). The absence of these characteristics makes candidiasis less likely (range of LRs, 0.28 to 0.86). Women whose discharge is judged normal (LR, 0.11; 95% CI, 0.01-0.86) to mild (LR, 0.53; 95% CI, 0.37-0.75) are less likely to have bacterial vaginosis than women with moderate (LR, 2.5; 95% CI, 1.7-3.8) to profuse (LR, 3.0; 95% CI, 0.32-28) discharge. A white discharge makes bacterial vaginosis less likely (range of LRs,

Table 52-5 Accuracy of Signs for the Diagnosis of Bacterial Vaginosis or Vaginal Trichomoniasis

Sign	Diagnosis	No. of Patients With Diagnosis	Sensitivity, %	Specificity, %	LR+ (95% CI)	LR- (95% CI)	Reference
Type of discharge noted by clinician							
Any	BV	27 ^a	100 (NS)	... ^b	8
Vaginal discharge on vulvae	BV	67	64	38
Normal	BV	81	1	89	0.11 (0.01-0.86)	1.1 (1.0-1.2)	9
Mild	BV	81	33	37	0.53 (0.37-0.75)	1.8 (1.3-2.5)	9
Moderate	BV	81	62	75	2.5 (1.7-3.8)	0.51 (0.38-0.69)	9
Profuse	BV	81	4	99	3.0 (0.32-28)	0.98 (0.93-1.0)	9
Color or appearance							
Bloodstained	BV	81	1	99	1.0 (0.06-16)	1.0 (0.97-1.0)	9
Clear	BV	81	0	85	0.01 (0-0.16)	2.9 (1.6-5.4)	9
Green	BV	81	1	99	1.0 (0.06-16)	1.0 (0.97-1.0)	9
Mucoid	BV	33	3	100	1.6 (0.10-24)	0.99 (0.92-1.1)	35
Purulent, frothy	BV	33	30	51	0.62 (0.34-1.1)	1.4 (0.96-1.9)	35
Yellow	BV	81	60	85	4.1 (2.4-7.1)	0.46 (0.35-0.62)	9
	BV	27 ^a	30 (NS)	8
	VT	8 ^a	50 (NS)	8
	VT	9	89	93	14 (6.1-31)	0.12 (0.02-0.75)	40
White	BV	81	37	32	0.55 (0.40-0.75)	2.0 (1.4-2.8)	9
	BV	27 ^a	41 (NS)	8
	VT	8 ^a	13 (NS)	8
Curdy	BV	33	3	71	0.10 (0.01-0.74)	1.4 (1.1-1.7)	35
Consistency							
Homogeneous	VT	10	100	60	2.2 (1.7-2.8)	0.15 (0.02-1.0)	10
Thick	BV	27 ^a	12 (NS)	8
	VT	8 ^a	0 (NS)	8
Thin	BV	27 ^a	88 (NS)	8
	VT	8 ^a	100 (NS)	8
Transparent	BV	33	0	96	0.31 (0.02-6.3)	1.0 (0.97-1.1)	35
Inflammation							
Erythema or edema	VT	17	18	97	6.4 (1.6-26)	0.85 (0.68-1.1)	39
Vulvar	BV	67	1	38
	BV	67	12	38
Cervical	BV	67	10	38
Vaginal	BV	67	15	38
Vaginal wall	BV	27 ^a	33 (NS)	8
	VT	8 ^a	63 (NS)	8
Uterine/adnexal tenderness	BV	67	12	38
Odor noted by clinician							
Any	BV	27 ^a	78 (NS)	8
	VT	8 ^a	87 (NS)	8
	VT	8 ^a	50 (NS)	8
High cheese	BV	81	78	75	3.2 (2.1-4.7)	0.30 (0.19-0.45)	9

Abbreviations: BV, bacterial vaginosis; CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio; NS, reported by author to be not significantly associated with diagnosis; VT, vaginal trichomoniasis.

^aPatient may have had more than 1 diagnosis.

^bEllipses indicate data not reported.

Table 52-6 Accuracy of Office Laboratory Tests for the Diagnosis of Vaginal Candidiasis or Bacterial Vaginosis

Laboratory Test	Diagnosis	No. of Patients					Reference
		With Diagnosis	Sensitivity, %	Specificity, %	LR+ (95% CI)	LR- (95% CI)	
Microscopy							
Clue cells	VC	23 ^a	17	40	0.29 (0.12-0.73)	2.0 (1.4-3.0)	12
	VC	24	17	16	0.20 (0.08-0.49)	5.4 (3.0-9.5)	32
	VC	32 ^b	19	... ^c	8
Curved rods	BV	34	86	36
Mobiluncus-type rods	BV	67	53	38
Bacilli with corkscrew motility	BV	34	65	100	44 (6.2-310)	0.36 (0.23-0.57)	36
Lactobacilli scant or absent	BV	91	90	68	3.1 (2.4-3.9)	0.02 (0-0.11)	10
Yeast seen with potassium hydroxide	VC	23 ^a	61	77	2.7 (1.4-4.9)	0.51 (0.30-0.86)	12
	VC	186	56	34
	VC	32 ^b	63	8
	VC	23	83	40
	VC	21	38	94	6.5 (2.5-17)	0.66 (0.47-0.92)	6
	BV	27 ^b	19 (NS)	8
Yeast seen with saline	VC	23 ^a	65	75	2.6 (1.5-4.6)	0.46 (0.26-0.83)	12
Yeast seen with saline and methylene blue	VC	23 ^a	64	83	3.7 (1.9-7.6)	0.44 (0.25-0.77)	12
Yeast seen with Gram stain	VC	23 ^a	65	100	31 (4.4-220)	0.36 (0.20-0.62)	12
Trichomonads seen with saline	VC	32 ^b	0 (NS)	8
	BV	27 ^b	11 (NS)	8
Leukocytes more than epithelial cells	VC	23 ^a	13	75	0.52 (0.16-1.7)	1.2 (0.92-1.5)	12
	BV	34	36	36
Leukocytes on slide	VC	32 ^b	25 (NS)	8
	BV	27 ^b	15 (NS)	8
pH Level							
<4.5	VC	140	59	23	0.77 (0.66-0.90)	1.8 (1.3-2.4)	7
	VC	32 ^b	67	8
	VC	23	96	40
<4.9	VC	24	71	90	7.2 (3.4-15)	0.32 (0.17-0.61)	32
>5.0	VC	23 ^a	77	35	12
Leukocyte count (cells/high-power field)							
<10	BV	92	77	31
10-50	BV	92	18	31
>50	BV	92	4	31
Whiff test result positive	VC	23 ^a	17	45	0.31 (0.12-0.79)	1.9 (1.3-2.7)	12
	VC	32 ^b	13 (NS)	8

Abbreviations: BV, bacterial vaginosis; CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio; NS, reported by author to be not significantly associated with diagnosis; VC, vaginal candidiasis.

^aFor most tests, 1 to 2 patients had missing data for methylene blue, Gram stains, and whiff tests. For immunofluorescence tests, 16 patients had vaginal candidiasis.

^bA patient may have had more than 1 diagnosis.

^cEllipses indicate data not reported.

0.10 to 0.55). One study reports that bloodstained, green, clear, and purulent and frothy discharges are uncommon with bacterial vaginosis.³⁴ A yellow discharge increases the likelihood of both bacterial vaginosis (LR, 4.1; 95% CI, 2.4-7.1) and trichomoniasis (LR, 14; 95% CI, 6.1-31). All patients in one study with trichomoniasis had a homogeneous discharge.¹⁰

Inflammation

Signs included a general impression of vulvar inflammation by the clinician and specific signs such as vulvar or vaginal edema, erythema, fissures, or excoriations. The presence of these signs is associated with candidiasis (range of LRs, 2.1 to 8.4), although they can also occur in trichomoniasis (LR, 6.4; 95% CI, 1.6-26). The absence of

these signs does not exclude the diagnosis of either candidiasis or trichomoniasis. No studies allow calculation of the LR of inflammation for bacterial vaginosis, but the prevalence of a variety of inflammatory signs was low.

Odor

The presence of a “fishy” odor perceived by the clinician makes candidiasis unlikely (LR, 0.03; 95% CI, 0-0.47), whereas the absence of an odor increases the likelihood (LR, 2.9; 95% CI, 2.4-5.0). In contrast, the presence of a “high cheese” odor makes bacterial vaginosis more likely (LR, 3.2; 95% CI, 2.1-4.7). Data on clinically perceived odors in trichomoniasis are limited.

Accuracy of Office Laboratory Tests

Tables 52-6 and 52-7 present the sensitivity, specificity, and LRs for all office laboratory tests. We evaluated (1) microscopy for clue cells and other findings associated with bacterial vaginosis, (2) microscopy for yeast (using saline or potassium hydroxide), (3) microscopy for trichomonads, (4) microscopic evidence of inflammation, (5) measurement of vaginal pH, and (6) the whiff test.

Microscopy

The sensitivity of microscopy for yeast varies from 38% to 83%. Consequently, the absence of yeast rules against candidiasis but cannot exclude it (range of LRs, 0.46 to 0.66).

Because clue cells are part of the diagnostic criteria for bacterial vaginosis,⁴⁰ it is not possible to calculate LRs in this

condition. Bacilli with corkscrew motility are highly associated with bacterial vaginosis (LR, 44; 95% CI, 6.2-310). The finding of scant or no lactobacilli is common in bacterial vaginosis (LR, 3.1; 95% CI, 2.4-3.9), whereas finding normal levels of lactobacilli makes bacterial vaginosis unlikely (LR, 0.02; 95% CI, 0-0.11). The presence of clue cells makes candidiasis unlikely (range of LRs, 0.20 to 0.29) but has no effect on the diagnosis of trichomoniasis.

The identification of trichomonads in the wet mount diagnoses trichomoniasis, but their absence does not eliminate the diagnosis (range of LRs, 0.34 to 0.96).

Microscopic Evidence of Inflammation

The presence of many leukocytes seems relatively uncommon in candidiasis and bacterial vaginosis. One study, however, found all 9 patients with trichomoniasis had more leukocytes than epithelial cells.³⁹

pH Level

Four of 5 studies on pH in vaginal candidiasis reported that a majority of patients (59%-96%) had a normal pH level (variably defined as ≤ 4.5 or ≤ 4.9). A fifth study found 77% of candidiasis patients had a pH of greater than 5.0.¹² Thus, a majority, but not all, of the studies report that candidiasis is associated with a normal pH level. The pH in bacterial vaginosis should be high ($\text{pH} > 4.5$) and is incorporated into the case definition. A majority of patients (>90%) with trichomoniasis will have an increased pH level, but the specificity (51%) has been evaluated in only

Table 52-7 Accuracy of Office Laboratory Tests for the Diagnosis of Vaginal Trichomoniasis

Laboratory Test	No. of Patients With Diagnosis	Sensitivity, %	Specificity, %	LR+ (95% CI)	LR- (95% CI)	Reference
Microscopy						
Clue cells	13	69	33	1.0 (0.70-1.5)	0.93 (0.39-2.2)	32
	8 ^a	75 (NS)	... ^b	8
Yeast seen with potassium hydroxide	8 ^a	13 (NS)	8
Trichomonads seen with saline	8 ^a	75 (NS)	8
	9	78	40
	18	67	100	100 (14-740)	0.34 (0.17-0.64)	23
	10	0	100	4.5 (0.1-217)	0.96 (0.84-1.1)	33
	88	60	100	310 (43-2200)	0.40 (0.31-0.52)	37
	55	49	100	51 (7.1-360)	0.51 (0.40-0.67)	11
Leukocytes more numerous than epithelial cells	9	100	74	3.5 (2.3-5.2)	0.14 (0.02-0.87)	40
Leukocytes on slide	8 ^a	25	8
pH Level						
<4.5	8 ^a	17	8
>4.9	9	100	40
>5.4	13	92	51	1.9 (1.4-2.5)	0.15 (0.02-1.0)	32
Whiff test result positive	8 ^a	25 (NS)	8
	9	67	65	1.9 (1.1-3.3)	0.51 (0.20-1.3)	40

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio; NS, reported by author to be not significantly associated with diagnosis.

^aA patient may have had more than 1 diagnosis.

^bEllipses indicate data not reported.

1 study. Unfortunately, given the overlap between the pH levels in various conditions, it is hard to draw firm conclusions from the existing literature.

Whiff Test

A positive whiff test result makes candidiasis less likely (LR, 0.31; 95% CI, 0.12-0.79) but is positively associated with trichomoniasis (LR, 1.9; 95% CI, 1.3-2.7). A positive whiff test result is one of the diagnostic criteria for bacterial vaginosis.

Are These Symptoms and Signs Ever Normal?

The distinction between normal and abnormal in terms of vaginal symptoms is problematic. The primary literature on normal vaginal discharge is scant.⁴¹ It appears that a normal vaginal discharge increases at midcycle (because of an increase in cervical mucus),^{42,43} can be malodorous,⁴⁴ and may be accompanied by irritative symptoms (such as itch).⁴⁵ This problem is compounded by the fact that the vaginal pathogens identified by the current diagnostic approach can be found in asymptomatic women.^{46,47} *Gardnerella* is part of the normal vaginal flora.⁴⁸ Thus, the identification of microbes in a vaginal discharge does not prove that they create symptoms.

CLINICAL SCENARIOS—RESOLUTIONS

CASE 1 What is the appropriate diagnostic evaluation? No symptom has enough predictive power to allow the confident diagnosis of any of the 3 main causes of vaginitis. The wet mount examination remains the best way to make a diagnosis.

Symptoms and signs can suggest a particular diagnosis. Candidiasis is associated with itching, a cheesy discharge, redness, and self-diagnosis, whereas bacterial vaginosis is associated with increased discharge and a complaint of odor. A watery discharge makes candidiasis unlikely.

Inflammatory signs are relatively specific for vaginal candidiasis but are not always present and do occur in trichomoniasis. An absent or mild discharge makes bacterial vaginosis unlikely. Odor observed on examination occurs in bacterial vaginosis but not in candidiasis.

Most diagnoses are made by microscopy and the whiff test. Most studies (but not all) would support that candidiasis is associated with a normal pH level. Although the microscopic identification of yeast or trichomonads is diagnostic, these causes cannot be ruled out by negative findings on microscopy. The presence of clue cells makes candidiasis less likely. A lack of lactobacilli and the presence of bacilli with corkscrew motility are 2 findings highly associated with bacterial vaginosis.

CASE 2 What do you do when the diagnostic evaluation fails? Despite a full medical history, physical examination, and microscopy, the evaluation in this case does not pinpoint a cause of the patient's symptoms. There are several possibilities to consider in patients for whom the diagnostic evaluation is inconclusive. It is possible that the algorithm has failed to diagnose vaginal candidiasis or trichomoniasis; clinicians

should consider empirical therapy or further testing for trichomonads or *Candida*. Clinicians may want to consider less common causes of vaginal symptoms, including gonorrhea, chlamydia, herpes, or genital warts. Finally, there may be no pathologic condition causing the discharge, and the clinician may elect, after discussion with the patient, an approach of watchful waiting.

THE BOTTOM LINE

Our conclusions are subject to 2 important limitations. First, the LRs in these studies are not particularly robust. Second, despite dozens of articles devoted to the diagnosis of vaginal symptoms, we could locate only 18 that were useful in this review and none was of the highest methodologic quality.

Current research on vaginitis has a number of weaknesses. Studies on vaginitis often mix together women with symptoms and those presenting for follow-up examinations or routine care. By analyzing data from these distinct patient groups as if they were one, the research fails to address either the question of how to diagnose patients with symptoms or how to screen for asymptomatic disease. The vocabulary of physical findings is not standardized (ie, what is a cheesy discharge?), case definitions for candidiasis and trichomoniasis are not clear, and multiple criterion standards are used. Scant attention has been paid to interobserver variability, which is a key issue in the clinical examination. Furthermore, most studies concentrate on diagnosing one particular etiology. However, the task facing the clinician is to choose among different etiologies. When 2 pathogens are identified in a study (mixed infections), it is conceptually difficult to clarify whether one, both, or neither is responsible for the symptoms. Finally, the studies on trichomonas, with only one exception, had fewer than 20 patients; this is not a good base on which to draw solid conclusions (a fact emphasized by the large 95% CIs of the LRs).

In addition to these limitations, the existing diagnostic approach fails to diagnose approximately 30% of women with vaginal symptoms. The time is ripe for new approaches to these complaints.

Despite these limitations, primary care clinicians need to be skilled in the diagnosis of vaginal candidiasis, bacterial vaginosis, and trichomoniasis. Patients may also have concerns regarding the meaning of these symptoms for their health and personal relationships⁴⁹ and these concerns need to be addressed sensitively. Recognizing that the clinical examination is a limited tool in this setting presents the problem of finding ways to better diagnose and treat patients with vaginal symptoms. Vaginal symptoms may be the most common gynecologic complaint in primary care, but much remains to be learned about their clinical diagnosis.

Author Affiliations at the Time of the Original Publication

Department of Family and Social Medicine, Albert Einstein College of Medicine, Bronx, New York (Dr Anderson); Center for Family Medicine in the College of Physicians and Surgeons, Columbia University, New York, New York (Dr Klink);

and the Department of Family Practice, Beth Israel Medical Center/The Institute for Urban Family Health, New York, New York (Dr Cohrssen).

Acknowledgments

Jean Abbott, MD (University of Colorado, Denver), kindly provided additional data from her study, which we gratefully incorporated in this review. Clyde Schechter, MD, provided invaluable assistance with the statistical analysis. Kathleen O'Toole assisted in reviewing data points from the included articles. Lori Bastian, MD, MPH, Joanne Piscitelli, MD, and David L. Simel, MD, MHS, provided extensive and invaluable advice in the preparation of this article.

REFERENCES

- Kent HL. Epidemiology of vaginitis. *Am J Obstet Gynecol*. 1991;165(4 pt 2):1168-1176.
- Mou S. Vulvovaginitis. In: Rakel RE, Bopé ET, eds. *Conn's Current Therapy* 2003. Philadelphia, PA: WB Saunders; 2003:1149-1152.
- Technical Bulletin No. 226: Vaginitis*. Washington, DC: American College of Obstetricians & Gynecologists; 1996.
- Bickley LS. Acute vaginitis. In: Black ER, Bordley DR, Tape TG, Panzer RJ, eds. *Diagnostic Strategies for Common Medical Problems*. Philadelphia, PA: American College of Physicians-American Society of Internal Medicine; 1999:255-268.
- Mulley AG. Approach to the patient with a vaginal discharge. In: Goroll AH, Mulley AG, eds. *Primary Care Medicine: Office Evaluation and Management of the Adult Patient*. Philadelphia, PA: Lippincott Williams & Wilkins; 2000:702-707.
- Carlson P, Richardson M, Paavonen J. Evaluation of the Oricut-N dipslide for laboratory diagnosis of vaginal candidiasis. *J Clin Microbiol*. 2000;38(3):1063-1065.
- Bro F. The diagnosis of *Candida* vaginitis in general practice. *Scand J Prim Health Care*. 1989;7(1):19-22.
- Schaaf VM, Perez-Stable EJ, Borchardt K. The limited value of symptoms and signs in the diagnosis of vaginal infections. *Arch Intern Med*. 1990;150(9):1929-1933.
- O'Dowd TC, West RR. Clinical prediction of *Gardnerella vaginalis* in general practice. *J R Coll Gen Pract*. 1987;37(295):59-61.
- Chandeying V, Skov S, Kemapumuanus M, Law M, Geater A, Rowe P. Evaluation of two clinical protocols for the management of women with vaginal discharge in southern Thailand. *Sex Transm Infect*. 1998;74(3):194-201.
- Bennett JR, Barnes WG, Coffman S. The emergency department diagnosis of *Trichomonas* vaginitis. *Ann Emerg Med*. 1989;18(5):564-566.
- Abbott J. Clinical and microscopic diagnosis of vaginal yeast infection: a prospective analysis. *Ann Emerg Med*. 1995;25(5):587-591.
- Burstein GR, Murray PJ. Diagnosis and management of sexually transmitted diseases among adolescents. *Pediatr Rev*. 2003;24(4):119-127.
- Ryan CA, Courtois BN, Hawes SE, Stevens CE, Eschenbach DA, Holmes KK. Risk assessment, symptoms, and signs as predictors of vulvovaginal and cervical infections in an urban US STD clinic: implications for use of STD algorithms. *Sex Transm Infect*. 1998;74(suppl 1):S59-S76.
- Stuart-Harris C. The epidemiology and clinical presentation of herpes virus infections. *J Antimicrob Chemother*. 1983;12(suppl B):1-8.
- Chiu A, Kelly K, Thomason J, Otte T, Mullins D, Fink J. Recurrent vaginitis as a manifestation of inhaled latex allergy. *Allergy*. 1999;54(2):184-186.
- Ludman B. Human seminal plasma protein allergy: a diagnosis rarely considered. *J Obstet Gynecol Neonatal Nurs*. 1999;28(4):359-363.
- Bachmann G, Nevadunsky N. Diagnosis and treatment of atrophic vaginitis. *Am Fam Physician*. 2000;61(10):3090-3096.
- Berg AO, Heidrich FE, Fihn SD, et al. Establishing the cause of genitourinary symptoms in women in a family practice: comparison of clinical examination and comprehensive microbiology. *JAMA*. 1984;251(5):620-625.
- Mayaud P, ka-Gina G, Cornelissen J, et al. Validation of a WHO algorithm with risk assessment for the clinical management of vaginal discharge in Mwanza, Tanzania. *Sex Transm Infect*. 1998;74(suppl 1):S77-S84.
- Wiesenfeld HC, Macio I. The infrequent use of office-based diagnostic tests for vaginitis. *Am J Obstet Gynecol*. 1999;181(1):39-41.
- Allen-Davis JT, Beck A, Parker R, Ellis JL, Polley D. Assessment of vulvovaginal complaints: accuracy of telephone triage and in-office diagnosis. *Obstet Gynecol*. 2002;99(1):18-22.
- Briselden AM, Hillier SL. Evaluation of affirm VP microbial identification test for *Gardnerella vaginalis* and *Trichomonas vaginalis*. *J Clin Microbiol*. 1994;32(1):148-152.
- Gwyther RE, Addison LA, Spottswood S, Bentz EJ, Evens S, Abrantes A. An innovative method for specimen autocollection in the diagnosis of vaginitis. *J Fam Pract*. 1986;23(5):487-488.
- Nugent RP, Krohn MA, Hillier SL. Reliability of diagnosing bacterial vaginosis is improved by a standardized method of Gram stain interpretation. *J Clin Microbiol*. 1991;29(2):297-301.
- Gardner HL, Dukes CD. *Haemophilus vaginalis* vaginitis: a newly defined specific infection previously classified "nonspecific" vaginitis. *Am J Obstet Gynecol*. 1955;69(5):962-976.
- Seattle STD/HIV Prevention Training Network. Examination of vaginal wet preps. http://depts.washington.edu/nptc/online_training/wet_prep.html. Accessed June 15, 2008.
- Association of Professors of Gynecology and Obstetrics Educational Series on Women's Health Issues. *Diagnosis of Vaginitis*. Washington, DC: Association of Professors of Gynecology and Obstetrics; 1996.
- Clinical Laboratory Improvements Amendments (CLIA). How to obtain a CLIA Certificate. Centers for Medicare and Medicaid Services. <http://www.cms.hhs.gov/CLIA/downloads/HowObtainCLIACertificate.pdf>. Accessed June 15, 2008.
- Abu Shaqra QM. Bacterial vaginosis among a group of married Jordanian women: occurrence and laboratory diagnosis. *Cytobios*. 2001;105(408):35-43.
- Bleker OP, Folkertsma K, Dirks-Go SI. Diagnostic procedures in vaginitis. *Eur J Obstet Gynecol Reprod Biol*. 1989;31(2):179-183.
- Borchardt KA, Hernandez V, Miller S, et al. A clinical evaluation of trichomoniasis in San Jose, Costa Rica using the InPouch TV test. *Genitourin Med*. 1992;68(5):328-330.
- Eckert LO, Hawes SE, Stevens CE, Koutsy LA, Eschenbach DA, Holmes KK. Vulvovaginal candidiasis: clinical manifestations, risk factors, management algorithm. *Obstet Gynecol*. 1998;92(5):757-765.
- Fule RP, Kulkarni K, Jahagirdar VL, Saoji AM. Incidence of *Gardnerella vaginalis* infection in pregnant and non-pregnant women with non-specific vaginitis. *Indian J Med Res*. 1990;91:360-363.
- Holst E, Wathne B, Hovelius B, Mardh PA. Bacterial vaginosis: microbiological and clinical findings. *Eur J Clin Microbiol*. 1987;6(5):536-541.
- Krieger JN, Tam MR, Stevens CE, et al. Diagnosis of trichomoniasis: comparison of conventional wet-mount examination with cytologic studies, cultures, and monoclonal antibody staining of direct specimens. *JAMA*. 1988;259(8):1223-1227.
- Livengood CH III, Thomason JL, Hill GB. Bacterial vaginosis: diagnostic and pathogenetic findings during topical clindamycin therapy. *Am J Obstet Gynecol*. 1990;163(2):515-520.
- Ryu JS, Chung HL, Min DY, Cho YH, Ro YS, Kim SR. Diagnosis of trichomoniasis by polymerase chain reaction. *Yonsei Med J*. 1999;40(1):56-60.
- Wathne B, Holst E, Hovelius B, Mardh PA. Vaginal discharge—comparison of clinical, laboratory and microbiological findings. *Acta Obstet Gynecol Scand*. 1994;73(10):802-808.
- Amsel R, Totten PA, Spiegel CA, Chen KC, Eschenbach D, Holmes KK. Nonspecific vaginitis: diagnostic criteria and microbial and epidemiologic associations. *Am J Med*. 1983;74(1):14-22.
- Anderson M, Karasz A, Friedland S. Can vaginal symptoms ever be normal? a review of the literature. Presented at: North American Primary Care Research Group Annual Meeting; November 3, 2002; New Orleans, LA.
- Moghissi KS, Syner FN, Evans TN. A composite picture of the menstrual cycle. *Am J Obstet Gynecol*. 1972;114(3):405-418.
- Billings EL, Brown JB, Billings JJ, Burger HG. Symptoms and hormonal changes accompanying ovulation. *Lancet*. 1972;1(7745):282-284.
- Doty RL, Ford M, Preti G, Huggins GR. Changes in the intensity and pleasantness of human vaginal odors during the menstrual cycle. *Science*. 1975;190(4221):1316-1318.

45. Priestley CJ, Jones BM, Dhar J, Goodwin L. What is normal vaginal flora? *Genitourin Med.* 1997;73(1):23-28.
46. Bergman JJ, Berg AO. How useful are symptoms in the diagnosis of *Candida* vaginitis? *J Fam Pract.* 1983;16(3):509-511.
47. Blake DR, Duggan A, Quinn T, Zenilman J, Joffe A. Evaluation of vaginal infections in adolescent women: can it be done without a speculum? *Pediatrics.* 1998;102(4 pt 1):939-944.
48. Sobel JD. Vaginitis. *N Engl J Med.* 1997;337(26):1896-1903.
49. Karasz A, Anderson M. The vaginitis monologues: women's experiences of vaginal complaints in a primary care setting. *Soc Sci Med.* 2003;56(5):1013-1021.
50. Spiegel CA, Amsel R, Holmes KK. Diagnosis of bacterial vaginosis by direct Gram stain of vaginal fluid. *J Clin Microbiol.* 1983;18(1):170-177.

This page intentionally left blank

Prepared by Joanne T. Piscitelli, MD,
and David L. Simel, MD, MHS

Reviewed by Matthew Anderson, MD, MSc

CLINICAL SCENARIO

A 25-year-old woman who recently became sexually active presents with concerns about her new vaginal discharge and vaginal itching. She has not noticed an odor. When you do a speculum vaginal examination, should the discharge be examined microscopically for bacterial vaginosis, yeast, and trichomonas or will the appearance of the discharge be sufficient for diagnosis?

UPDATED SUMMARY ON VAGINITIS

Original Review

Anderson MR, Klink K, Cohrssen A. Evaluation of vaginal complaints. *JAMA*. 2004;291(11):1368-1379.

UPDATED LITERATURE SEARCH

Our literature search replicated that of the original article, confined to 2003 to April 2006. We identified 92 potential articles and reviewed the abstracts to find articles that included consecutive, prospectively identified patients with vaginal complaints in a primary care setting (primary care, general gynecology, or sexually transmitted disease clinics). Our focus was on identifying clinical studies that evaluated symptomatic women. We found 1 new article that met these standards. The literature search also uncovered 2 recent articles that assessed new bedside tests for bacterial vaginosis and trichomoniasis and that had data suitable for summarizing in likelihood ratios (LRs).

NEW FINDINGS

- The patient's symptom of an abnormal vaginal odor is a useful finding, but distinguishing bacterial vaginosis from vaginal candidiasis is not as efficient as proposed in the original report. Fortunately, the LR for bacterial vaginosis when the woman perceives an odor and for candidiasis when an odor is absent make perceived odor a useful symptom for clinical diagnosis. The patient's perception of an odor increases her likelihood of bacterial vaginosis

(summary LR, 2.2; 95% confidence interval [CI], 1.4-3.6), whereas the absence of an odor has the same effect in increasing the likelihood of vaginal candidiasis (summary LR, 2.2; 95% CI, 1.9-2.5).

- When clinicians do not have microscopes, point-of-care testing may prove useful for bacterial vaginosis and vaginal trichomoniasis.

Details of the Update

A recent study¹ includes the largest patient sample in which all 3 diagnoses were systematically evaluated. For each of the target conditions, the investigators reported data that allow calculation of the LRs for abnormal discharge, change in discharge, odor, vaginal pruritus, vaginal burning, and dysuria. A vaginal odor is the most useful symptom for distinguishing patients with bacterial vaginosis (odor symptoms present) from those with vaginal candidiasis (no perceived odor). No symptom worked for identifying women with vaginal trichomoniasis, because the LR CI for every symptom (both positive and negative LRs) includes 1. For both candidiasis and trichomoniasis, microscopic tests by the clinician are much more useful than the symptoms. The presence of yeast on a potassium hydroxide (KOH) preparation had an LR of 7.4 (95% CI, 3.8-15) vs culture, whereas the absence of yeast forms is less useful in identifying women who will have positive yeast culture results (LR, 0.80; 95% CI 0.74-0.87). The presence of trichomonads on a wet preparation slide was virtually diagnostic (LR, 22; 95% CI, 13-37). The absence of trichomonads does not rule out vaginal trichomoniasis because a culture result can still be positive (LR, 0.39; 95% CI, 0.29-0.53).

Although not reviewed in the original Rational Clinical Examination article on vaginitis, point-of-care testing for both bacterial vaginosis and vaginal trichomoniasis is gathering increased attention. Approved products are now available and marketed toward clinics that do not have access to microscopes or trained personnel for assessing the presence of clue cells (bacterial vaginosis) or trichomonads. Compared with the Amsel criteria,² the BVBlue Test (Gryphus Diagnostics, LLC, Birmingham, Alabama) has a positive LR of 9.8 (95% CI, 6.0-16) and a negative LR of 0.13 (95% CI, 0.08-0.21).³ The test uses a chromogenic assay for vaginal fluid sialidase produced by bacteria. Although the test takes

fewer than 10 minutes to perform, in this study the test kits were taken to a laboratory for processing. The findings require further study in a setting in which the clinic personnel interpret the results as a true “bedside” test, rather than sending the sample to a trained laboratory technician. A second type of point-of-care test for bacterial vaginosis incorporates a pH test and a test for amines (both of these are part of the Amsel criteria²). In a resource-poor environment, Azerbaijani women at a health fair were screened with the FemExam (Litmus Concepts, Inc, Santa Clara, California).⁴ Compared with the Amsel criteria,² a FemExam result positive for both pH and amines has a sensitivity of 92% for bacterial vaginosis, suggesting that it may be a reasonable substitute for the complete Amsel criteria² (positive LR, 7.5; 95% CI, 4.0-14). However, finding that both the pH and amine results are negative has an LR that is 0.45 (95% CI, 0.34-0.57), which is not low enough to rule out bacterial vaginosis, given its high pretest probability. Although most of the women in the study did have an abnormal vaginal discharge, not all were specifically seeking care for vaginitis. A point-of-care test for trichomoniasis (Xenostrip-Tv; Xenotope Diagnostics, San Antonio, Texas) identifies antigen to the protozoan. The test is highly efficient at confirming infection, with a positive LR of 361 (95% CI, 22-5845), but a normal result does not rule out vaginal trichomoniasis, with a negative LR of 0.52 (95% CI, 0.40-0.67).⁵

IMPROVEMENTS IN THE DATA PRESENTED IN THE ORIGINAL PUBLICATION

With data from the original Table 52-8 and from articles identified in the update,^{1,6} the prevalence of vaginal candidiasis, bacterial vaginosis, and vaginal trichomoniasis among women with vaginal complaints and presenting for care can be summarized. The summary estimates provide a reasonable anchor for making clinical decisions, though the data suggest geographic variability, which means that each provider needs a sense of prevalence in his or her own practice setting. The summary prevalences are as follows: bacterial vaginosis, 34% (95% CI, 28%-41%); vaginal candidiasis, 26% (95% CI, 22%-30%); and vaginal trichomoniasis, 10% (95% CI, 7%-15%). These prevalences support the notion that approximately 30% of women will have less common infections or remain undiagnosed after their evaluation.

We calculated summary LR for several of the symptoms in which the results were clinically consistent across studies. When considering the CI associated with these summary LRs, the clinician should have a better sense for the utility of the findings.

CHANGES IN THE REFERENCE STANDARD

None.

RESULTS OF LITERATURE REVIEW

Table 52-8 Univariate Findings for Vaginitis

Finding	Condition (No. of Studies) ^a	Summary LR+ (95% CI)	Summary LR- (95% CI)
Patient Symptoms			
Vaginal odor	Bacterial vaginosis (2)	2.2 (1.4-3.6)	0.30 (0.24-0.38)
	Candidiasis (3)	0.29 (0.20-0.43)	2.2 (1.9-2.5)
Vaginal itching	Candidiasis (5)	1.5 (1.3-1.8)	0.53 (0.33-0.86)
Microscopic Tests			
Yeast forms on a KOH preparation	Candidiasis (3)	4.8 (2.7-8.4)	0.78 (0.71-0.85)
Trichomonads seen with a saline preparation	Trichomoniasis (5)	46 (17-121)	0.50 (0.36-0.71)

Abbreviations: CI, confidence interval; KOH, potassium hydroxide; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

^aData are combined from that in Table 2 of the original Rational Clinical Examination article by Anderson et al⁷ and Table 6 in the article by Landers et al.¹

EVIDENCE FROM GUIDELINES

The Centers for Disease Control and Prevention funds an online training program developed by the Seattle STD/HIV Prevention Training Center that can be reviewed by clinicians who do office microscopy to diagnose vaginitis (http://depts.washington.edu/nnptc/online_training/wet_prep_video.html; accessed June 15, 2008).

Although bacterial vaginosis in pregnancy was not a focus of the review, the US Preventive Health Services Task Force⁸ evaluated the condition and found the evidence lacking to recommend for or against screening high-risk pregnant women for bacterial vaginosis. For clinicians who choose to screen, the task force observed that the Amsel criteria² are the accepted clinical criteria even though the “optimal” test has not been determined.

CLINICAL SCENARIO—RESOLUTION

The diagnosis of vaginitis requires microscopic examination of the vaginal discharge. Although you may not be able to determine a diagnosis in about 30% of patients, approximately 33% will have bacterial vaginosis, 25% will have candidiasis, and 10% will have trichomonas. The lack of a perceived odor makes candidiasis more likely (LR, 2.2), but the absence of the symptom is not conclusive. A thick or “curdy” discharge would be compatible with yeast, but women may have multiple infections. Thus, a diagnosis is best established by obtaining a specimen for: (1) measuring the pH; (2) preparing a slide for KOH assessment (evaluate the odor after application of KOH for the whiff test [bacterial vaginosis] and use the microscope to identify yeast forms); and (3) preparing a separate wet saline microscopic slide (for clue cells and trichomoniasis).

VAGINITIS—MAKE THE DIAGNOSIS

PRIOR PROBABILITY

Among women with vaginal symptoms, the most common diagnoses are bacterial vaginosis (34%), vaginal candidiasis (26%), and vaginal trichomoniasis (10%). The prevalence changes across regions, so clinicians should be familiar with the findings in their own clinics.

POPULATION FOR WHOM VAGINITIS SHOULD BE CONSIDERED

Vaginitis should be considered in any woman with concerns about a vaginal symptom that typically includes a combination of vaginal discharge, odor, irritation, or pruritus.

DETECTING THE LIKELIHOOD OF CAUSES OF VAGINITIS

Although the presence of odor helps identify women more likely to have bacterial vaginosis versus candidiasis, no symptoms reliably identify those with trichomoniasis (see Table 52-9). Thus, unless point-of-care tests become validated, a microscopic evaluation is required for identifying clue cells (bacterial vaginosis), yeast forms (vaginal candidiasis), or trichomonads (vaginal trichomoniasis). Clinicians who do office microscopy need appropriate training to recognize the findings (http://depts.washington.edu/nnpctc/online-training/wet_prep_video.html; accessed June 15, 2008).

REFERENCE STANDARD TESTS

Bacterial Vaginosis

The pragmatic reference standard consists of the Amsel criteria.² These require 4 different tests, of which at least 3 must have positive results: (1) a thin, homogenous vaginal discharge; (2) clue cells on microscopic examination; (3) positive whiff test; and (4) vaginal pH higher than 4.5.

Table 52-9 Likelihood Ratios of Symptoms and Microscopy for Vaginitis

Finding	Condition	LR+ (95% CI)	LR- (95% CI)
Patient Symptoms			
Vaginal odor (symptoms)	Bacterial vaginosis	2.2 (1.4-3.6)	0.30 (0.24-0.38)
	Candidiasis	0.29 (0.20-0.43)	2.2 (1.9-2.5)
Vaginal itching	Candidiasis	1.5 (1.3-1.8)	0.53 (0.33-0.86)
Odor, itching, vaginal burning, dysuria	Trichomoniasis	The LR+ and LR- have narrow CIs that include 1, suggesting they are of no value	
Microscopic Tests			
Yeast forms on a KOH preparation	Candidiasis (n = 3)	4.8 (2.7-8.4)	0.78 (0.71-0.85)
Trichomonads seen with a saline preparation	Trichomoniasis (n = 5)	46 (17-121)	0.50 (0.36-0.71)

Abbreviations: CI, confidence interval; KOH, potassium hydroxide; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

The reference standard test requires culture, though culture cannot distinguish between infections and colonization.

Trichomoniasis

The reference standard test in clinical research studies typically requires culture. However, in clinical practice the presence of trichomonads on a saline microscopic preparation is considered diagnostic, though the absence of trichomonads does not definitively rule out the condition.

REFERENCES FOR THE UPDATE

- Landers DV, Wiesenfeld HC, Heine P, Krohn MA, Hillier SL. Predictive value of the clinical diagnosis of lower genital tract infection in women. *Am J Obstet Gynecol.* 2004;190(4):1004-1010.^a
- Amsel R, Totten PA, Spiegel CA, Chen KC, Eschenbach D, Holmes KK. Nonspecific vaginitis: diagnostic criteria and microbial and epidemiologic associations. *Am J Med.* 1983;74(1):14-22.
- Bradshaw CS, Morton AN, Garland SM, Horvath LB, Kuzevska I, Fairley CK. Evaluation of a point-of-care test, BVBlue, and clinical and laboratory criteria for diagnosis of bacterial vaginosis. *J Clin Microbiol.* 2005;43(3):1304-1308.
- Posner SF, Kerimova J, Aliyeva F, Duerr A. Strategies for diagnosis of bacterial vaginosis in a resource-poor setting. *Int J STD AIDS.* 2005;16(1):52-55.
- Pillay A, Lewis J, Ballard RC. Evaluation of XenoStrip-Tv, a rapid diagnostic test for *Trichomonas vaginalis* infection. *J Clin Microbiol.* 2004;42(8):3853-3856.
- Luni Y, Munim S, Qureshi R, Tareen AL. Frequency and diagnosis of bacterial vaginosis. *J Coll Physicians Surg Pak.* 2005;15(5):270-272.
- Anderson MR, Klink K, Cohrsen A. Evaluation of vaginal complaints. *JAMA.* 2004;291(11):1368-1379.
- Berg AO. Screening for bacterial vaginosis in pregnancy: recommendations and rationale. *Am J Prev Med.* 2001;20(3 suppl):59-61.

^aFor the Evidence to Support the Update for this topic, see <http://www.JAMAEvidence.com>.

This page intentionally left blank

EVIDENCE TO SUPPORT THE UPDATE:

52

Vaginitis

TITLE Predictive Value of the Clinical Diagnosis of Lower Genital Tract Infection in Women.

AUTHORS Landers DV, Wiesenfeld HC, Heine P, Krohn MA, Hillier SL.

CITATION *Am J Obstet Gynecol.* 2004;190(4):1004-1010.

QUESTION Can experienced midlevel practitioners correctly diagnosis vaginitis among women with vaginal complaints?

DESIGN Prospective, independent.

SETTING Three sites in Pittsburgh, Pennsylvania: a student health center, a public sexually transmitted disease clinic, and a suburban public health clinic. Two of the clinicians were physician assistants and 1 was a nurse practitioner. Each clinician underwent specific instruction for the study and had competency testing in the bedside tests and microscopic studies.

PATIENTS Women aged 18 to 45 years and with untreated genital complaints consisting of vaginal discharge, odor, itching, or lower genital tract burning.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

Each patient filled out a questionnaire and then received a speculum examination. The clinician recorded evidence of mucopurulent cervicitis and evaluated vaginal secretions for color, viscosity, homogeneity, and odor after the addition of potassium hydroxide (KOH) to a sample of the vaginal secretions. The secretions were used to perform a KOH microscopic evaluation, pH testing, Gram stain, trichomonas, and yeast culture, along with endocervical cultures for sexually transmitted diseases and a Papainicolaou test. A clinical diagnosis for yeast was established from the microscopic KOH slide preparation that showed yeast. Trichomoniasis was diagnosed by observation of motile bacteria on the microscopic slide. Bacterial vaginosis was established by applying Amsel criteria.¹

The laboratory reference standard diagnosis for trichomonas and yeast was established by culture, and bacterial vaginosis was established by Gram stain examined for Nugent criteria.²

MAIN OUTCOME MEASURES

Sensitivity and specificity of the clinical diagnosis compared with the laboratory diagnosis. The sensitivity and specificity of the various vaginal complaints for bacterial vaginosis could be calculated from data in the article.

MAIN RESULTS

Among these 598 women with vaginal complaints, at least 1 microbiologic diagnosis was established in 79%. The distribution was bacterial vaginosis, 49%; vaginal yeast, 29%; trichomoniasis, 12%; and chlamydia or gonorrhea, 11%. Women could be coinfected by multiple organisms. Tables 52-10, 52-11, and 52-12 show the value of symptoms for bacterial vaginosis, candidiasis, and trichomoniasis. Table 52-13 displays the likelihood ratio (LR) of the clinical diagnosis for each infection compared to a laboratory criterion standard.

CONCLUSIONS

LEVEL OF EVIDENCE Level 2.

STRENGTHS The criteria for the clinicians' diagnoses are well outlined. Not only can the likelihood ratios (LRs) for the individual symptoms be reported but also the LRs for the bedside tests.

Table 52-10 Likelihood Ratios for Symptoms of Bacterial Vaginosis Compared With Amsel Criteria¹

Symptoms	LR+ (95% CI)	LR- (95% CI)
Vaginal odor	3.2 (2.6-3.9)	0.31 (0.25-0.39)
Change in discharge	2.2 (1.8-2.6)	0.38 (0.31-0.47)
Abnormal discharge	1.9 (1.7-2.2)	0.26 (0.19-0.35)
Dysuria	1.5 (0.97-2.3)	0.95 (0.89-1.1)
Vaginal burning	1.3 (0.96-1.9)	0.93 (0.86-1.0)
Vaginal pruritus	1.2 (0.97-1.5)	0.91 (0.81-1.0)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

Table 52-11 Likelihood Ratios for Symptoms of Vaginal Candidiasis Compared With Culture

Symptoms	LR+ (95% CI)	LR- (95% CI)
Vaginal pruritus	1.1 (0.87-1.4)	0.95 (0.84-1.1)
Vaginal burning	0.58 (0.37-0.90)	1.1 (1.0-1.2)
Change in discharge	0.47 (0.37-0.60)	1.9 (1.6-2.2)
Abnormal discharge	0.40 (0.32-0.50)	3.0 (2.4-3.7)
Dysuria	0.37 (0.19-0.71)	1.1 (1.0-1.2)
Vaginal odor	0.22 (0.15-0.32)	2.3 (2.0-2.7)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

Table 52-12 Likelihood Ratios for Symptoms of Vaginal Trichomoniasis Compared With Culture

Symptom	LR+ (95% CI)	LR- (95% CI)
Vaginal burning	1.1 (0.86-1.8)	0.98 (0.87-1.1)
Dysuria	1.1 (0.58-2.1)	0.99 (0.90-1.1)
Vaginal odor	1.0 (0.78-1.3)	1.0 (0.79-1.3)
Vaginal pruritus	1.0 (0.70-1.4)	1.0 (0.84-1.2)
Change in discharge	0.9 (0.7-1.2)	1.1 (0.87-1.4)
Abnormal discharge	0.81 (0.65-1.0)	1.3 (1.0-1.8)

Abbreviations: CI, confidence interval; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

Table 52-13 Likelihood Ratios for the Clinician's Diagnosis Compared With Laboratory Diagnosis

Clinical Diagnosis	LR+ (95% CI)	LR- (95% CI)
Trichomonas (wet preparation)	21 (13-37)	0.39 (0.29-0.53)
Candidiasis (KOH preparation)	7.4 (3.8-15)	0.80 (0.74-0.87)
Bacterial vaginosis (Amsel criteria)	4.0 (3.2-4.9)	0.11 (0.07-0.17)

Abbreviations: CI, confidence interval; KOH, potassium hydroxide; LR+, positive likelihood ratio; LR-, negative likelihood ratio.

LIMITATIONS These results were those of experienced midlevel practitioners who were specifically trained to do the clinical and microscopic examination. Not only were they trained but also they demonstrated competency in the performance of the bedside tests. Generalist physicians would have to ensure their competency in microscopic examinations of vaginal secretions to replicate the results. However, the authors provide accuracy data for these 2 microscopic studies compared with cultures.

A patient's symptom of an abnormal vaginal odor makes bacterial vaginosis more likely, with an LR of 3.2 (95% confidence interval [CI], 2.6-3.9), whereas the absence of the odor makes vaginal candidiasis more likely, with an LR of 2.3 (95% CI, 2.0-2.7). A patient's symptoms of a "change" in her vaginal discharge worked similarly (though not as well) to the presence of an odor: a change in the vaginal discharge made bacterial vaginosis more likely (LR, 2.2; 95% CI, 1.8-2.6), whereas no change in the discharge despite vaginal complaints increased the likelihood of candidiasis (LR, 1.9; 95% CI, 1.6-2.2).

Vaginal pruritus was an inefficient finding for candidiasis.

The symptoms have almost no value for diagnosing trichomoniasis. Although trichomoniasis is the least common of the 3 diagnoses, examination of a microscopic preparation for the organism is necessary. The presence of trichomonads on a microscopic specimen makes the diagnosis of trichomoniasis almost certain. The Amsel criteria¹ for bacterial vaginosis and the presence of yeast on a KOH preparation are also much more useful than the individual clinical findings.

Reviewed by David L. Simel, MD, MHS

REFERENCES FOR THE EVIDENCE

1. Amsel R, Totten PA, Spiegel CA, Chen KC, Eschenbach D, Holmes KK. Nonspecific vaginitis: diagnostic criteria and microbial and epidemiologic associations. *Am J Med.* 1983;74(1):14-22.
2. Nugent RP, Krohn MA, Hillier SL. Reliability of diagnosing bacterial vaginosis is improved by a standardized method of Gram stain interpretation. *J Clin Microbiol.* 1991;29(2):297-301.

CHAPTER 53

Does This Dizzy Patient Have a Serious Form of Vertigo?

David A. Froehling, MD

Marc D. Silverstein, MD

David N. Mohr, MD

Charles W. Beatty, MD

CLINICAL SCENARIOS

Common Causes of Vertigo

CASE 1 A 52-year-old woman was admitted to the hospital because of nausea, a constant spinning sensation, and vomiting of 24 hours' duration. Any movement of her head made these symptoms worse. On examination, she had bilateral horizontal spontaneous nystagmus. Two days later, after symptomatic improvement, she was discharged. At follow-up 2 weeks later, her symptoms and nystagmus had completely resolved.

CASE 2 A 70-year-old woman had a 4-month history of an intermittent whirling sensation when turning her head and especially when rolling over in bed. On examination, a left-side-down head-hanging maneuver elicited rotatory nystagmus, with the fast component to the left ear ([Figure 53-1](#)). There was a latency of about 3 seconds before the onset of nystagmus, which lasted approximately 10 seconds.

WHY EVALUATE VERTIGO?

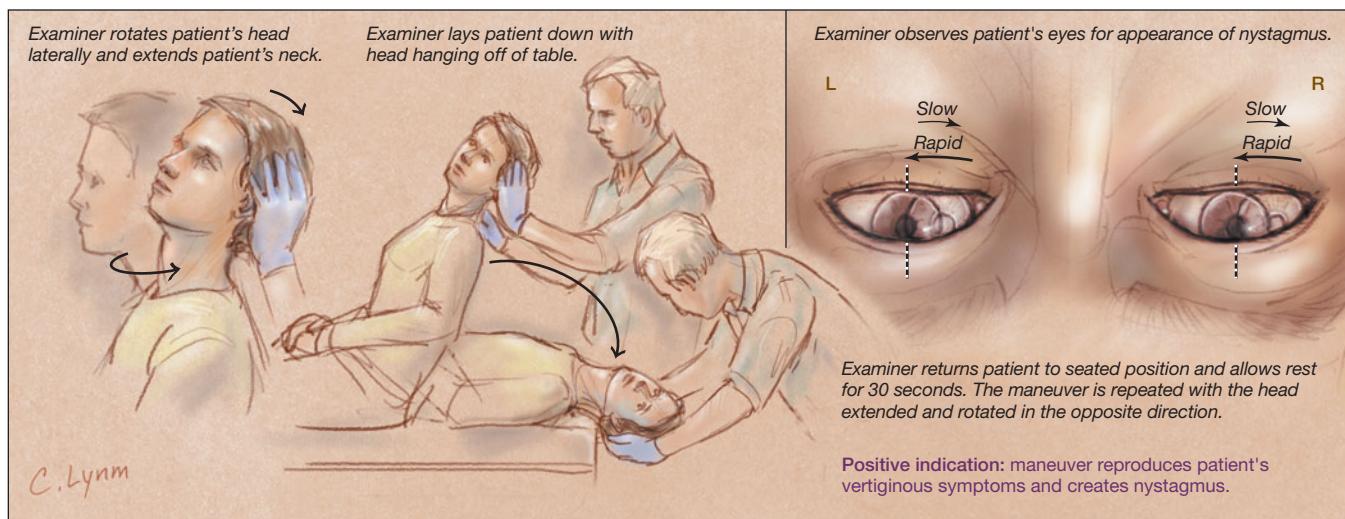
Vertigo is defined in Merriam-Webster's dictionary¹ as a disturbance "in which the external world seems to revolve around the individual or in which the individual seems to revolve in space." Vertigo is an illusion of motion² and is one of several forms of dizziness. The word *dizziness* is derived from the old English word *dysig*, meaning foolish or stupid. The modern usage of the word includes "a whirling sensation in the head with a tendency to fall," "mentally confused or dazed," and "giddiness."¹

In one study³ from a general internal medicine outpatient clinic, dizziness was the third most frequent complaint of patients. In a national survey reported in 1989,⁴ it was the 13th most frequent reason for visits to internists in the United States. Dizziness is often a diagnostic problem in the emergency department.⁵ Among patients treated in an emergency department,⁵ in an outpatient clinic,⁶ and in 2 subspecialty dizziness clinics,^{7,8} vertigo was the most frequent category of dizziness.

Most patients with dizziness can be classified as having one of the following syndromes:

1. impaired perfusion of the central nervous system or near syncope (eg, orthostatic hypotension, cardiac presyncope)
2. dysequilibrium, a sensation of imbalance when standing or walking⁶ (eg, multiple sensory deficits)
3. psychogenic dizziness (eg, major depression, anxiety disorder, and somatization disorder)
4. vertigo (eg, Meniere disease and vestibular neuronitis)⁷

Usually dizziness can be classified according to information obtained from the medical history and physical examination. In this article, we concentrate on the evalua-

**Figure 53-1** How to Test for Positional Nystagmus

The Dix-Hallpike maneuver for positional vertigo is performed by the examiner, who stands at the head of the bed. As the patient is supported and lowered into a position whereby his or her rotated and extended head hangs off the end of the examining table, the examiner observes for nystagmus. In this view, the patient's head has been rotated to the left and expresses nystagmus with a slow response to the right and a rapid response to the left. Repeating the maneuver with the head rotated in the opposite direction would reverse the direction of the nystagmus. A maneuver (with positive indication) will reproduce the patient's symptoms.

tion of vertigo, the most common category of dizziness. Serious forms of vertigo are due to conditions associated with increased mortality or long-term disability. Vertigo severe enough to impair daily functioning and lasting for more than a month would be included as a serious form of vertigo.

The importance of recognizing a patient's complaint of dizziness as vertigo is that it narrows the list of possible causes. Customarily, the causes of vertigo are divided into central

causes (lesions of the central nervous system) and peripheral causes (lesions of the vestibular labyrinth or nerve or both) (Table 53-1). Because of the importance of detecting lesions or diagnosing syndromes that can be treated and because of the need to determine prognosis, physicians should attempt to make a specific diagnosis for patients with vertigo.

Most cases of vertigo are due to lesions of the vestibular nerve or labyrinth.⁵⁻⁸ In 2 dizziness clinics, the most common cause of vertigo was benign paroxysmal positional vertigo.^{7,8}

Table 53-1 Common Causes of Vertigo

Peripheral
Benign paroxysmal positional vertigo
Vestibular neuritis
Recurrent vestibulopathy
Classic Meniere disease
Head trauma (labyrinthine concussion)
Acoustic neuroma
Otosclerosis
Herpes zoster oticus
Cholesteatoma
Perilymph fistula
Aminoglycoside ototoxicity
Central
Vertebrobasilar transient ischemic attacks
Cerebellar or brainstem stroke
Brain tumors
Multiple sclerosis
Vertebobasilar migraine

PATHOPHYSIOLOGY OF VERTIGO AND NYSTAGMUS

Origins of Vertigo

The maintenance of the sense of balance and spatial orientation depends on input from the vestibular labyrinth, visual system, and proprioceptive nerves arising from tendons, muscles, and joints.⁹ The vestibular nuclei, which are in the medulla and lower pons, receive input from the vestibular labyrinth via the vestibular branch of cranial nerve VIII and from the cerebellum.¹⁰ The vestibular nuclei, in turn, send efferent fibers to the cerebellum, the medial longitudinal fasciculus, and the vestibulospinal tract. Visceral manifestations of vertigo (such as nausea and vomiting) are caused by altered input to the dorsal nucleus of the vagus nerve from the vestibular nuclei. Conscious awareness of vertigo resides in the superior temporal gyrus of the cerebral cortex⁹ and involves a mismatch between input to the cerebral cortex from the visual, proprioceptive, and vestibular systems.¹¹ Lesions in various locations, including the inner ear, brain stem, and cerebellum, may all be manifested as vertigo.

Origins of Nystagmus

Nystagmus is the objective accompaniment of vertigo and is defined best as a “rhythmical oscillation of the eyes, with a fast movement in one direction and a slow movement in the other.”¹² The fast component may be horizontal, vertical, rotatory, or any combination of these.¹³

There are 2 clinically relevant kinds of nystagmus in evaluating vertigo: Spontaneous nystagmus is elicited by having the patient look straight ahead, up, down, to the right, and to the left. This type of nystagmus is not influenced by head position.¹⁴ It is normal to have a few beats of nystagmus with extreme lateral gaze.¹³ Positional nystagmus is elicited by a head-hanging maneuver (Figure 53-1).¹³

Altered input passing from the vestibular nuclei to the nuclei of the extraocular muscles through the medial longitudinal fasciculus and related pathways in the reticular formation produces nystagmus. This input may be modified by information arising from the cerebral cortex and the cerebellum.¹³ For example, the fast component of spontaneous nystagmus depends on interaction between the vestibular system and the cerebral cortex.¹⁵

HOW TO ELICIT THE SYMPTOMS AND SIGNS OF VERTIGO

First, Distinguish Vertigo From Other Causes of Dizziness

Patients often have difficulty describing symptoms of dizziness, and even those who have disorders that produce vertigo may not clearly describe a hallucination of movement. As Olsson and Atkins¹⁶ pointed out, “A person is so rarely conscious of his own vestibular system, he has a great deal of trouble describing his symptoms to a doctor.” Thus, clues must be gathered from the medical history and physical examination to classify the dizziness properly.

Dizziness when standing may be due to vertigo, decreased cerebral perfusion,¹⁷ or dysequilibrium.⁶ If the patient reports having symptoms of dizziness primarily while standing, the blood pressure should be checked with the patient in the supine position and also after standing for 5 minutes. If there is an orthostatic decrease in blood pressure, the symptom is likely due to impaired central nervous system perfusion.

Unsteadiness while walking, especially in elderly patients, is often due to dysequilibrium (a feeling of imbalance). The cause is usually multifactorial. On examination, the findings of decreased visual acuity and signs of peripheral neuropathy or abnormal vestibular function support a diagnosis of dysequilibrium.^{6,7}

Dizziness when turning, and especially when rolling over in bed, is usually due to vertigo.

Psychogenic dizziness is a diagnosis of exclusion that should be considered especially in patients with psychiatric illnesses, such as major depression, anxiety disorder, and a somatization disorder. In this setting, the patient should be asked to hyperventilate for 2 minutes and then asked whether the feeling associated with hyperventilation is exactly the

same as the dizzy symptom. The physician should initially hyperventilate along with the patient; this approach encourages the patient and demonstrates the desired rate and depth of breathing for the test.¹⁸ If hyperventilation reproduces the symptom, the dizziness is often psychogenic. However, the usefulness of hyperventilation in diagnosing psychogenic dizziness is unclear. In a study by Kroenke et al⁶ of 100 ambulatory patients with a chief complaint of dizziness, symptoms of dizziness were reproduced by hyperventilation in 21; however, only 1 of these patients had hyperventilation as the primary cause of dizziness. Most of them had dizziness inducible by other maneuvers in addition to hyperventilation. Further studies of the hyperventilation maneuver in the evaluation of patients with suspected psychogenic dizziness are needed. In this study of 100 patients, only 16% had pure psychogenic dizziness, but 24% had other causes of dizziness exacerbated by psychiatric illness.⁶

Second, Take a Proper Medical History From Patients With Vertigo

After it is clear that the patient is describing vertigo, further questions help elicit clues about its specific cause.

Ask When the Dizziness Occurs

It is probably more important to ask a patient about the circumstances in which the dizziness occurs than to ask for a description of the dizziness. Dizziness related to early-morning activities is somewhat helpful in distinguishing between peripheral and central vertigo. Matutinal vertigo (vertigo on first arising in the morning) is usually due to a peripheral vestibular disorder.¹⁹

Ask About Other Otologic Symptoms

Associated otologic symptoms can be helpful in identifying a peripheral cause of vertigo. Hearing loss and vertigo are common in patients with otosclerosis.²⁰ Episodes of hearing loss with vertigo, tinnitus, and a sensation of fullness in the ear occur in patients with Meniere disease.²¹ Patients with acoustic neuromas usually present with hearing loss rather than vertigo. Most of these patients notice dizziness but complain of unsteadiness rather than vertigo.²²

Ask About Other Neurologic Symptoms

Symptoms of neurologic disease, such as weakness, difficulty with speech, or diplopia, in addition to vertigo suggest a central cause.

Ask About Symptom Patterns

Patients with vestibular neuronitis (also called labyrinthitis), benign paroxysmal positional vertigo, and recurrent vestibulopathy (also called benign recurrent vertigo and vestibular Meniere disease) have normal hearing.²³⁻²⁶ Patients with benign paroxysmal positional vertigo²³ (also called benign paroxysmal positional nystagmus²⁷ and cupulolithiasis²⁸) have intermittent episodes of vertigo with head turning.^{23,29} Vestibular neuronitis is characterized by a relatively sudden onset of severe, constant vertigo (made worse by head movement) that resolves after days or weeks.^{23,30} Patients with recurrent vestibulopathy have intermittent episodes of constant vertigo lasting for minutes or hours.^{24,25} Vertigo (with or

Table 53-2 Accuracy of Signs and Symptoms for Diagnosing Peripheral Vertigo in an Emergency Department^a

No. of Patients With Peripheral Vertigo (Not an Emergency)	No. of Patients With Other Causes of Dizziness That Might Be an Emergency	Total	Predictive Value, %	Likelihood Ratio
Positive cluster of signs and symptoms ^b	23	4	27	Positive 85 (23/27) 7.6
Lack of one or more elements in cluster	31	67	98	Negative 68 (67/98) 0.6
Total	54	71	125	... ^c ...

^aData from Herr et al.⁵

^bPositive cluster includes positive results on head-hanging maneuver plus either vertigo or vomiting.

^cEllipses indicate not applicable.

without hearing loss) in a patient who has recently received aminoglycoside antibiotics may be due to the toxic effect these agents have on the vestibular labyrinth.³¹

How to Examine Patients With Vertigo

Findings on physical examination can help physicians detect abnormalities that can be used to determine the cause of vertigo.

Perform a Brief Neurologic Examination

Look for cranial nerve palsies, weakness, reflex changes, ataxia, decreased sensation in the feet, and abnormalities of gait and station. Vertical nystagmus is associated with lesions of the vestibular nuclei or of the cerebellar vermis.¹³ Neurologic findings other than pathologic nystagmus suggest that the lesion is central.

Examine the Ears

Hearing should be checked.³² Cholesteatoma, a complication of chronic otitis media that can present with hearing loss, drainage from the ear, and vertigo, may be found³³; the usual treatment for this is surgery. Alternatively, vesicles associated with herpes zoster oticus (also called Ramsay Hunt syndrome) may be present; patients with this condition often have facial palsy and deafness, together with vertigo.³⁴

Check for Spontaneous Nystagmus

Patients with vestibular neuronitis usually have spontaneous horizontal nystagmus or a mixture of spontaneous horizontal nystagmus and rotatory nystagmus.³⁰ Patients with disorders of the central nervous system may also have spontaneous nystagmus.³⁵ In most of the patients examined by Silvoniemi,³⁰ Lachman and Stahle,³⁶ and Aantaa and Virolainen,³⁷ nystagmus was readily apparent, but in some, detection required Frenzel glasses or electronystagmographic monitoring with the patients' eyes closed. Patients with vestibular neuronitis may also have positional nystagmus.³⁰ Patient 1 in the clinical scenarios had vestibular neuronitis.

Perform a Head-Hanging Maneuver

Most physicians test for positional nystagmus with a method first outlined by Dix and Hallpike²³ and more recently by Mohr.²⁹ The head-hanging maneuver begins with the patient in a sitting position, with gaze fixed on the examiner's forehead (Figure 53-1). The examiner firmly grasps the patient's head and has the patient quickly lie supine, with the head turned about 30 degrees to one side and about 30 degrees below the level of the examining table. Next, the patient sits

up, and the maneuver is repeated with the head turned to the opposite side. In 1979, Baloh et al³⁸ observed that if the maneuver was performed slowly (during a period of 20 seconds), nystagmus was not induced; thus, they recommended performing the position change in about 2 seconds. After each head-hanging maneuver, the physician should observe the patient's eyes for 5 to 15 seconds to determine whether nystagmus has been induced.²⁹ Overall, it takes about 3 to 5 minutes to explain the head-hanging maneuver to the patient, to perform the position changes, and to observe for nystagmus.

Benign paroxysmal positional vertigo is the most common cause of vertigo^{7,8} and can usually be suspected on the basis of the medical history alone. Features of this syndrome include vertigo that occurs only with positional changes and an associated positional nystagmus that is usually rotatory, with a vertical or horizontal component. Also, the nystagmus usually begins 5 to 15 seconds after the head-hanging maneuver, lasts 2 to 30 seconds, and, if the patient is repeatedly returned to the provocative position, occurs less and less until it cannot be induced.^{23,29} Positional nystagmus cannot always be elicited in a patient with a history otherwise compatible with the diagnosis of benign paroxysmal positional vertigo.³⁹⁻⁴¹ Its occurrence during a head-hanging maneuver occasionally makes a vague description of dizziness clearer. Rarely, patients with central nervous system lesions may present with positional vertigo and nystagmus and with no other neurologic abnormality.⁴² Patient 2 in the clinical scenarios had benign paroxysmal positional vertigo.

Learning how to check for positional nystagmus usually requires practice. Always explain to the patient what you are going to do before performing a head-hanging maneuver. Specifically, ask the patient to keep the eyes open if he or she becomes vertiginous; many patients close their eyes if vertigo develops. The head-hanging maneuver should be performed quickly but not so rapidly as to injure the patient. Be observant because the nystagmus may last only a few seconds.

Accuracy of the Symptoms and Signs of Vertigo

Data are available on 3 clinically relevant questions about the accuracy of the clinical examination in patients with vertigo.

1. Can positional nystagmus identify patients with benign paroxysmal positional vertigo? The answer is, not very well. Only 198 of 255 patients with positional vertigo

Table 53-3 Accuracy of Signs and Symptoms for Detecting Serious Causes of Dizziness in an Emergency Department^a

	No. of Patients With Serious Causes of Dizziness ^b	No. of Patients With Nonserious Causes of Dizziness	Total	Predictive Value, %	Likelihood Ratio
Absence of vertigo, age >69 y, or neurologic deficit	33	50	83	Positive 40 (33/83)	1.5
Presence of vertigo, age ≤69 y, and no neurologic deficit	5	37	42	Negative 88 (37/42)	0.3
Total	38	87	125	... ^c	...

^aData from Herr et al.⁵

^bSerious causes of dizziness include medication adverse effects, seizures, stroke, and cardiac arrhythmia.

^cEllipses indicate not applicable.

examined in a dizziness clinic had positional nystagmus during initial and subsequent examinations (sensitivity, 78%).³⁹ In an epidemiologic study of positional vertigo, only 13 of 26 patients tested had positional nystagmus (sensitivity, 50%).⁴¹

- Can matutinal vertigo distinguish peripheral causes from central causes of vertigo? Again, the answer is, not very well. In a study of 100 neurology patients (48 of whom had matutinal vertigo), matutinal vertigo had a sensitivity of 51% and a specificity of 69% for peripheral disorders,⁴³ and in an epidemiologic study, symptoms of vertigo when rolling over in bed generated a sensitivity of 40% for benign paroxysmal positional vertigo.⁴¹
- Can any set of symptoms and signs distinguish urgent causes from nonurgent causes of dizziness? Symptoms and signs can help identify patients in need of an urgent evaluation, as shown in Tables 53-2 and 53-3, which are from a study of 125 emergency department patients with the complaint of dizziness.⁵ Patients who had the highly specific cluster of positive results on the head-hanging test and either vertigo or vomiting almost always had a nonurgent peripheral vertigo (a finding with high specificity, if positive, tends to rule in the target disorder). In Table 53-3, the high sensitivity (87%) of the absence of vertigo or age older than 69 years or the presence of a neurologic deficit for a serious cause of dizziness meant that younger patients with vertigo but no neurologic deficit were unlikely to have an urgent cause of dizziness (a finding with high sensitivity, if negative, tends to rule out the target disorder).

These reassuring results of the accuracy of the clinical examination come from a single study in an emergency department with rates of peripheral vertigo and serious disease characteristic of such settings; they need independent confirmation in different settings. Although the nonurgent causes of dizziness may not require immediate hospitalization, some of the causes of peripheral vertigo (eg, acoustic neuroma) deserve further diagnostic study.

THE BOTTOM LINE

The following are our recommendations on useful symptoms and signs in the evaluation of patients with dizziness:

- In patients with suspected vertigo, ask whether they have dizziness when changing body position (rolling over in bed, looking up at the ceiling, or bending over to tie shoelaces) and perform a head-hanging maneuver to check for positional nystagmus.
- In combination with other data (including a brief neurologic examination) in an emergency department setting, the presence of positional nystagmus can be useful when evaluating for serious causes of dizziness.

Author Affiliations at the Time of the Original Publication

Division of Area General Internal Medicine (Drs Froehling, Silverstein, and Mohr), Department of Health Sciences Research (Dr Silverstein), and Department of Otorhinolaryngology (Dr Beatty), Mayo Clinic and Mayo Foundation, Rochester, Minnesota.

REFERENCES

- Webster's Third New International Dictionary of the English Language, Unabridged. Springfield, MA: Merriam-Webster Inc; 1986:664.
- Adams RD, Victor M. Deafness, dizziness, and disorders of equilibrium. In: *Principles of Neurology*. 3rd ed. New York, NY: McGraw-Hill Book Co; 1985:216-218.
- Kroenke K, Mangelsdorff AD. Common symptoms in ambulatory care: incidence, evaluation, therapy, and outcome. *Am J Med*. 1989;86(3):262-266.
- Woodwell DA. Office visits to internists, 1989. *Adv Data*. 1992;(209):1-10.
- Herr RD, Zun L, Mathews JJ. A directed approach to the dizzy patient. *Ann Emerg Med*. 1989;18(6):664-672.
- Kroenke K, Lucas CA, Rosenberg ML, et al. Causes of persistent dizziness: a prospective study of 100 patients in ambulatory care. *Ann Intern Med*. 1992;117(11):898-904.
- Drachman DA, Hart CW. An approach to the dizzy patient. *Neurology*. 1972;22(4):323-334.
- Nedzelski JM, Barber HO, McIlroy L. Diagnoses in a dizziness unit. *J Otolaryngol*. 1986;15(2):101-104.
- Frederick MW. Central vertigo. *Otolaryngol Clin North Am*. 1973;6(1):267-285.
- Kelly JP. Vestibular system. In: Kandel ER, Schwartz JH, eds. *Principles of Neural Science*. 2nd ed. New York, NY: Elsevier Science Publishing Co; 1985:591-595.
- Lehrer JF, Poole DC. Diagnosis and management of vertigo. *Compr Ther*. 1987;13(9):31-40.
- Rowland LP. Clinical syndromes of the brain stem. In: Kandel ER, Schwartz JH, eds. *Principles of Neural Science*. 2nd ed. New York, NY: Elsevier Science Publishing Co; 1985:599.
- Mayo Clinic Department of Neurology. *Clinical Examinations in Neurology*. 5th ed. Philadelphia, PA: WB Saunders Co; 1981:63-95.

14. Mylén CO. Positional nystagmus: a review and future prospects. *J Laryngol Otol.* 1950;64(6):295-318.
15. Plum F, Posner JB. *The Diagnosis of Stupor and Coma.* 3rd ed. Philadelphia, PA: FA Davis Co; 1980:58.
16. Olsson JE, Atkins JS. Vestibular disorders. *Otolaryngol Clin North Am.* 1987;20(1):83-111.
17. Barber HO. Positional vertigo and nystagmus. *Otolaryngol Clin North Am.* 1973;6(1):169-187.
18. Magarian GJ. Hyperventilation syndromes: infrequently recognized common expressions of anxiety and stress. *Medicine (Baltimore).* 1982;61(4):219-236.
19. Fisher CM. Vertigo in cerebrovascular disease. *Arch Otolaryngol.* 1967;85(5):529-534.
20. Thomas JE, Cody DTR. Neurologic perspectives of otosclerosis. *Mayo Clin Proc.* 1981;56(1):17-21.
21. Pulec JL. Meniere's disease: etiology, natural history, and results of treatment. *Otolaryngol Clin North Am.* 1973;6(1):25-39.
22. Harner SG, Laws ER Jr. Diagnosis of acoustic neurinoma. *Neurosurgery.* 1981;9(4):373-379.
23. Dix MR, Hallpike CS. The pathology, symptomatology, and diagnosis of certain common disorders of the vestibular system. *Proc R Soc Med.* 1952;45(6):341-354.
24. Leliever WC, Barber HO. Recurrent vestibulopathy. *Laryngoscope.* 1981;91(1):1-6.
25. Slater R. Benign recurrent vertigo. *J Neurol Neurosurg Psychiatry.* 1979;42(4):363-367.
26. Alford BR. Meniere's disease: criteria for diagnosis and evaluation of therapy for reporting (report of Subcommittee on Equilibrium and Its Measurement). *Trans Am Acad Ophthalmol Otolaryngol.* 1972;76(6):1462-1464.
27. Harbert F. Benign paroxysmal positional nystagmus. *Arch Ophthalmol.* 1970;84(3):298-302.
28. Schuknecht HF. Cupulolithiasis. *Arch Otolaryngol.* 1969;90(6):765-778.
29. Mohr DN. The syndrome of paroxysmal positional vertigo: a review. *West J Med.* 1986;145(5):645-650.
30. Silvoniemi P. Vestibular neuronitis: an otoneurological evaluation. *Acta Otolaryngol Suppl (Stockh).* 1988;453:1-72.
31. Jackson GG, Arcieri G. Ototoxicity of gentamicin in man: a survey and controlled analysis of clinical experience in the United States. *J Infect Dis.* 1971;124(suppl):S130-S137.
32. Bagai A, Thavendiranathan P, Detsky AS. Does this patient have hearing impairment? *JAMA.* 2006;295(4):416-428.
33. Vernick DM, Branch WT Jr. The painful or discharging ear. In: Branch WT Jr, ed. *Office Practice of Medicine.* 2nd ed. Philadelphia, PA: WB Saunders Co; 1987:291-293.
34. Adams RD, Victor M. Viral infections of the nervous system. In: *Principles of Neurology.* 3rd ed. New York, NY: McGraw-Hill Book Co; 1985:552-553.
35. Nylen CO. The oto-neurological diagnosis of tumours of the brain. *Acta Otolaryngol.* 1939;33(suppl):5-151.
36. Lachman J, Stahle J. Vestibular neuritis: a clinical and electronystagmographic study. *Neurology.* 1967;17(4):376-380.
37. Aantaa E, Virolainen E. Vestibular neuronitis: a follow-up study. *Acta Otorhinolaryngol Belg.* 1979;33(3):401-404.
38. Baloh RW, Sakala SM, Honrubia V. Benign paroxysmal positional nystagmus. *Am J Otolaryngol.* 1979;1(1):1-6.
39. Katsarkas A, Kirkham TH. Paroxysmal positional vertigo: a study of 255 cases. *J Otolaryngol.* 1978;7(4):320-330.
40. Baloh RW, Honrubia V, Jacobson K. Benign positional vertigo: clinical and oculographic features in 240 cases. *Neurology.* 1987;37(3):371-378.
41. Froehling DA, Silverstein MD, Mohr DN, Beatty CW, Offord KP, Ballard DJ. Benign positional vertigo: incidence and prognosis in a population-based study in Olmsted County, Minnesota. *Mayo Clin Proc.* 1991;66(6):596-601.
42. Watson P, Barber HO, Deck J, Terbrugge K. Positional vertigo and nystagmus of central origin. *Can J Neurol Sci.* 1981;8(2):133-137.
43. Berkowitz BW. Matutinal vertigo: clinical characteristics and possible management. *Arch Neurol.* 1985;42(9):874-877.

Prepared by David L. Simel, MD, MHS

Reviewed by David A. Froehling, MD,
and Richard Bedlack, MD, PhD**CLINICAL SCENARIO**

A 58-year-old healthy man presents with dizziness. One week ago, he had an upper respiratory illness consisting of a slight fever, cough, and rhinorrhea. During the previous 2 days, he has had 3 episodes of extreme unbalance lasting less than 3 to 4 minutes, when he felt as if he were “drunk.” During these episodes, he felt nauseated, which caused him to lie down and close his eyes until the symptoms resolved. He has had no hearing loss. Your neurologic examination reveals no focal findings in the cranial or peripheral nerves.

Original Review

Froehling DA, Silverstein MD, Mohr DN, Beatty CW. Does this dizzy patient have a serious form of vertigo? *JAMA*. 1994;271(5):385-388.

UPDATED LITERATURE SEARCH

The focus of the original Rational Clinical Examination article and this update is on the vestibular disorders characterized by true vertigo. True vertigo creates a sensation of rotation. Although the initial publication approached vertigo from a general perspective, we sought to find updated information on the diagnosis of benign positional vertigo, the most common cause of vertiginous symptoms. We used the search terms “vertigo/di,” “exp dizziness,” and the text words “\$Hallpike,” “Eply,” or “benign positional vertigo” to identify English-language articles on vertigo in adults, published between 1993 and November 2004. After excluding case reports, letters, and general reviews, we were left with 154 articles. These were searched to identify studies using prospective data collection and that reported the sensitivity, specificity, or predictive values of clinical findings in patients who presented to their physician with complaints of dizziness. A systematic review¹ evaluated the distribution of diagnoses among patients with dizziness. A second general systematic review² without any quantitative formal research question provides a useful reference list for clinical descriptions of the common causes of vertigo. We found 1 additional article that prospectively evaluated patients in a

clinical population, using a patient questionnaire for diagnosing vertigo.

NEW FINDINGS

- The response to the Dix-Hallpike maneuver serves as a reasonable reference standard for benign positional vertigo because it identifies patients who will respond to canalith repositioning maneuvers.
- Hearing loss, part of the examination of the dizzy patient, has been reviewed in The Rational Clinical Examination series and can be assessed with the whispered voice test.³

Details of the Update

Patients with dizziness may have a variety of disorders so that diagnosing benign positional vertigo requires an understanding of its overall incidence in relation to other etiologies. Peripheral vestibular disorders are the most common causes for dizziness (about 40% of patients with dizziness), of which benign positional vertigo and vestibular neuritis are the most frequent diagnoses. Retrospective studies tend to find a higher incidence of benign positional vertigo than those that enroll dizzy patients prospectively.

Clinicians (and patients) may be overly concerned with brain tumors when there is a new symptom of vertigo, but the likelihood that a dizzy patient without hearing loss will have a cerebellopontine angle mass responsible for the symptoms is low (probability, 1×10^{-4}).⁴ Among patients with dizziness associated with asymmetric hearing loss, a clinician would need to perform 638 scans to detect 1 cerebellopontine angle mass (compared with 9307 scans for dizzy patients without hearing loss). Thus, the approach to clinical diagnosis should more appropriately focus on attempts to rule in less serious causes of vertigo (eg, benign positional vertigo), rather than an initial effort to rule out serious causes such as tumors.

We found a systematic review¹ that identified 2 retrospective studies suggesting that the clinical history alone allows proper diagnosis of 69% to 76% of dizzy patients. We also found a prospective study in a small group of patients referred to an otolaryngologist where patient history was collected through a questionnaire.⁵ The questionnaire directs the clinician to the more common causes of vertigo and would have allowed correct

Box 53-1 Establish the Initial Diagnosis After Understanding the Patient's History

Patient Symptoms	Initial Diagnosis
No hearing loss + episodic vertigo	Benign positional vertigo
No hearing loss + persistent vertigo	Vestibular neuronitis
Hearing loss + episodic vertigo	Meniere disease
Hearing loss + persistent vertigo	Labyrinthitis

categorization of 61% of the patients with true vertigo according to whether they had episodic (<5 minutes, 5 minutes to 24 hours, 1 day to 1 week) vs persistent vertigo (>1 week) and hearing loss or no hearing loss. See [Box 53-1](#).

The questionnaire requires validation in a much larger population of patients and in different clinical settings (emergency departments and primary care clinics) because the patient may not belong clearly in one category, requiring clinical judgment. However, the questions do provide a reasonable paradigm for the initial line of questioning for the vertiginous patient.

Once the medical history is obtained, perhaps narrowing the diagnosis to the most likely causes, specialists use a variety of clinical maneuvers. The maneuvers assess the vestibuloocular reflex through the nystagmus response to a head thrust, through fixation suppression, after a headshake, through caloric testing, or through visual acuity during head shaking.⁶ Unfortunately, the maneuvers have not been assessed in primary care clinics or emergency departments to evaluate whether they add information to the Dix-Hallpike during a patient's initial presentation for care and before referral.

IMPROVEMENTS IN THE DATA PRESENTED IN THE ORIGINAL PUBLICATION

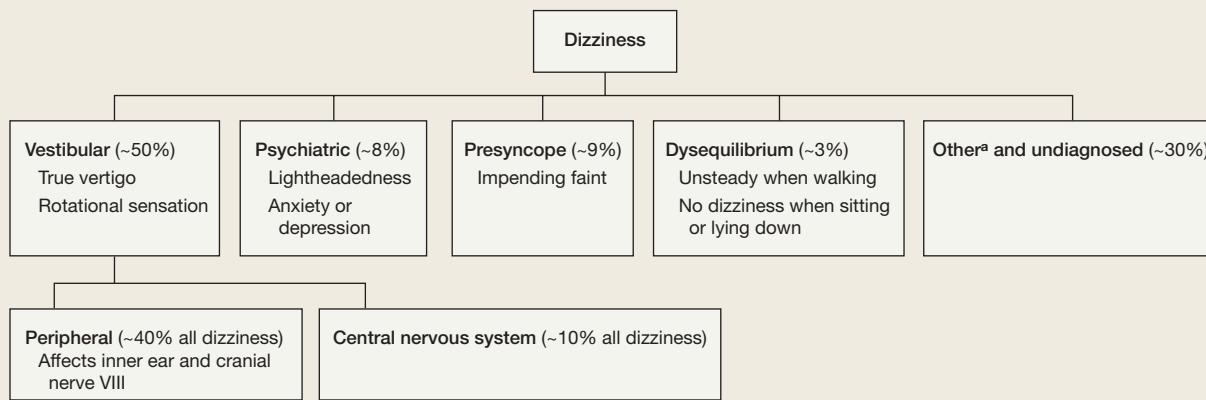
A systematic review provides a useful taxonomy for patients with disorders creating dizziness, improving the information provided in Table 53-1 of the original article ([Figure 53-2](#)).¹ The vestibular disorders are further sorted by those that represent peripheral vestibular problems (“less serious” in terms of the underlying etiology, though often creating a significant problem with activities of daily living) vs central vestibular disorders ([Figure 53-3](#)).

CHANGES IN THE REFERENCE STANDARD

The diagnosis of vestibular disorders relies on the direct observation of eye movements during positional testing in a patient with no focal neurologic findings or central nervous system disease. The clinical definition of benign positional vertigo that requires a positive Dix-Hallpike maneuver result is supported by a meta-analysis of randomized trials of canalith repositioning procedures.⁷ The randomized trials demonstrated that, within 1 month of treatment, patients with a positive Dix-Hallpike maneuver result benefit from the repositioning procedures with symptom resolution (number needed to treat = 3). Furthermore, the positive Dix-Hallpike maneuver result returns to normal at a rate similar to that of the symptom improvement.

RESULTS OF LITERATURE REVIEW

The Dix-Hallpike maneuver can be done in most patients, but some cannot tolerate it. A small study of patients with benign positional vertigo showed that the maneuver could be performed with a different motion by having the patient lie down on his or her side.⁸ The examiner supports the head while the patient looks to the left at a 45-degree angle and rapidly lies down on the right side. The maneuver is repeated with the patient looking to the right and rapidly going from the sitting position to lying on the left side. The patient should cross the

**Figure 53-2 Dizziness Taxonomy**

^a“Other” includes drug toxicity, substance abuse, and a variety of medical illnesses.

arms to prevent inadvertently stopping the motion as the physician helps with the maneuver. The agreement with the Dix-Hallpike maneuver is moderate ($\kappa = 0.60$; 95% confidence interval, 0.32–0.89). However, patients with back or neck problems may not be able to perform the side-lying maneuver any easier than the Dix-Hallpike maneuver.⁹ A partial list of the absolute contraindications to either maneuver includes a history of neck surgery, severe rheumatoid arthritis, cervical myelopathy, cervical radiculopathy, carotid syncope, neck trauma, or vascular diseases of the neck.

EVIDENCE FROM GUIDELINES

No federal guidelines address the systematic evaluation of dizzy patients.

CLINICAL SCENARIO—RESOLUTION

The patient's clinical history is informative. He almost certainly has benign positional vertigo or vestibular neuritis related to his previous viral infection. A Dix-Hallpike maneuver result would likely be positive. No additional laboratory studies or radiologic imaging is necessary with this initial presentation of true vertigo.

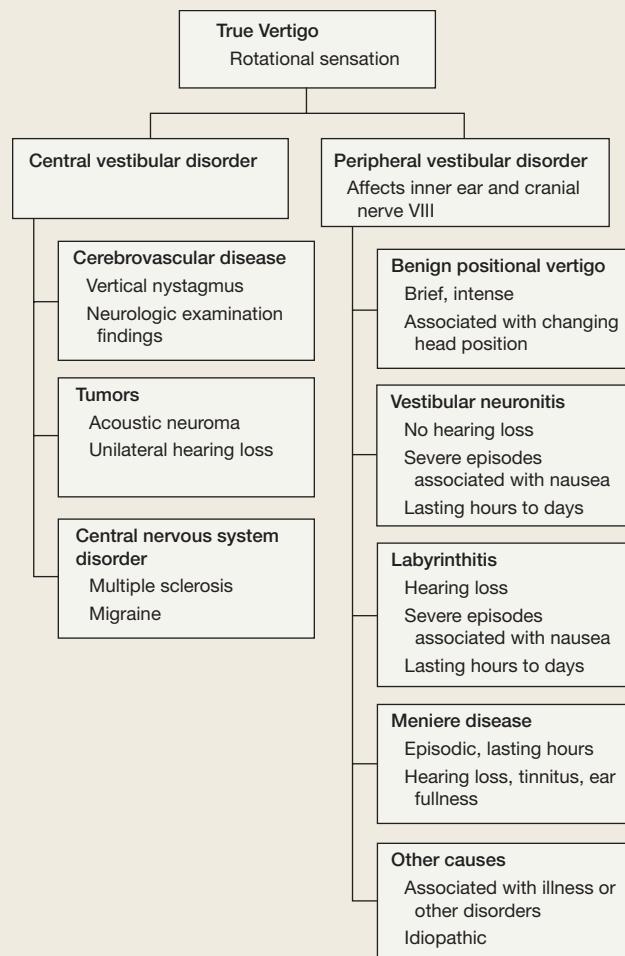


Figure 53-3 Vestibular Disorders

VERTIGO—MAKE THE DIAGNOSIS

PRIOR PROBABILITY

Once the medical history confirms vertigo in a patient with dizziness, most affected patients will have a peripheral vestibular disorder (40%). The prior probability of benign positional vertigo among dizzy patients is 10%.

POPULATION FOR WHOM VERTIGO SHOULD BE CONSIDERED

- Benign positional vertigo should be considered only in patients who volunteer that they have dizziness symptoms.

DETECTING THE LIKELIHOOD OF VERTIGO

The medical history identifies the patient with true vertigo, whereas the clinical examination results identify patients with benign positional vertigo. The responses to the maneuvers are not screening tests with an associated sensitivity and specificity because they define the diagnosis of benign positional vertigo.

REFERENCE STANDARD TESTS

The diagnosis requires direct observation of eye movements during positional testing in a patient with no focal neurologic findings or central nervous system disease. Prospective clinical studies might put more weight on the observations by a specialist, but no comparison studies between generalist physicians and specialist physicians have evaluated the accuracy of generalist clinicians.

REFERENCES FOR THE UPDATE

1. Hoffman RM, Einstdter D, Kroenke K. Evaluating dizziness. *Am J Med*. 1999;107(5):468-478.^a
2. Hanley K, O'Dowd T, Considine N. A systematic review of vertigo in primary care. *Br J Gen Pract*. 2001;51(469):666-671.
3. Bagai A, Thavendiranathan P, Detsky AS. Does this patient have hearing impairment? *JAMA*. 2006;295(4):416-428.
4. Gizzi M, Riley E, Molinari S. The diagnostic value of imaging the patient with dizziness. *Arch Neurol*. 1996;53(12):1299-1304.
5. Kentala E, Rauch SD. A practical assessment algorithm of diagnosis of dizziness. *Otolaryngol Head Neck Surg*. 2003;128(1):54-59.^a
6. Goebel JA. The ten-minute examination of the dizzy patient. *Semin Neurol*. 2001;21(4):391-398.
7. Woodworth BA, Gillespie MB, Lambert PR. The canalith repositioning procedure for benign positional vertigo: a meta-analysis. *Laryngoscope*. 2004;114(7):1143-1146.
8. Cohen HS. Side-lying as an alternative to the Dix-Hallpike test of the posterior canal. *Otol Neurotol*. 2004;25(2):130-134.
9. Humphriss RL, Baguley DM, Sparkes V, Peerman SE, Moffat DA. Contraindications to the Dix-Hallpike manoeuvre: a multidisciplinary review. *Int J Audiol*. 2003;42(3):166-173.

^aFor the Evidence to Support the Update for this topic,
see <http://www.JAMAEvidence.com>.

EVIDENCE TO SUPPORT THE UPDATE:

53

Vertigo

TITLE Evaluating Dizziness.

AUTHORS Hoffman RM, Einstadter D, Kroenke K.

CITATION *Am J Med.* 1999;107(5):468-478.

QUESTION What are the frequencies of various causes of dizziness?

DESIGN Formal systematic review without meta-analysis.

DATA SOURCE MEDLINE database.

STUDY SELECTION AND ASSESSMENT The authors identified studies of adults with dizziness, published in English between 1966 and 1996, indexed with the following search terms: “dizziness” and “vertigo” with “vestibular function tests,” “electronystagmography,” “calorics,” “nystagmus,” “Barany,” “Hallpike,” “caloric testing,” and “brainstem auditory evoked responses.” An initial 1755 references were identified and then filtered down to 229 references that met the initial criteria; an additional 44 articles were retrieved from the reference lists. The review was based on 12 etiology studies, 16 prognosis studies, and 38 studies of diagnostic tests. The studies of etiology used a variety of diagnostic tests. Each article was reviewed by 2 investigators; disagreements were resolved by a third person.

MAIN RESULTS

The clinical setting, study design, sample size, age and sex of patients, symptom duration, and diagnostic tests used were reported for the 12 etiology studies. Quality scores were not reported. The authors provide a framework for the taxonomy of the dizzy patient (see Figures 53-2 and 53-3).

The authors report that the medical history and physical examination led to a probable diagnosis for dizziness in about 75% of patients, but the details of this assessment are not provided. According to 2 retrospective studies, the investigators found that the diagnoses could be based on the history alone in 69% to 76% of patients. Among all patients with dizziness, the Dix-Hallpike maneuver (suggesting

Table 53-5 Frequency of Various Causes of Dizziness

Disorder (n = 7 Studies)	Prevalence (95% CI)
Peripheral vestibular disorder	0.40 (0.27-0.54)
Central vestibular disorder	0.09 (0.06-0.13)
Presyncope	0.09 (0.06-0.13)
Psychiatric	0.08 (0.05-0.12)
Dysequilibrium	0.03 (0.001-0.10)

Abbreviation: CI, confidence interval.

benign peripheral vertigo) was present in 16% (median), though the range was 7% to 44%.

The authors did not conduct a meta-analysis of any results. However, the sample size and frequency of disorders are presented for each etiology study. The data in Table 53-5 represent the prevalence of each disorder for the studies that were done with prospective data collection. The settings for these prospective data were primary care clinics (n = 2 studies, 240 patients), neurology clinics (n = 2 studies, 217 patients), emergency departments (n = 2 studies, 218 patients), or a dizziness clinic (n = 1 study, 104 patients).

Approximately 10% of all dizzy patients had benign positional vertigo, whereas 11% had vestibular neuronitis. The frequency of other causes of true vertigo, iatrogenic causes, and undiagnosed dizzy patients is high and approximately 25% to 30%.

CONCLUSIONS

LEVEL OF EVIDENCE Systematic review.

STRENGTHS The systematic review included data from primary care clinics, emergency departments, neurology clinics, and specialized dizziness clinics. The sample sizes across these clinics were well balanced, representing a typical spectrum of dizzy patients.

LIMITATIONS No quality scores or formal methodologic assessments were reported, though the study design (retrospective vs prospective) is reported. The review required that studies have a reference standard for diagnostic tests, but the reference standard that was used is not reported. The authors

acknowledge that there is no objective reference standard for most causes of dizziness.

This systematic review provides a useful taxonomy for the dizzy patient. By combining the estimates for the prospective studies only, we find that about 50% of dizzy patients had vestibular disorders. This is compatible with the frequency reported in nonsystematic reviews. Peripheral vestibular disorders include patients with benign positional vertigo, vestibular neuronitis, Meniere disease, and true vertigo of unknown cause. About 10% of dizzy patients will have benign positional vertigo, and a similar number will have vestibular neuronitis.

Reviewed by David L. Simel, MD, MHS

TITLE A Practical Assessment Algorithm for Diagnosis of Dizziness.

AUTHORS Kentala E, Rauch SD.

CITATION *Otolaryngol Head Neck Surg.* 2003;128(1): 54-59.

QUESTION Does a simple questionnaire do as well as a clinician for diagnosing the cause of vertigo?

DESIGN Prospective, nonconsecutive patients.

SETTING Otolaryngology clinic with a specialist in vertigo.

PATIENTS Fifty-seven patients (42 women and 15 men) referred for dizziness.

DESCRIPTION OF TESTS AND DIAGNOSTIC STANDARD

The patient-completed questionnaire followed the paradigm of categorizing dizzy patients presented in the original Rational Clinical Examination article on vertigo.¹ The questionnaire involves first asking about the presence of self-assessed hearing loss and vertigo (defined for the patient as “false sense of motion, floating, bobbing, swaying, rocking, tilting, or spinning”). The patients with true vertigo assessed the duration of episodes as episodic (<5 minutes, 5 minutes to 24 hours, 1 day to 1 week) or persistent vertigo (>1 week). The questionnaire also asked single questions to assess for (1) dysequilibrium (“Do you have a sense of being off balance, tipsy, wobbly, feeling you might fall?”); (2) presyncope (“Do you have a feeling you might faint, black out, or lose consciousness?”); or (3) psychiatric diagnosis (“Do you feel disconnected or distanced from the world around you, feel panicky, or have tingling about the mouth or hands?”).

The otolaryngologist, blinded to the patient’s self-assessed questionnaire results, diagnosed the patient according to the medical history elicited, clinical examination results, and results from audiometric and otoneurologic tests. The specific tests and maneuvers were not reported.

MAIN OUTCOME MEASURES

For patients with true vertigo, the clinician’s diagnosis was compared with the patient’s questionnaire, categorized as shown in Box 53-1.

MAIN RESULTS

A total of 35 of the 57 patients had true vertigo. The questionnaire alone would have allowed correct categorization of 61% of the patients with true vertigo according to whether they had episodic (<5 minutes, 5 minutes to 24 hours, 1 day to 1 week) vs persistent vertigo (>1 week) and hearing loss or no hearing loss.

CONCLUSIONS

LEVEL OF EVIDENCE Level 4.

STRENGTHS Simplified approach to recording the patient medical history.

LIMITATIONS Although the clinician did not have the questionnaire answers, the clinician developed the questionnaire and was thus aware of the study hypotheses. This incorporation bias may have made the questionnaire appear to work better than it would once generalized to other settings. The questionnaire requires evaluation in a primary care and emergency department setting. The details of the clinical examination and other tests are not provided. The sample size is small.

Although the overall quality of the study means that the results cannot be applied with confidence, the questionnaire does provide a reasonable paradigm for the initial line of questioning the vertiginous patient.

Reviewed by David L. Simel, MD, MHS

REFERENCE FOR THE EVIDENCE

- Froehling DA, Silverstein MD, Mohr DN, Beatty CW. Does this patient have a serious form of vertigo? *JAMA.* 1994;271(5):385-388.

INDEX

Page numbers followed by a *t* or *f* indicate locations of Tables or Figures, respectively.

A

AAA. *See* abdominal aortic aneurysm

AAP. *See* American Academy of Pediatrics

ABCD(E) criteria for melanoma, 392, 393*t*

abdominal aortic aneurysm (AAA), 17-22, 25-27

abdominal palpation for, 17, 18

asymptomatic AAA, 19-20, 21*t*

factors, affecting, 20, 21*t*

method, 20, 22

ruptured AAA, 19

evidence from guidelines, 26

findings for, 25-26

univariate, 26

likelihood ratio, 27*t*

literature research, 25

methods, 18-19

original publication data, improvements in, 26

physical diagnosis of, 17-18

importance of, 18

prior probability, 27

pulsatile mass of, 18

reference standard, changes in, 26, 27

abdominal auscultation, for bruits

accuracy of

in renovascular hypertension, 31

areas of, 30

precision of, 31

abdominal bruits, 35-37

anatomic and physiologic origin of, 29

auscultatory characteristics of, 32

evidence from guidelines, 36

examination for, 30-31

findings of, 35

literature research, 35

nonrenovascular causes of, 31*t*

original publication data,

improvements in, 36

presence of, 31-32

prevalence of, 29-30

prognosis of, 32

reference standard, changes in, 36

abdominal palpation, 25. *See also* palpation

for abdominal aortic aneurysm, 17, 18

asymptomatic AAA, 19-20, 21*t*

factors, affecting, 20, 21*t*

method, 20, 22

ruptured abdominal aortic aneurysm, 19

abdominojugular reflux sensitivity, specificity, or likelihood ratio

left ventricular dysfunction, 213*t*

venous waveforms in, 126*t*

central venous pressure assessment, 134*t*

abdominojugular reflux test, 128, 134

abduction, of thumb testing, 113

abduction stress test, 361

abnormal monofilament testing, 112*t*

abnormal vibratory sensation, 112*t*

accuracy

of clinical examination, 1, 9

characteristics, 4-5

confidence interval, 12

“good” symptom or sign, 11-12

likelihood ratio, 9-11

meta-analysis, 12-13

pretest probability, 11

“sensitivity-only” studies, 13

ACE. *See* angiotensin-converting enzyme

acetaminophen, 247, 343, 357

acetylcholine receptor (AChR), 449, 450

acetylcholine receptor antibody-positive myasthenia gravis, 450*f*

test for, 450

Achilles tendon reflex. *See* reflexes

AChR. *See* acetylcholine receptor

ACI. *See* acute cardiac ischemia

ACI-TIPI. *See* Acute Cardiac Ischemia Time-Insensitive Predictive Instrument

ACL. *See* anterior cruciate ligament

acoustic neuromas, 711

acoustic reflectometry, 495

action tremor, 506, 507

active compression test

for labral tears, 586*t*

acute blood loss

physical signs, accuracy of, 319-320

acute cardiac ischemia (ACI), 475

multivariate findings for, 473-474

Acute Cardiac Ischemia Time-

Insensitive Predictive

Instrument (ACI-TIPI),

475, 476

acute chest pain, diagnosis of, 462-463, 463*f*

acute cholecystitis, 137-143, 145-147, 561

definition, 137-138

diagnostic imaging, accuracy of, 138

findings of, 145

guidelines, evidence from, 146

likelihood ratio, 147

literature review, results of, 146

literature search, 145

methods, 138-139

original publication data, improvements in, 146

prior probability, 147

reference standard, changes in, 146, 147

results, 139-142

signs and symptoms, 138

accuracy of, 141

precision of, 140-141

acute otitis media (AOM), in children, 493-499, 501-503

anatomic/physiologic origins, 494

definition of, 494

findings of, 501

guidelines, evidence from, 502

improvement of, 498-499

likelihood ratio of, 503*t*

literature review, results of

multivariate findings for, 502, 502*t*

univariate findings for, 502

literature search, 501

and otitis media with effusion,

distinguishing between, 493

prior probability, 503

reference standard, changes in, 503

search strategy and quality review, 495-496

symptoms and signs

accuracy of, 496-498

- acute otitis media (AOM), in children
(Continued)
 elicitation of, 494-495
 precision of, 496
- acute respiratory illness, 539. *See also* pneumonia
- adenoviruses, 344
- aerobic exercise, 306
- age, 152
 as an indicator of prevalence for bruits, 30f
 perimenopause, 408f
 postmenopause, 408f
- sensitivity, specificity, or likelihood ratio
 in back pain, 76t
 in breast cancer, 101t
 in carpal tunnel syndrome, 115t
 in obstructive airways disease, 161t
 in osteoporosis, 491t
 in perimenopause, 416t
 in temporal arteritis, 654
- Agency for Health Care Policy and Research, 47
- Agency for Healthcare Research and Quality (AHRQ), 494, 498, 502
- agreement
 calculation of, 8f
- agreement beyond chance, 3, 4, 15, 104, 294, 333, 617. *See also* agreement
- AHRQ. *See* Agency for Healthcare Research and Quality
- airflow limitation, 149-156
 clinical examination for, 155
 accuracy of, 155
 measures of, 151-152, 153, 155
 pathophysiologic characteristics of, 150-151
 signs of
 accuracy of, 153-155
 medical history, 151
 physical examination, 151-152
 precision of, 153
 symptoms of
 medical history, 151
 accuracy of, 152-153
 precision of, 152
 physical examination, 151-152
- airflow obstruction, 153
- alcohol abuse, 1-2, 39, 249. *See also* alcoholism
 CAGE questionnaire for, 2-3, 4-5, 7, 8t
 diagnostic standards for, 39-41
 diagnostic tests of, 41-42
 at-risk drinking, problems of
 in pregnant women, 44-45
- AUDIT questionnaire, 41, 42, 43t, 43-44
 biochemical and hematologic tests, 41-42
 CAGE questionnaire, 41, 42, 43t, 44
 MAST questionnaire, 41, 42, 42t, 44
- alcohol dependency. *See* alcohol abuse
- alcohol drinking, problems of, 47-52
 evidence from guidelines, 49
 findings of, 47
 literature research, 47
 original publication data, improvements in, 48
 prior probability, 50
 reference standard, changes in, 48, 50
- alcohol screening
 instruments. *See* AUDIT questionnaire; AUDIT-C questionnaire; CAGE questionnaire; MAST questionnaire; T-ACE questionnaire; TWEAK questionnaire
 web resources for, 49
- Alcohol Use Disorders Identification Test (AUDIT), 41
 accuracy of, 43t, 43-44
 reliability of, 42
 questionnaire, 47, 49, 51t
 sensitivity, specificity, or likelihood ratio
 for alcohol abuse, 50t
- Alcohol Use Disorders Identification Test, Consumption Questions (AUDIT-C), 49, 51t
 sensitivity, specificity, or likelihood ratio
 for alcohol abuse, 50t
- alcohol withdrawal syndromes, 5
- alcoholism, 2-3. *See also* alcohol abuse
 algorithm-driven analyses
 for appendicitis, 54
 alignment, of knee, 361
 Alvarado clinical decision rule (MANTRELS), 61, 62, 62t, 63t.
See also clinical prediction rules and scores
 sensitivity, specificity, or likelihood ratio
 in appendicitis, 63t
- Alzheimer disease, 509
- ambulatory carotid bruit, 104-105
- amenorrhea, 552, 554
- American Academy of Family Practice, 502
- American Academy of Neurology, 634
- American Academy of Pediatrics (AAP), 330, 498, 502
- American College of Cardiology, 211, 430, 445
- American Heart Association, 302, 430, 445
- American Medical Association, 267
- American Psychiatric Association, 40
- American Society of Clinical Oncologists policy statement, on genetic testing, 266
- American Thoracic Society, 155
- Amoss sign, 396
- amoxicillin, 515, 516, 517, 523
- amphetamine withdrawal, 249
- ampicillin, 515, 516, 517
- Amsel criteria, 694t
 sensitivity, specificity, or likelihood ratio
 for vaginitis, 707t, 694t
- anabolic steroids, 249
- anatomic origin, of abdominal bruit, 29
- aneroid instruments, 305, 306
- angina pectoris, 462
 grading of, 462t
 unstable, 462
- angiotensin-converting enzyme (ACE) inhibitor, 35, 133, 179, 183
- anhedonia, 249
- ankle dorsiflexion. *See* strength testing
- ankle dorsiflexor, 78
- ankle edema
 history of, and ascites, relationship between, 5-6
 sensitivity, specificity, or likelihood ratio
 ascites, 6, 6f, 7, 7f, 68t, 73t
- ankle plantar flexion, 79
- ankle reflexes, 78
- ankle swelling. *See* ankle edema
- ankylosing spondylitis, 77
- anorexia
 sensitivity, specificity, or likelihood ratio
 in acute cholecystitis, 140t
 in appendicitis, adult, 57t
 in temporal arteritis, 648t
- anterior apprehension test
 for labral tears, 586t
- anterior cruciate ligament (ACL), 358, 359, 361, 362
- physical examination
 accuracy of, 363t
 maneuvers, 363t

- anterior drawer test, 360f, 361
 sensitivity, specificity, or likelihood ratio
 knee ligament and meniscus injury, 359f, 370t
- anterior Q waves, 187, 188
- anterior release test
 for shoulder instability, 581f, 585t, 589, 590
 sensitivity, specificity, or likelihood ratio
 for shoulder instability or labral tear, 585t, 591t
- anterior slide test
 for labral tears, 586t
- anticholinesterase test
 sensitivity, specificity, or likelihood ratio
 myasthenia gravis, 460t
- anticholinesterase tests, for myasthenia gravis, 451-453, 455
- anticoagulant therapy, 235, 561, 562
- antidepressants, 247, 249
- antihypertensive therapy, 174
- antimicrobial therapy, 395
- AOM. *See* acute otitis media
- aortic regurgitation (AR), 419, 430t
 anatomic and physiologic origins of, 419
 cardiac auscultation, 420-421
 causes, 420t
 evidence from guidelines, 430
 examination, 419-420, 425
 features, 421f
 findings of, 429
 likelihood ratios, 430t
 of physical examination, 429t
 literature review, 429-430
 literature search, 429
 maneuvers, 421
 original publication data, improvements in, 429
 likelihood ratios, 430t
 in patients with renal failure, 425
 peripheral hemodynamic signs, 421-422
 physical examination
 accuracy of, 423
 precision of, 422-423
 physical examination signs in diagnosis, 430
 prior probability, 430
 reference standard, changes in, 429, 430
- aortic stenosis (AS), 437, 445, 446
 physical examination
 accuracy of, 438t, 445t
 likelihood ratio, 446t
- apical impulse, 187, 190
- Apley compression test, 360f, 361
 sensitivity, specificity, or likelihood ratio
 in knee ligament and meniscus injury, 360f, 364t
- appendectomy
 delayed, 62
 unneeded, 62
- appendiceal anatomy, of appendicitis, 54
- appendicitis, 53-58, 61-63
 appendiceal anatomy of, 54
 diagnostic modalities, accuracy of, 54
 findings of, 61-62
 likelihood ratio, 63
 literature research, 61
 original publication data, improvements in, 62
 pathophysiology of, 54
 prior probability, 63
 reference standard, changes in, 62, 63
 symptoms and signs, 54-55
 accuracy of, 56-57
 precision of, 55
- apprehension test
 for shoulder instability, 581f, 583, 585t
 sensitivity, specificity, or likelihood ratio
 for shoulder instability, 585t
 for labral (shoulder) tear, 586t
- AR. *See* aortic regurgitation
- arm drift
 sensitivity, specificity, or likelihood ratio
 in stroke, 641t
- arm span-height difference test
 for occult vertebral fracture, 479, 482-483
- arrhythmias, 304
- arterial blood gas analysis, 561
- arterial bruit, compared to venous hum, 292t
- AS. *See* aortic stenosis
- ascites, 1, 2, 65, 71
 and ankle swelling, relationship between, 5-6
 evidence from guidelines, 72
 example, 65
 findings of, 71
 history and symptoms, accuracy of, 67, 68t
 information, 71-72
 literature search, 71
 original publication data, improvements in, 71
 pathophysiology of, 66
 physical examination, 66
- likelihood ratios for, 69t
 sensitivity and specificity of, 68t
- physical signs, 72t
- prior probability, 73
- reference standard, changes in, 66, 71, 73
- signs, 66-67, 73t
 accuracy of, 68-69
 precision of, 67-68
- symptoms, 66-67, 71t, 73t
- aspirin, 183
- asthma, 150, 152, 166, 202, 203, 205, 206, 549
- asymmetric skin lesion. *See* ABCD(E) criteria
- asymptomatic abdominal aortic aneurysm
 abdominal palpation for, 19-20, 21t
- asystole, 452
- atherosclerosis
 abdominal bruit in, 30, 32
- atrial fibrillation, 304
- at-risk drinking, 47
 problems of
 in pregnant women, 44-45
- atypical (dysplastic) nevi, 384
- AUDIT. *See* Alcohol Use Disorders Identification Test
- AUDIT-C. *See* Alcohol Use Disorders Identification Test, Consumption Questions
- auscultation, 528
 for airflow limitation, 151, 153, 154-155
- auscultatory characteristics, of abdominal bruits, 32
- auscultatory percussion, 66
 sensitivity, specificity, or likelihood ratio
 for ascites, 72t
- autoimmune disorders, 249
- aztreonam, 517

B

- bacampicillin, 517
- bacterial pneumonia, 540
- balloon angioplasty, 35
- barium enemas
 for appendicitis, 54
- barrel chest sign, 153-154
 sensitivity, specificity, or likelihood ratio
 in obstructive airways disease, 154t
- BCDDP. *See* Breast Cancer Detection Demonstration Project
- BDI. *See* Beck Depression Inventory

- Beck Depression Inventory (BDI), 250, 251. *See also* clinical prediction rules and scores
- bedside ultrasonography
testing, for acute cholecystitis, 145, 146
- benign moles, 383
- benign paroxysmal positional vertigo, 711, 712
- benzyl penicilloyl, 520
- benzylpenicillin, 517
- β -blocker, 133, 506
- β -hemolytic streptococcal pharyngitis, 615, 616, 617, 624, 625
likelihood of, 625
- β -lactam antibiotics
penicillin, cross-reactivity with, 518
- bias
incorporation, 498
reverse workup, 520
spectrum, 141, 497
in thyroid size estimation, 282
verification, 16, 138, 141, 498, 582, 589
- biceps load test I
for labral tears, 586*t*, 590
- biceps load test II
for labral tears, 582*f*, 590
- biceps load tests
for labral tears, 591*t*
- biochemical and hematologic tests, 41-42
- blood pressure (BP). *See also* hypertension
classification of, 302*t*, 311*t*
diastolic, 301, 302
errors in measurement, 303*t*-304*t*
indirect
vs direct blood pressure, 304-305
technical inaccuracies of, 305
measurement, 302*t*
office
factors, affecting, 303-304*t*
vs usual blood pressure, 305
systolic, 301, 302
- blunt abdominal trauma, 72
- BMAST. *See* Brief Michigan Alcoholism Screening Test
- BMD. *See* bone mineral density
- BNP. *See* brain natriuretic peptide
- Boas sign, 138
- bone mineral density (BMD), 477, 478, 484-485
- BP. *See* blood pressure
- brachial-popliteal pulse gradient (Hill sign), 422
- brachioradial delay
for aortic stenosis, 438*t*
- bradycardia, 320, 452
- bradykinesia, 507
maneuvers, detecting, 507*f*
sensitivity, specificity, or likelihood ratio
in Parkinson disease, 507, 514*t*
- bradykinin, 616
- brain natriuretic peptide (BNP), 196, 197, 203, 209
accuracy of, 201-202
sensitivity, specificity, or likelihood ratio
left ventricular dysfunction, 213*t*
- Brain Resuscitation Clinical Trials (BRCTs), 220-221
- BRCTs. *See* Brain Resuscitation Clinical Trials
- breast cancer, 87, 99, 265, 266, 267, 268, 272. *See also* clinical breast examination
evidence from guidelines, 100
findings of, 99-100
incidence of, 88*t*
literature search, 99
pathophysiology of, 99
prior probability, 101
reference standard tests, 101
risk factors for, 88
- Breast Cancer Detection Demonstration Project (BCDDP), 88, 92
- Breast-ovarian cancer syndrome, 275
- breast tenderness, 551, 552
- Breathing Not Properly Multinational Study, 202
- breath sounds
sensitivity, specificity, or likelihood ratio
pneumonia, adult, 531*t*, 536*t*
- Breslow-Day test, 554
- Brief Michigan Alcoholism Screening Test (BMAST), 41, 44
- British Hypertension Society, 312
- British Thoracic Society, 574
- bronchial lavage, 540
- bronchiolitis, 540, 541*f*
- bronchitis, 527
- bronchoconstriction, 452
- bronchodilators, 149
- Brudzinski signs, 396, 399
sensitivity, specificity, or likelihood ratio
in meningitis, adult, 404*t*
- bruit
abdominal, 29-32, 35-37
carotid, 103-106
perumbilical, 30
- systolic, 35, 36
- systolic-diastolic abdominal, 30, 31, 32
- sensitivity, specificity, or likelihood ratio
of abdominal, in renovascular hypertension, 31*t*, 37*t*
of carotid, in carotid stenosis, 109*f*, 110*t*
- vascular, 29
- bulbar weakness, 451
- bulging flanks, 66
sensitivity, specificity, or likelihood ratio
ascites, 69*t*

C

- C rating, 18. *See also* evidence, level of CAGE questionnaire. *See also* cut down, annoyed by criticism, guilty about drinking, eye-opener drinks
sensitivity, specificity, or likelihood ratio
for alcohol abuse, 5*f*, 41*t*
calcium-channel blocker, 183
calf vein thrombosis, 227
Canadian Cardiovascular Society, 462
Canadian class II angina, 183
Canadian Hypertension Education Program, 312
Canadian National Breast Screening Study (NBSS), 88
Canadian Preventive Health Services Task Force, 490
Canadian Task Force, 99, 109
on the Periodic Health Examination, 18
on Preventive Health Care, 49, 261, 393
for malignant melanoma, 393
cancer, 76, 86
family history of, 265-272, 275-276
accuracy of, 268-270
data collection, improvement in, 270-272
elicitation of, 266-267
false-negative reports, reasons for, 270
false-positive reports, reasons for, 270
findings of, 275
guidelines, evidence from, 275
information, collection of, 270-272
likelihood ratio, 276

- literature review, results of, 275
 literature search, 275
 methods, 267-268
 precision of, 268
 prevalence of, 266
 prior probability, 276
 reference standard, changes in, 275, 276
candidiasis, 706*t*, 707*t*, 694*f*, 697*t*
 sensitivity, specificity, or likelihood ratio
 in pharyngitis, 616*t*
 in vaginitis, 707*t*, 694*t*
capillary refill time, 318, 331
 sensitivity, specificity, or likelihood ratio
 in hypovolemia, adult, 327*t*
 in hypovolemia, child, 341*t*
carbapenems, 515, 518, 524
cardiac arrest, 215
 comatose survivors of, 215-223, 225-226
 literature search, 225
 prior probability, 226
 reference standard tests, 226
 ventricular fibrillation, 215
cardiac bradyarrhythmias, 452
cardiac dullness. *See* percussion
cardiac ischemic chest pain, 462, 463.
See also chest pain; myocardial infarction
carotid arterial pulse
 and jugular venous pulse,
 distinguishing between, 127
carotid bruit, 103-106, 110*t*
 ambulatory bruit, 104-105
 auscultation, precision of, 104
 carotid artery cause, in neck, 103-104
 clinical significance, 103
 evidence from guidelines, 109
 findings of
 asymptomatic patients, 107
 symptomatic patients, 107
 likelihood ratio, 108, 109, 110
 literature review
 asymptomatic patients, 108-109
 symptomatic patients, 108
 literature search, 107
 original publication data,
 improvements in, 107
 preoperative bruit, 105
 prior probability, 110
 reference standard, changes in, 107-108
 symptomatic bruit, 105
carotid pulse, for aortic stenosis, 446*t*
carotid volume, for aortic stenosis, 446*t*
- carpal tunnel syndrome (CTS)**, 111-117
 diagnostic standard, 112-113
 evidence from guidelines, 123
 findings of, 121-122
 history and physical examination
 accuracy of, 114-117
 precision of, 114
 importance of, 111-112
 likelihood ratio, 124
 literature review, 123
 literature search, 121
 methods, 113
 normal anatomy of, 112*f*
 original publication data,
 improvements in, 122
 prior probability, 124
 reference standard, changes in, 122-123, 124
 sensitivity and specificity of electrodiagnosis, 112
 signs, 113
 symptoms, 113
case-control study, 88
case-finding, 248
 instruments, 250
 performance, in primary care settings, 252-253*t*
Castell method, 607, 607*f*, 608
Castell sign
 sensitivity, specificity, or likelihood ratio
 in splenomegaly, 612*t*
cauda equina syndrome, 80
CBE. *See* clinical breast examination
CDC. *See* Centers for Disease Control and Prevention
Center for Epidemiologic Studies Depression Screen (CES-D), 250, 251, 260. *See also* clinical prediction rules and scores
Centers for Disease Control and Prevention (CDC), 267, 330, 355, 356, 405, 524, 706
US Influenza Sentinel Providers Surveillance Network, 344
Centor clinical prediction rule
 for sore throat, 619, 619*f*
 sensitivity, specificity, or likelihood ratio
 adults, 619*f*
 modified for age, 623-624, 623*t*
central venous pressure (CVP), 125-130, 133-135
abdominojugular reflux, 134*t*
abnormal, 127
 clinical assessment of, 126-128
 abdominojugular reflux test, 128
 accuracy of, 129-130
- carotid arterial pulse and jugular venous pulse**, distinguishing between, 127
jugular veins, 130
Kussmaul sign, 128
neck veins, position of, 126-127
 precision of, 128-129
 estimation of, 127, 128*f*
 findings of, 133
 guidelines, evidence from, 134
 likelihood ratio, 135
 literature review, results of, 134
 literature search, 133
 original publication data,
 improvements in, 134
 prior probability, 135
 reference standard, changes in, 134, 135
 venous waveforms in, 126*t*
cephalosporins, 515, 516, 517, 524
cerebral infarction
 Oxfordshire classification of, 634
 cerebrospinal fluid (CSF), 395-396
CES-D. *See* Center for Epidemiologic Studies Depression Screen
Chadwick sign, 552, 556, 557
 for pregnancy, 556*t*
chance agreement, 3-4
chest hyperresonance, 154
chest pain. *See also* cardiac ischemic chest pain; myocardial infarction; pain
 acute, 462-463
 cardiac ischemic, 462, 463
 mechanism of, 463
chest radiograph, 195, 202, 211, 661
 accuracy of, 201
 for community-acquired pneumonia, 528, 530, 532
 for community-required pneumonia, 535, 536, 537
 for pneumonia, 540, 548
 for pulmonary embolism, 561
 for thoracic aortic dissection
 accuracy of, 666
 sensitivity of, 667*t*
chest retractions. *See* retractions, chest
chest x-ray. *See* radiographic findings
 χ^2 test, 464, 617
 for abdominal aortic aneurysm, 19
 2-tailed χ^2 test, 529
chlamydia, 541
Chlamydia pneumoniae, 344
cholecystectomy, 39, 138
cholesteatoma, 712
chronic bronchitis, 150, 151, 152

- chronic obstructive pulmonary disease (COPD), 163, 166, 168, 169, 195, 202, 203, 205, 206, 489, 490
- chronic thromboembolic pulmonary hypertension, 235
- cigarette smoking, 151
- Cincinnati Prehospital Stroke Scale (CPSS), 628, 630, 631
- for stroke, 641*t*
- cirrhosis, 292
- classic essential tremor, 506, 514
- clinical agreement, 3-4
- clinical assessment
- accuracy of, 9-16
 - for airflow limitation, 155
 - accuracy of, 155
- CAGE questionnaire
- accuracy characteristics of, 4-5
- of central venous pressure, 126-128
- abdominojugular reflux test, 128
 - accuracy, 129-130
 - carotid arterial pulse and jugular venous pulse, distinguishing between, 127
 - jugular veins, 130
 - Kussmaul sign, 128
 - neck veins, position of, 126-127
 - precision, 128-129
- for clubbing
- accuracy, 168-169
 - precision, 168
- for coma
- accuracy of, 219
 - interobserver agreement of, 218*t*
 - precision of, 219
- for congestive heart failure, 195-206
- accuracy of, 199, 200*t*
 - precision of, 199
- of deep vein thrombosis, 228-229, 235-236
- importance of, 173-174
 - for internal derangement of knee, 359-361
 - function, 360-361
 - inspection, 359
 - palpation, 359-360
- for mitral regurgitation, 438-439
- accuracy of, 439*t*
- for mitral valve prolapse, 439-440
- accuracy of, 440*t*
 - precision of, 9-16
- for spider nevi
- precision of, 3*f*
- for systolic murmurs
- accuracy of, 436, 437*t*
 - precision of, 435-436, 436*t*
- for thoracic aortic dissection
- accuracy of, 662*t*
- clinical breast examination (CBE), 87.
- See also* breast cancer
- accuracy, 90-91
- bottom line, 91
 - Anatomic basis, 87-88
- bottom line
- priorities for research, 95-96
 - resolution of scenarios, 95
- effectiveness, 88-90
- bottom line, 90
- examiner factors
- bottom line, on accuracy, 92
 - duration, 91
 - experience, 91-92
 - techniques, 91
- methods, 88
- patient factors
- age, 92
 - bottom line
 - on accuracy, 92
 - of suggested approach, 94
- breast boundaries, 93
- breast characteristics, 92
- cancer characteristics, 92
- duration, 94
- examiner pattern, 93
- fingers, 93-94
 - inspection, 94
 - issues, 94
 - normal from abnormal (cancerous)
 - lumps, distinguish, 95
- palpation, 92
- patient position, 93
 - techniques, 94-95
- precision, 90
- bottom line, 90
- sensitivity, specificity, or likelihood ratio
- for acute cholecystitis, 147*t*
 - for aortic aneurysm, 24*t*, 134*t*, 147*t*
 - for aortic regurgitation, 429*t*
 - for central venous pressure, 134*t*
 - for chronic obstructive airways disease, 159*t*
 - for hypovolemia, child, general appearance, 341*t*
 - for left ventricular dysfunction, 188*t*, 213*t*
 - for pneumonia, infant and child, 548*t*
 - for pulmonary embolus, 572*t*, 575*t*
 - for valvular heart disease, 446*t*
- clinical interview, for depression
- accuracy and reliability of, 254, 255
- clinical prediction guide, for deep vein thrombosis, 230
- development of, 230-232
- clinical prediction rules and scores
- ABCD(E) criteria, for melanoma, 392*t*, 393*t*
 - Alvarado score, for appendicitis, adult, 63*t*
 - for aortic stenosis, 438*t*
 - Beck Depression Inventory, for major depression, 252*t*
 - Center for Epidemiologic Studies Depression (CES-D), for major depression, 252*t*
 - Cincinnati Prehospital Stroke Scale, for stroke, 641*t*
 - for deep vein thrombosis, 235-236, 237-238
 - Glasgow Coma Scale, in recovery from coma, 220*t*
 - Malnutrition Screening Tool, for malnutrition, adult, 381*t*

for myocardial infarction, 476*t*, 468-469
 for osteoporosis in men, 490*t*
 for osteoporosis in women, 491*t*
 Patient Health Questionnaire (PHQ-9)
 for depression, 261*t*
 for dysthymia, 261*t*
 for pneumonia, adult, 536*t*
 pneumonia, infant and child, 549*t*
 PRIME-MD, for depression, 261*t*
 of pulmonary embolism, 564-566, 567*f*, 572*t*, 575*t*
 accuracy of, 565*t*, 568*t*
 components of, 566
 validation of, 566-567
 for sinusitis, 603*t*, 625*t*
 for sore throat, 618-620, 619*t*
 Centor clinical prediction rule, 619, 619*f*
 McIsaac clinical prediction rule, 620, 620*f*
 Walsh algorithm, 621*f*
 subjective global assessment (SGA),
 for malnutrition, adult, 376*t*
 for temporal arteritis, 654
 for urinary tract infection, women, 689*t*
 Wells Prediction Rule, for deep vein thrombosis, 246*t*
 Wicki model, 566*t*, 567
 closed fist sign, 112*t*
 clubbing, 163-169
 clinical examination for
 accuracy of, 168-169
 precision of, 168
 congenital, 163
 data analysis, 166
 digital, 163
 findings of, 171
 guidelines, evidence from, 172
 inspection
 general appearance, 164, 165*f*
 nailfold angles, 164, 165*f*
 palpation of, 165-166
 phalangeal depth ratio, 164-165, 165*f*
 Schamroth sign, 165, 165*f*
 literature review, results of, 172
 literature search, 171
 methods, 166
 original publication data,
 improvements in, 171
 pathophysiology of, 164
 prevalence in associated conditions, 172*t*
 prevalence of, 162
 reference standard, changes in, 171-172, 172*t*

results
 quality of evidence, 166
 quantitative indices, 166-168
 signs of, 164
 study characteristics, 166
 symptoms of, 164
 clunk test
 for shoulder instability, 585*t*
 sensitivity, specificity, or likelihood ratio
 for shoulder instability, 585*t*
 cocaine, 249, 301, 660
 cog wheeling, 506. *See also* rigidity
 sensitivity, specificity, or likelihood ratio
 in Parkinsonism, 506
 cold, 527, 541*f*, 593, 596,
 cold water caloric testing, 217
 colon cancer, 265, 267
 colors, multiple in a skin lesion. *See* ABCD(E) criteria
 coma
 clinical examination for
 accuracy of, 219*t*
 interobserver agreement of, 218*t*
 precision of, 219
 hypoxic-ischemic, 216
 methods
 likelihood ratios, 218
 search strategy and quality review, 217-218
 statistical methods, 218
 motor response and brainstem reflexes, 219-221
 pathophysiology of, 216
 physical examination of, 216-217
 postcardiac arrest, 215, 216
 search results and quality of evidence, 218-219
 combination chemotherapy, 228
 community-acquired pneumonia, adult, 527-533, 535-537
 diagnosis of
 clinical history, accuracy of, 530
 physical examination findings, accuracy of, 530-532
 findings of, 535
 guidelines, evidence from, 536
 likelihood ratio test for, 537
 literature review, results of, 536
 literature search, 535
 methods
 data analysis, 529
 literature search, 528-529
 quality review, of articles, 529
 multivariate findings for, 536*t*
 original publication data,
 improvements in, 535
 pathophysiology of, 528
 prediction of
 algorithm evaluation, 532-533
 prior probability, 537
 reference standard, changes in, 535-536, 537
 symptoms and signs
 elicitation of, 528
 precision of, 529-530
 compliance, 173-174. *See also* noncompliance
 clinical measures, accuracy of, 175-176
 Compliance Questionnaire
 Rheumatology, 180
 Comprehensive Meta-Analysis software (version 2.197), 236
 compression rotation test
 for labral tears, 586*t*
 compression ultrasonography, 230
 computed tomography (CT), 17, 18
 for appendicitis, 54
 computed tomography (CT)
 angiography
 for pulmonary embolism, 571, 572
 computed tomography (CT) scanning
 for acute cholecystitis, 138
 chest CT, for community-acquired pneumonia, 536
 for paranasal sinuses, 594
 computer-guided analyses
 for appendicitis, 54
 computerized genograms, 271
 confidence interval, 6, 12
 congenital clubbing, 163
 congestive heart failure, dyspnea in, 195-206
 in emergency department patients
 brain natriuretic peptide, accuracy of, 201-202, 203
 chest radiographs, accuracy of, 201, 202
 clinical examination and investigations
 accuracy of, 199, 200*t*
 precision of, 199
 clinical gestalt, 199, 202
 clinician's assessment, 204
 limitations, 204-205
 electrocardiogram, accuracy of, 201, 202
 historical items, 199, 202
 pathophysiology of, 196
 physical examination, 200-201, 202
 physiological categories and mechanisms of, 196*t*
 pulmonary diseased patients, 202
 search strategy, 196-197

- congestive heart failure, dyspnea in
(Continued)
 statistical methods, 197
 study characteristics, 198, 198-199^t
 study quality, assessment of, 197
 study selection, 197
 symptoms, 199-200, 202
 and signs, elicitation of, 196
COPD. *See* chronic obstructive pulmonary disease
Cope's Early Diagnosis of the Acute Abdomen, 55, 138
Copenhagen Stroke Scale, 634
coronary heart disease, 249
corneal reflex. *See* reflexes
Corrigan water hammer pulse, 422
 sensitivity, specificity, or likelihood ratio
 in aortic regurgitation, 425^t
coryza, 571, 574
costovertebral angle tenderness
 sensitivity, specificity, or likelihood ratio
 in influenza, 355^t
 in obstructive airways disease, 152^t
 in otitis media, child, 497^t
 in pneumonia, adult, 530^t
 in streptococcal pharyngitis, 618^t
Courvoisier sign, 138
cover-uncover test, 451
CPSS. *See* Cincinnati Prehospital Stroke Scale
crank test
 for labral tears, 586^t
crossed straight-leg raising sign (CSLR)
 sign, 78
 sensitivity, specificity, or likelihood ratio
 for disk herniation, 86^t
CSF. *See* cerebrospinal fluid
CSLR. *See* crossed straight-leg raising sign
CT. *See* computed tomography
CTS. *See* carpal tunnel syndrome
curtain sign, 451
Cushing disease, 249
cut down, annoyed by criticism, guilty about drinking, eye-opener drinks (CAGE questionnaire)
 accuracy of, 42
 predictive, 44
questionnaire, 41, 43^t, 49, 52^t
 accuracy characteristics of, 4-5
 for alcohol abuse or dependency, 2-3, 4-5, 7, 8^t
reliability of, 42
sensitivity, specificity, or likelihood ratio
 for alcohol abuse, 5^f, 41^t
CVP. *See* central venous pressure
cyanotic congenital heart disease, 163
cystic fibrosis, 163
- D**
- DADS.** *See* Duke Anxiety and Depression Scale
DBP. *See* diastolic blood pressure
D-dimer assay, 230, 236, 238-239, 561, 562, 563, 566, 567-568, 569, 571, 572, 573, 575. *See also* laboratory findings
 in deep vein thrombosis diagnosis, 239
 high-sensitivity, 240, 241^t
 likelihood ratio of, 573^t
 moderate-sensitivity, 239-240
de Musset head bobbing sign, 421
decision analytic model, 352
deep vein thrombosis (DVT), 227-232, 245-246
 clinical assessment of, 228-229, 235-236
 clinical prediction guide, 230
 development of, 230-232
 clinical prediction rules, 235-236
 data extraction, 236
D-dimer testing for, 239
 high-sensitivity, 240, 241^t
 moderate-sensitivity, 239-240
 diagnosis of, 228, 232
 likelihood ratio, 246
 objective assessment of, 229-230
 original data publication, improvements in, 245
 prevalence of, 238^f
 prior probability, 246
 reference standard tests, 246
 search strategy, 228
 statistical analysis, 236-237
 study identification, 236
 study selection, 236
 symptoms and signs, frequency of, 229^t
 ultrasonography testing for, 240-241
dehydration, 315
 in children, 329-330
- anatomic and physiologic origins** of, 330
evidence from guidelines, 340
examination signs, precision of, 333-335
findings of, 339-340
laboratory tests, 335
likelihood ratio, 341
limitations, 335-336
literature review, 340
literature search, 339
methods
 search strategy and quality review, 331-332
 statistical analyses, 332-335
original publication data, improvements in, 340
prior probability, 341
reference standard tests, 341
symptoms and signs, 330-331
 accuracy of, 333
 precision of, 333
Dehydration Assessment Scale, for hypovolemia, child, 334^t
delayed menses, for pregnancy, 560
depressed mood, perimenopausal, 409
Depression Scale (DEPS), 250, 251
DEPS. *See* Depression Scale
DerSimonian-Laird random-effects method, 509
deviated nasal septum, 594, 595
diabetes mellitus, 249
Diagnostic and Statistical Manual of Mental Disorders (Third Edition Revised) (DSM-III-R), 40, 259
Diagnostic and Statistical Manual of Mental Disorders (Fourth Edition) (DSM-IV), 48, 247, 249, 259
diagnostic odds ratio (DOR), 12, 13
diaphoresis
 sensitivity, specificity, or likelihood ratio
 in myocardial infarction, 467^t
diastolic blood pressure (DBP), 301, 302, 303
diastolic dysfunction, 184
 and systolic dysfunction, difference between, 189, 211
diastolic murmur. *See* aortic regurgitation
digital clubbing, 163
diplopia, 455
 sensitivity, specificity, or likelihood ratio
temporal arteritis, 656

dipstick urinalysis
for urinary tract infection
accuracy of, 678

direct blood pressure
vs indirect blood pressure, 304-305

diuretic, 179

therapy, 184

Dix-Hallpike maneuver, 710*f*, 715, 716, 717

dizziness. *See also* vertigo
sensitivity, specificity, or likelihood ratio
postural, in hypovolemia, adult, 327*t*

Doppler echocardiography, 430

Doppler effect, 553

DOT. *See* diagnostic odds ratio

drink, 41, 48

dry axilla
sensitivity, specificity, or likelihood ratio
in hypovolemia, adult, 327*t*

dry mucous membranes
sensitivity, specificity, or likelihood ratio
in hypovolemia, adult, 327*t*
in hypovolemia, child, 341*t*

DSM-III-R. *See Diagnostic and Statistical Manual of Mental Disorders* (Third Edition Revised)

DSM-IV. *See Diagnostic and Statistical Manual of Mental Disorders* (Fourth Edition)

Duke Anxiety and Depression Scale (DADS), 250, 251

Duroziez double intermittent femoral bruit, 422, 425*t*
sensitivity, specificity, or likelihood ratio
in aortic regurgitation, 425*t*

DVT. *See* deep vein thrombosis

dyskinesias, 506

dyspnea, 125, 149, 152, 153, 183, 186, 187, 215, 225
in congestive heart failure, 195-206
sensitivity, specificity, or likelihood ratio
in congestive heart failure, 200*t*
in pneumonia, adult, 530*t*

dyspnea on exertion. *See* dyspnea

dysthymia, 248

dysuria
sensitivity, specificity, or likelihood ratio
in urinary tract infection, women, 689*t*
in vaginitis, 707*t*

E

ear rubbing
sensitivity, specificity, or likelihood ratio
in otitis media, child, 503*t*

ECG. *See* electrocardiogram

echocardiogram, 209
for left ventricular systolic dysfunction, 210-211, 212

edema, 187, 617
sensitivity, specificity, or likelihood ratio
in left ventricular dysfunction, 200*t*

edrophonium chloride, 451-452

edrophonium test, 452, 453*t*. *See also* anticholinesterase test

effectiveness score, 250, 253

effusion, 359-360

egophony, 528

ejection fraction, detection of, 187-189

electrocardiogram (ECG), 202
accuracy of, 201

left bundle-branch block on, 187

for myocardial infarction
accuracy of, 468
precision of, 465-466

for pulmonary embolism, 561-562

sensitivity, specificity, or likelihood ratio
for left ventricular dysfunction, 213*t*

for thoracic aortic dissection, 665*t*

for myocardial infarction, 468*t*

ELISA. *See* enzyme-linked immunosorbent assay

ELISA Vidas DD, 573

emphysema, 150

endemic iodine deficiency, 285

endometrial cancer, 265, 267

enhanced ptosis. *See* curtain sign

enlarging skin lesion. *See* ABCD(E) criteria

enzyme-linked immunosorbent assay (ELISA), 230, 239, 572

epiglottitis, 540, 541*f*

erythema, 617

erythrocyte sedimentation rate (ESR), 649

erythromycin, 524

ESR. *See* erythrocyte sedimentation rate

estradiol, perimenopausal, 408, 410

ethmoid sinus, 594, 595, 595*f*

ethnicity
sensitivity, specificity, or likelihood ratio
in osteoporosis, 491*t*

in perimenopause, 416*t*

in thoracic aortic dissection,
enlarged aorta or wide mediastinum, 673*t*

European Influenza Surveillance Scheme, 344

European Society of Cardiology, 212

European Society of Hypertension, 312

Evaluation du Scanner Spirale dans l'Embolie Pulmonaire study

group, 564

evidence, level of, 15*t*

expected agreement. *See* agreement

Extended Wells scoring system, 571, 573

extraocular muscles, asymmetric

weakness of, 451

eye movements, in coma, 220*t*. *See also* reflexes

F

facial paresis

sensitivity, specificity, or likelihood ratio

in stroke, 641*t*

facial weakness, 451, 455

family history

sensitivity, specificity, or likelihood ratio

in cancer, 275*t*

in early menopause, 417*t*

female. *See* sex

femoral pistol shot murmur, 422, 425*t*

femur, 358, 359, 360

fever

sensitivity, specificity, or likelihood ratio

in acute cholecystitis, 140*t*

in appendicitis, adult, 57*t*

in influenza, 356*t*

in meningitis, adult, 404*t*

in otitis media, child, 497*t*

in pneumonia, adult, 536*t*

in pneumonia, infant and child, 550*t*

in streptococcal pharyngitis, 618*t*

in temporal arteritis, 648*t*

in urinary tract infection, women, 679*t*

fibromuscular hyperplasia

abdominal bruit in, 30

finger-flicking percussion, 66-67

Fisher exact test, 529

flank dullness, 66, 67. *See also* percussion

flexicurve measurement, 479, 485

flick sign, 112*t*

Flint murmur, 420-421
 in aortic regurgitation, 420
 fluctuating weakness. *See* reduced muscle power
 fluid loss, subjective global assessment, 374-375
 fluid wave, 66, 67. *See also* percussion
 follicle-stimulating hormone (FSH), perimenopausal, 408, 410, 416, 417t
 fontanelle, sunken
 sensitivity, specificity, or likelihood ratio
 in hypovolemia, child, 334t
 Food and Drug Administration, 197, 416
 forced expiratory time, 151, 159t
 in obstructive airways disease, 161t
 Fracture Intervention Trial, 482
 fractures, spinal compression, 77
 Framingham study, 463
 frequent urination
 sensitivity, specificity, or likelihood ratio
 in urinary tract infection, women, 689t
 frontal sinus, 594f, 595, 595f
 surface palpation for, 596f
 FSH. *See* follicle-stimulating hormone

G

gag reflex. *See* reflexes
 gait
 sensitivity, specificity, or likelihood ratio
 in Parkinsonism
 heel to toe walking, 514t
 rising from a chair, 514t
 shuffling, 514t
 gastroenterologist, 293
 gastrointestinal (GI)
 symptoms, 374
 tract hemorrhage, 316
 GCS. *See* Glasgow Coma Scale
 GDS. *See* Geriatric Depression Scale
 General Health Questionnaire (GHQ-12), 260
 genetic testing
 family history assessment tools for, 270
 policy statement on, 266
 Geneva rule, 572, 573
 Geriatric Depression Scale (GDS), 250, 251, 260
 GHQ-12. *See* General Health Questionnaire

GI. *See* gastrointestinal
 girth, increased abdominal
 sensitivity, specificity, or likelihood ratio
 in ascites, 68t, 73t
 glabella tap reflex test, 508, 508f, 510, 513, 514. *See also* reflexes
 Glasgow Coma Scale (GCS), 216, 216t, 221. *See also* clinical prediction rules and scores
 Glasgow-Pittsburgh Cerebral Performance Categories, 217
 in coma, 217
 Global Initiative for Chronic Obstructive Lung Disease, 161
 glucocorticoids, 249
 goiter, 277-282, 285-287
 accuracy of, 281-282
 anatomic basics of landmarks, 277-278, 278f
 normal size, 278
 examination, 278-279
 bias in, 282
 false-negative results, 280
 false-positive results, 279-280
 findings of, 285
 guidelines, evidence from, 286
 likelihood ratio, 287
 literature review, results of, 286
 literature search, 285
 original publication data, improvements in, 285
 precision of
 interobserver variability, 280-281
 intraobserver variability, 281
 prior probability, 287
 reference standard, changes in, 285-286, 287
 size of, 277
 Goldman chest pain protocol, 474f, 475
 "good" clinical finding, 11-12
 Goodell sign, 552
 sensitivity, specificity, or likelihood ratio
 in pregnancy, 556
 grades of evidence. *See* levels of evidence
 Graves disease, 277
 great toe extensor weakness
 sensitivity, specificity, or likelihood ratio
 in sciatica, 79t
 grind test
 sensitivity, specificity, or likelihood ratio
 for knee ligament and meniscus injury, 370t

grunting
 sensitivity, specificity, or likelihood ratio
 in pneumonia, infant and child, 550t
 guarding, 55
Guide to Clinical Preventive Services, Third Edition, Periodic Updates, 47

H

HADS. *See* Hospital Anxiety and Depression Scale
Haemophilus influenzae, 400, 494, 501
 hand diagram. *See* Katz hand diagram
 hand grip strength test
 for occult vertebral fracture, 479, 485
 harmful drinking, 41, 47
 Hawkins grading scheme, 579
 hazardous drinking, 41, 47
 HCG test. *See* human chorionic gonadotropin test
 headache. *See* pain
 head-hanging maneuver, vertigo, 712
 Health Canada, 344
 Health Insurance Plan (HIP) study, 89
 HealthSTAR database, 361
 hearing loss
 in benign positional vertigo, 716
 in labyrinthitis, 716
 in Ménière disease, 716
 in vertigo, 716
 in vestibular neuronitis, 716
 heart failure, 195
 ascites, 65
 heart sounds
 sensitivity, specificity, or likelihood ratio
 S2 (second heart sound) for aortic stenosis, 446t
 S3 (third heart sound)
 for aortic regurgitation, 430t
 for myocardial infarction, 467t
 for the breathless emergency patient, 200t
 for ventricular dysfunction, 213t
 S4 (fourth heart sound)
 for aortic stenosis, 438t
 for ventricular dysfunction, 200t
 heel-to-toe test, 513
 Hegar sign, 552, 553f
 sensitivity, specificity, or likelihood ratio
 for pregnancy, 556
 height loss
 in osteoporosis, 482, 483t, 490t

hemagglutination-inhibition method, 555
 hematuria, in urinary tract infection, women, 679t
 hemorrhagic stroke, 635
 heparin-induced thrombocytopenia, 235
 hepatojugular reflux, 128
 hepatomegaly
 findings of, 299
 guidelines, evidence from, 300
 likelihood ratio, 300
 literature review, results of, 299
 literature search, 299
 original publication data,
 improvements in, 299
 prior probability, 300
 reference standard, changes in, 299, 300
 hereditary cancer syndrome, 265, 267
 hereditary nonpolyposis colon cancer (HNPCC), 266, 267, 268
 herniated disk
 with radiculopathy in low back pain, 86
 herpes zoster oticus, 712
 HIP. *See* Health Insurance Plan study
 HIV. *See* human immunodeficiency virus
 history
 of penicillin allergy, 525t
 HNPCC. *See* hereditary nonpolyposis colon cancer
 Homans sign, 228
 sensitivity, specificity, or likelihood ratio
 in deep vein thrombosis, 229t
 home pregnancy test (HPT), 559, 560, 560t
 accuracy of, 555-556
 likelihood ratios of, 559, 560t
 Hoover sign, 542
 Hopkins Symptom Check List (HSCL), 250, 251, 252
 hormone replacement therapy (HRT), 411
 Hospital Anxiety and Depression Scale (HADS), 259, 260
 hot flashes, perimenopausal, 409
 Hoyne sign, 396
 HPT. *See* home pregnancy test
 HRT. *See* hormone replacement therapy
 HSCL. *See* Hopkins Symptom Check List
 human chorionic gonadotropin (HCG) test, 553, 554, 555, 557
 human immunodeficiency virus (HIV), 611

humped back, in osteoporosis, 481t
 hydrochlorothiazide, 316
 hypalgesia, 112t
 hypernatremia, 316
 hypertension, 183, 301-307. *See also* blood pressure
 classification of, 311
 diagnosis of
 guidelines for, 301-302
 potential improvements in, 306-307
 findings of, 311
 guidelines, evidence from, 312-313
 likelihood ratio, 313
 literature research, results of, 312
 literature search, 311
 measurement of
 accuracy of, 304
 techniques for, 302t
 variation in, 303-304
 original publication data,
 improvements in, 311-312
 prediction, issue of
 blood pressure now vs blood pressure later, 305-306
 palpation, 306
 relative risk, 306
 prior probability, 313
 reference standard, changes in, 312, 313
 sensitivity, specificity, or likelihood ratio
 in left ventricular dysfunction, 211t
 in thoracic aortic dissection, 666t
 hyperthyroidism, 277
 hypertrophic cardiomyopathy, 439
 hypertrophic osteoarthropathy, 163, 164
 hypotension
 in myocardial infarction, 467t
 postural, 318, 319
 supine, 320
 hypothyroidism, 249, 277
 hypovolemia, 315-316
 acute blood loss
 physical signs, accuracy of, 319-320
 clinical study, 317t
 findings of, 325
 likelihood, 327t
 in ICU patients, 326t
 literature review, 326
 literature search, 325
 methods, 316-317
 multivariate findings for, 326
 pathogenesis, 318-319
 physical signs
 accuracy of, 320-321
 precision of, 319
 postural vital signs, 317-318

prior probability, 327
 reference standard, changes in, 325, 327
 ICD-10. *See International Statistical Classification of Diseases, 10th Revision*
 ice pack test, 451, 454
 precision of, 455
 sensitivity, specificity, or likelihood ratio
 for myasthenia gravis, 460t
 ICS. *See* Innsbruck Coma Scale
 idiopathic dilated cardiomyopathy, 183
 idiopathic penicillin hypersensitivity, 517
 IgE antibodies, 519
 imipenem, 517
 immediate penicillin hypersensitivity, 516-517
 impedance plethysmography, 229-230
 incorporation bias, 498
 increased abdominal girth. *See* girth incontinence
 sensitivity, specificity, or likelihood ratio
 in perimenopause, 412t
 independence, 13-14
 indirect blood pressure
 vs direct blood pressure, 304-305
 technical inaccuracies of, 305
 inelastic skin, 331
 infiltrative disorders, 292
 influenza, 343-344, 355-356
 clinical findings, 347t
 diagnosis test, 350-352
 likelihood of, 356
 methods
 diagnostic odds ratio, 346
 search strategy and quality review, 344-346
 statistical methods, 346
 prior probability, 356
 reference standard tests, 356
 signs and symptoms
 accuracy of, 346-350
 precision of, 346
 Innsbruck Coma Scale (ICS), 221
 Integrated Management of Childhood Illness Scale, 330
 intention tremor, 506
 interleukin 1, 616
 interleukin 6, 616
 internal rotation resistance strength test
 for labral tears, 582f, 586t

International Registry of Acute Aortic Dissection (IRAD), 671
International Statistical Classification of Diseases, 10th Revision (ICD-10), 40, 249
 ipsilateral straight-leg raising sign, 78
 IRAD. *See* International Registry of Acute Aortic Dissection
 irregular border skin lesion. *See* ABCD(E) criteria
 irritability, perimenopausal, 409
 ischemic stroke, 635
 subtype analysis, 635-636
 accuracy of, 635
 reliability of, 635-636
 transient, 627, 631-633
 itching
 sensitivity, specificity, or likelihood ratio
 in vaginitis, 707t

J

jaw claudication, in temporal arteritis, 656t
 JNC-VII. *See* Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure, seventh report of
 joint line tenderness, 360. *See also* pain
 Joint National Committee, 305
 Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure, seventh report of (JNC-VII), 312
 jolt accentuation of headache, 396, 400
 in meningitis, adult, 404t
 jugular veins, 190
 jugular venous distention, 186, 187
 sensitivity, specificity, or likelihood ratio
 in left ventricular dysfunction, 213t
 jugular venous pressure (JVP), 125, 133, 134
 anatomic and physiologic origins of, 125-126
 jugular veins, clinical examination of, 130
 waveforms, analysis of, 126
 abnormalities, 126
 jugular venous pulse
 and carotid arterial pulse, distinguishing between, 127
 JVP. *See* jugular venous pressure

K

κ statistic, 152, 153. *See also* precision calculation of, 8f
 weighted, 571
 Kartagener syndrome, 594
 Katz hand diagram, 113, 114f
 sensitivity, specificity, or likelihood ratio
 in carpal tunnel syndrome, 115t, 123t, 124
 Kernig signs, 396, 399
 sensitivity, specificity, or likelihood ratio
 in meningitis, adult, 404t
 knee effusion
 sensitivity, specificity, or likelihood ratio
 in knee ligament and meniscus injury, 370t
 knee, meniscal, and ligamentous injuries, 357-358
 anatomy, 358
 anterior cruciate ligament (ACL) examination, 362
 physical examination
 accuracy of, 363t
 maneuvers, 363t
 clinical examination for internal derangement, 359-361
 function, 360-361
 inspection, 359
 palpation, 359-360
 epidemiology of, 359
 findings of, 369
 likelihood of, 370
 limitations, 363, 365
 literature review, 369
 literature search, 369
 mechanism, 358-359
 methods
 analysis, 362
 search strategy, 361-362
 original publication data, improvements in, 369
 physical examination, 366t
 accuracy of, 364t
 maneuvers, 365t
 posterior cruciate ligament (PCL) examination, 362
 physical examination
 accuracy of, 364t
 prior probability, 370
 reference standard, changes in, 369, 370
 symptoms, 358
 Kussmaul sign, 128
 kwashiorkor, 372

kyphosis, 477, 478, 479, 485
 in osteoporosis, 490t

L

laboratory findings
 sensitivity, specificity, or likelihood ratio
 for acute cholecystitis, 140t
 for adult malnutrition, albumin, 380t
 for deep vein thrombosis, D-dimer, 246t
 for hypovolemia, adult, urine specific gravity, 327t
 for hypovolemia, child, 334t
 for influenza, rapid tests, 356t
 for left ventricular dysfunction, brain natriuretic peptide, 213t
 for malnourishment, adult 380t
 for perimenopause, 417t
 for pulmonary embolus, D-dimer, 575t
 for streptococcal pharyngitis, rapid streptococcal test, 625t
 for temporal arteritis, erythrocyte sedimentation rate, 656t
 for urinary tract infection, women, urinalysis, 689t
 for vaginitis, microscopic tests, 699t, 700t, 707t
 labral (shoulder) tears, 589-591
 findings of, 589
 guidelines, evidence from, 590
 likelihood ratio for, 591
 literature search, 589
 original publication data, improvements in, 590
 physical examination tests, 580t
 precision of
 laxity maneuvers, 590t
 provocation maneuvers, 590t
 prior probability, 591
 reference standard, changes in, 590, 591
 labyrinthitis, 711
 Lachman test, 360, 360f, 369
 sensitivity, specificity, or likelihood ratio
 for knee ligament and meniscus injury, 359f, 370t
 LACS. *See* lacunar infarction syndrome
 lacunar infarction syndrome (LACS), 634
 laparoscopy
 for appendicitis, 54

- LAPSS. *See* Los Angeles Prehospital Stroke Scale
- laryngeal height
- sensitivity, specificity, or likelihood ratio
 - in obstructive airways disease, 161t, 162t
- laryngitis, 540, 541f
- laryngotracheobronchitis, 540, 541f
- late penicillin hypersensitivity, 517-518
- lateral collateral ligament (LCL), 358
- lateral pivot shift test, 360f, 361
- sensitivity, specificity, or likelihood ratio
 - knee ligaments and menisci, 359f
- latex agglutination assays, 230
- LAW (lymphocyte count, albumin, percentage weight loss)
- discriminant model, for adult malnutrition, 381
- laxity tests
- for shoulder instability, 579, 580t
- LCL. *See* lateral collateral ligament
- lead pipe rigidity. *See* rigidity
- left-sided heart failure, in adults, 183
- clinical findings for
 - detection of, 186-187, 187t
 - precision of, 189
 - definition, 184
 - ejection fraction, detection of, 187-189, 188t
 - methods
 - data abstraction, 184-186
 - literature search, 184
 - pathophysiology of, 184
 - signs, elicitation of
 - apical impulse, 190
 - jugular veins, 190
 - radiographic cardiomegaly, 190
 - radiographic redistribution, 190
 - third heart sound, 190
 - vital signs, 189-190
- left ventricular dysfunction
- findings of, 209
 - guidelines, evidence from, 211-212
 - likelihood ratio, 213
 - literature review, results of, 211
 - literature search, 209
 - original publication data, improvements in, 211
 - prior probability, 213
 - reference standard, changes in, 211, 213
 - systolic dysfunction
 - diagnosis of, 210-211
 - echocardiograms, 210-211
 - postmyocardial infarction, 210
- and diastolic dysfunction, difference between, 211
- left ventricular hypertrophy, 189
- Legionella*, 344
- Legionella monocytogenes*, 400
- levels of evidence, 15t
- Levine grading system. *See* murmur intensity
- levodopa, 505, 506, 508
- Li-Fraumeni syndrome, 275
- likelihood ratio (LR) test, 218, 529, 617
- for abdominal aortic aneurysm, 27t
 - for acute cholecystitis, 147
 - for acute otitis media, 496, 501, 503t
 - for aortic regurgitation, 429t, 430t
 - for aortic stenosis, 446t
 - for appendicitis, 63
 - for β-hemolytic streptococcal pharyngitis, 625
 - calculation of, 8f
 - for cancer, 276
 - for carpal tunnel syndrome, 124
 - for central venous pressure, 135
 - for chest pain protocol, acute cardiac ischemia, 473t
 - for chest pain radiation, 471-472, 472t
 - for clinical assessment, 9-11
 - of deep vein thrombosis, 229, 231
 - for community-acquired pneumonia, 530t, 531, 531t, 537
 - for deep vein thrombosis, 246
 - for dehydration, 341
 - for goiter, 287
 - for home pregnancy test, 559, 560t
 - for hypovolemia, 327t
 - in ICU patients, 326t
 - for influenza, 356
 - for labral tears, 591
 - for left ventricular dysfunction, 213
 - for major depression, 262
 - for malnutrition, 380t, 381t
 - for medication nonadherence, 182t
 - for meningitis in adults, 404t, 406
 - for meniscal and ligamentous knee injuries, 370
 - for mitral regurgitation, 447t
 - for myasthenia gravis, 460t
 - for myocardial infarction, 476
 - for obstructive airways disease, 162
 - for osteoporosis, 491
 - for Parkinson disease, 514
 - for pediatric pneumonia, 550
 - for penicillin allergy, 523
 - for perimenopausal, 416t, 417t
 - for Phalen sign, 124t
- for pregnancy, 560
- for pulmonary embolism, 572t, 575
- for reference standard tests, 356t
- for renal artery stenosis, 37
- for shoulder instability, 591
- for sinusitis, 603
- for splenomegaly, 613
- for stroke, 629, 641
- for temporal arteritis, 648t, 649t, 654t, 656t
- for thoracic aortic dissection, 667t, 673, 673t
- for Tinel sign, 124t
- for urinary tract infection, 681t
- for vaginal complaints, 707
- for valvular heart disease, 446t
- for vertigo, 717
- limb weakness, 451
- liver, physical examination of, 289-296
- auscultation of, 291-292
 - examination of, 290-291
 - inspection of, 291
 - palpable liver edge, 292-293
 - physical findings of, 295
 - pulsatile liver edge, 293
 - topography, 289-290
 - vertical liver span, assessing, 293-295
- liver disease
- ascites, 65
 - liver edge. *See* palpation
 - liver span, normal, 291t
- load and shift anterior test
- for shoulder instability, 585t
- load and shift posterior test
- for shoulder instability, 585t
- logistic analysis, 14
- Los Angeles Prehospital Stroke Scale (LAPSS), 628, 630-631, 633
- low back pain
- anatomic/physiologic origins, 75
 - causes, 75
 - evidence from guidelines, 84
 - findings of, 83
 - history, 81
 - literature review, 84
 - literature search, 83
 - neurologic compromise, 77-80
 - cauda equina syndrome, 80
 - imaging tests, indications for, 80
 - lumbar disk herniations, 77-78
 - motor, reflex, and sensory dysfunction, assessment of, 78-80
 - spinal stenosis, 80
- original publication data, improvements in, 83
- physical examination, 81
- prevalence of diseases, 75-76

- low back pain (*Continued*)
 prior probability, 86
 reference standard, changes in, 84, 86
 social or psychological distress, 80-81
 systemic disease
 ankylosing spondylitis, 77
 cancer, 76
 compression fractures, 77
 spinal infections, 77
 spine range-of-motion measures, 77
 lower-extremity dermatomes, 79/
 lower respiratory tract illness (LRI), 539, 540. *See also* pneumonia
 LR. *See* likelihood ratio test
 LRI. *See* lower respiratory tract illness
 lumbar disk herniation, 77-78
 crossed straight-leg raise test, accuracy, 84*t*
 ipsilateral straight-leg raise test, accuracy, 84*t*
 physical examination, accuracy
 sciatica, patients with, 79/
 lumbar spine
 low back pain, 75
 lung scanning, 562
 lymphadenopathy
 sensitivity, specificity, or likelihood ratio
 in streptococcal pharyngitis, anterior cervical, 618*t*
 lymphocyte count, albumin, percentage weight loss (LAW)
 discriminant model, for adult malnutrition, 381
 lysosomal enzyme, 616
- M**
- major depression, 248
 likelihood ratio, 262
 malaise
 sensitivity, specificity, or likelihood ratio
 influenza, 356*t*
 male. *See* sex
 malignancy, 249
 ascites, 65
 malignant melanoma
 ABCD(E) criteria
 likelihood ratio, 393*t*
 multivariate findings, 392*t*
 univariate findings, 392*t*
 detection and prognosis, 383
 epidemiology, 383
 evidence from guidelines, 392-393
 findings of, 392
 literature review, 392
 literature search, 391
 methods
 search strategy and quality filter, 385
 original publication data, improvements in, 392
 prior probability, 393
 reference standard, changes in, 392, 393
 skin examination
 accuracy of
 ABCD(E) checklist, 385*t*, 385-387
 for detecting presence or absence, 387*t*, 387-388
 revised 7-point checklist, 386, 386*t*
 checklists as diagnostic aid, 384
 criterion standard for diagnosis, 385
 historical feature assessment, 384
 physical examination technique, 384
 precision of, 385
 signs and symptoms, 383-384
 skin type risk factors, 393*t*
 malignant neoplasm, 76. *See also* cancer
 malnutrition, 371. *See also* nutritional status assessment
 evidence from guidelines, 381
 findings of, 379
 likelihood ratio, 382
 of findings combinations, 380*t*
 of low albumin, 380*t*
 literature review, 380-381
 literature search, 379
 multivariate findings, 380*t*
 original publication data, improvements in, 379-380
 prior probability, 381
 reference standard, changes in, 380, 381
 subjective global assessment, 379-380
 Malnutrition Screening Tool, 381*t*
 Mammacare Method, 94, 95
 mammography, 89-91
 MANTRELS mnemonic, 61, 62*t*. *See also* Alvarado clinical decision rule
 marasmus, 372
 Marfan syndrome, 663
 in thoracic aortic dissection, 664*t*
 marginal cross-products, 3
 Massachusetts Women's Health Study, 408
 MAST. *See* Michigan Alcoholism Screening Test
 match test, 151-152
 maxillary sinus, 594, 595, 595*f*, 597
 surface palpation for, 596*f*
 transillumination of, 596*f*
 McIsaac clinical prediction rule for sore throat, 620, 620*f*
 MCL. *See* medial collateral ligament
 McMurray test, 360*f*, 361, 369
 sensitivity, specificity, or likelihood ratio
 knee ligament and meniscus injury, 359*f*, 370*t*
 MDM. *See* minor determinant mixture
 medial collateral ligament (MCL), 358, 359, 361
 medial-lateral grind test, 361
 Medical Research Council Thrombosis Prevention Trial, 26
 medication adherence, assessing, 179-182, 182*t*
 findings of, 179
 guidelines, changes in, 181
 literature review, results of, 180-181
 literature search, 179
 Morisky questions, 182*t*
 original publication data, improvements in, 180
 pill count, 176*t*
 prior probability, 182
 reference standard, changes in, 180, 182
 Medication Adherence Self-Report Inventory, 180
 Medication Event Monitoring System (MEMS) caps, 180
 medication response
 sensitivity, specificity, or likelihood ratio
 to anticholinesterase, for myasthenia gravis, 460*t*
 to decongestants, for sinusitis, 603*t*
 to levodopa, for Parkinsonism, 510*t*
 to penicillin skin test, for penicillin allergy, 525*t*
 melanocyte, 383-384
 MEMS. *See* Medication Event Monitoring System caps
 Ménière disease, 316, 711
 meningitis, 404*t*
 clinical examination, 395-396
 clinical history
 accuracy of, 398
 sensitivity of, 398*t*
 evidence from guidelines, 405
 findings of, 403
 likelihood ratios, 404*t*, 406
 literature review, 404
 prospective study, 404-405
 retrospective study, 404

- literature search, 403
 methods
 data analysis, 398
 literature search and selection, 396-397
 study characteristics, 397-398
 original publication data,
 improvements in, 403-404
 pathophysiology of, 396
 physical examination
 accuracy of, 398, 399-400
 sensitivity of, 399t
 prior probability, 406
 reference standard, changes in, 404, 406
 sensitivity of findings, 404t
 signs and symptoms, 396
 precision of, 398
 menopause. *See* perimenopause
 meta-analysis
 of clinical examination, 12-13
 methacholine, 150
 MI. *See* myocardial infarction
 Michigan Alcoholism Screening Test (MAST), 41, 42, 42t, 44. *See also* Brief Michigan Alcoholism Screening Test; Short Michigan Alcoholism Screening Test
 accuracy of, 42
 questionnaire, 49
 reliability of, 42
 sensitivity, specificity, or likelihood ratio
 in problem alcohol drinking, 41
 micrographia
 sensitivity, specificity, or likelihood ratio
 in Parkinsonism, 509t
 Middleton hooking maneuver
 sensitivity, specificity, or likelihood ratio
 in splenomegaly, 612t
 Mini MagLite, 597
 Mini Nutritional Assessment (MNA), 380-381
 Minnesota Multiphasic Personality Inventory, 80
 minor depression, 248
 minor determinant mixture (MDM), 520
 mitral regurgitation (MR), 438-439, 445
 clinical examination
 accuracy of, 439t
 and mitral valve prolapse, 446
 physical examination
 accuracy of, 445t
 mitral stenosis
 and pulmonic regurgitation, 423, 425
 mitral valve prolapse (MVP), 439-440, 445
 clinical examination
 accuracy of, 440t
 and mitral regurgitation, 446
 MNA. *See* Mini Nutritional Assessment
 moderate drinking, 48
 Modigliani syndrome, 280
 monobactams, 515, 518
 monofilament testing. *See* sensory change
 mood, 249
 Moraxella catarrhalis, 494
 Morisky measure, 180
 sensitivity, specificity, or likelihood ratio
 for medication adherence, 182t
 morning sickness, 552, 554, 555
 for pregnancy, 560
 Movement Disorder Society, 507
 MR. *See* mitral regurgitation
 multiple nevi, 384
 multivariate analysis, 14
 murmur
 intensity, 420
 sensitivity, specificity, or likelihood ratio
 in aortic regurgitation
 clinical impression, 429t
 intensity, 430t
 significant vs insignificant
 systolic murmurs, 446t
 in thoracic aortic dissection
 diastolic, 665t, 666t, 673t
 typical murmur, 430t
 in aortic stenosis
 carotid pulse, 446t
 intensity, 446t
 location of murmur, 446t
 radiation to carotids, 438t
 S2 (second heart sound), 446t
 systolic murmur, 446t
 timing, 438t
 in hypertrophic cardiomyopathy
 change with maneuvers, 439
 in mitral regurgitation
 during myocardial infarction, 439t
 intensity, 447t
 location, 439t
 timing, 439t
 in tricuspid regurgitation
 change with abdominal pressure, 439
 change with inspiration, 439
 Murphy sign, 138, 145, 146
 sensitivity, specificity, or likelihood ratio
 in acute cholecystitis, 147t
 muscle wasting, subjective global assessment (SGA), 374
 MVP. *See* mitral valve prolapse
 myalgia
 sensitivity, specificity, or likelihood ratio
 in influenza, 348t
 in pneumonia, adult 536t
 myasthenia gravis, 449-456, 459
 acetylcholine receptor antibody-positive myasthenia gravis, 449, 450f
 anticholinesterase tests, 451-453
 likelihood ratio of, 460t
 office tests, accuracy of, 454-455
 prior probability, 460
 reference standard tests, 460
 search strategy and quality review, 453
 statistical methods, 454
 symptoms and signs of
 accuracy of, 454
 anatomical and physiological origins of, 450-451
 elicitation of, 451
 Mycoplasma pneumoniae, 344
 myeloproliferative syndrome, 612
 Myerson sign, 508
 myocardial infarction (MI), 195, 461-469. *See also* cardiac ischemic chest pain; chest pain
 accuracy of, 466
 electrocardiogram, 468
 medical history, 467-468
 physical examination, 467-468
 acute cardiac ischemia, multivariate findings for, 473-474
 acute chest pain, diagnosis of, 462-463, 463f
 cardiac and noncardiac conditions, 463-464
 clinical findings of, 468-469
 clinical prediction rules, 468-469
 diagnostic criteria, 476
 findings of, 471
 guidelines, evidence from, 474-475
 likelihood ratio of, 476
 literature review, results of, 472-473
 literature search, 471
 mechanism of, 463
 methods
 analysis, 465
 precision and accuracy, test criteria for, 464

- myocardial infarction (MI) (*Continued*)
 quality assessment, 464-465
 search strategy, 464
 selection of articles, 464
 multivariate findings for, 476t
 original publication data,
 improvements in, 472
 precision of
 electrocardiogram, 465-466
 medical history, 465
 physical examination, 465
 pretest probability, 469
 prior probability, 476
 reference standard, changes in, 472,
 476
 symptoms and signs, 463
 univariate findings for, 476t
- myoclonus, 217, 221, 222
 sensitivity, specificity, or likelihood ratio
 in coma, 226t
- N**
- nafcillin, 519
 nailfold angles, 164, 166
 for clubbing, 165f
 nasal congestion
 differential diagnosis of, 594t
 sensitivity, specificity, or likelihood ratio
 in influenza, 356t
 in otitis media, child, 497t
 in pneumonia, adult, 536t
 in sinusitis, purulent, 603t
 nasal flaring, in pneumonia, infant and child, 544t
 nasal turbinates, 594, 595f
- NASCET. *See* North American Symptomatic Carotid Endarterectomy Trial
- National Ambulatory Medical Care Survey, 493, 615
- National Center for Health Statistics, 53
- National Health Interview Survey data, 99
- National Heart, Lung, and Blood Institute, 161
- National Institute of Neurological Disorders and Stroke, 634
- National Institute on Alcohol Abuse and Alcoholism, 48
- National Institutes of Health Stroke Scale (NIHSS), 630, 630t, 633, 637, 637t
 reliability of, 634, 635t
- National Osteoporosis Foundation, 478
- National Program of Cancer Registries, 267
- National Society of Genetic Counselors, 267
- nausea. *See* nausea and vomiting
- nausea and vomiting
 sensitivity, specificity, or likelihood ratio
 in acute cholecystitis, 140t
 in appendicitis, adult, 57t
 in meningitis, adult, 406t
 in myocardial infarction, 576t
- NBSS. *See* Canadian National Breast Screening Study
- neck
 carotid artery cause, for bruits, 103-104
 stiffness, 396, 399
- negative likelihood ratio (LR-), 19, 57, 197, 236, 454, 554, 617
 in clinical examination, 9, 12
 median, 253
 for test of Speed, 589
 for test of Yergason, 589
- negative predictive value, 4, 5
 calculation of, 8f
- Neisseria meningitidis*, 400
- neostigmine bromide, 452
- nephrotic syndrome
 ascites, 65
- nervous tension, perimenopausal, 409
- neurologic compromise, low back pain, 77-80
 cauda equina syndrome, 80
 imaging tests, indications for, 80
 lumbar disk herniations, 77-78
 motor, reflex, and sensory dysfunction, assessment of, 78-80
 spinal stenosis, 80
- neurologic deficit
 sensitivity, specificity, or likelihood ratio
 in stroke, 641t
 in thoracic aortic dissection, 673t
- nevi
 atypical (dysplastic), 384
 clinical assessment
 for spider nevi, precision of, 3f
 multiple nevi, 384
- night sweats, perimenopausal, 409
- NIHSS. *See* National Institutes of Health Stroke Scale
- nitric oxide, 616
- nitroglycerin, 473
- Nixon method, 606-607, 607f, 608
 sensitivity, specificity, or likelihood ratio
- in splenomegaly, 612t
- nomogram, 7
- noncompliance. *See also* compliance measurement, 174-175
 methods, 175
 nature of, 174
- normovolemic
 phlebotomy study, 316t
 postural vital signs, 318t
- North American Symptomatic Carotid Endarterectomy Trial (NASCET), 105, 108
- Nugent criteria, 694t
 sensitivity, specificity, or likelihood ratio
 for vaginitis, 707t, 694t
- nutrition-associated complications, 372
- nutritional status assessment, 371-372.
See also malnutrition
 accuracy of, 375-376
 anatomic and physiologic origin, 372
 components of, 371-372
 precision of, 375
 subjective global assessment, features, 373t
 dietary intake change, 373
 functional capacity, 374-375
 gastrointestinal symptoms, 374
 weight change, 372, 373
 symptoms and signs, 376-377
- nystagmus
 origin of, 711
- O**
- obesity, 20
- observed agreement. *See* agreement
- obstructive airways disease, 159-162
 findings of, 159-160
 guidelines, evidence from, 161
 likelihood ratio, 162
 literature review, results of, 160-161
 literature search, 159
 original data publication, improvements in, 160
 prior probability, 162
 reference standard, changes in, 160, 162
- obstructive lung disease, 528
- obturator sign
 in appendicitis, 55
 examination for, 55f
- occult vertebral fracture, 478
 hand grip strength test for, 479, 485
 rib-pelvis distance test for, 479, 485
 skinfold thickness test for, 479, 485

tooth count, 485-486
 wall-occiput distance test for, 479, 485
 ocular myasthenia, 450
 oculocephalic reflex. *See* reflexes
 odds ratio (OR), 9, 35, 228, 237
 diagnostic, 12
 odor
 sensitivity, specificity, or likelihood ratio
 in vaginitis, 707t
 office blood pressure
 factors, affecting, 303-304
 vs usual blood pressure, 305
 OME. *See* otitis media with effusion
 opening snap
 sensitivity, specificity, or likelihood ratio
 murmur, diastolic, 421
 OR. *See* odds ratio
 ORAI. *See* osteoporosis risk assessment instrument
 orbicularis oculi weakness, 451, 452f
 orthopnea, 153. *See also* dyspnea
 orthostatic hypotension, 320t
 Osler sign, 304-305
 OST. *See* osteoporosis self-assessment screening tool
 osteomyelitis, 594
 osteoporosis, 477-486, 489-491
 arm span-height difference, 479, 482-483
 definition of, 478
 diagnostic accuracy, 480, 483t, 484t
 elicitation of, 478-480
 findings of, 489
 hand grip strength test for, 479, 485
 height loss, 482
 likelihood ratio of, 491
 literature review, results of, 490
 multivariate findings for, 490t
 univariate findings for, 490t
 literature search, 489
 methods
 data analysis, 480
 quality assessment, of articles, 480
 original publication data,
 improvements in, 490
 pathophysiology of, 478
 precision of, 480, 482t
 prevalence of, 478, 491t
 prior probability, 491
 reference standard, changes in, 490, 491
 rib-pelvis distance test for, 479, 485
 skinfold thickness test for, 479, 485
 study characteristics, 480, 481t, 482t
 tooth count, 485-486

wall-occiput distance test for, 479, 485
 weight, 483-485
 osteoporosis risk assessment instrument (ORAI), 490, 491t
 osteoporosis self-assessment screening tool (OST), 489, 490, 491t
 otitis media
 diagnostic criteria, in children, 494
 otitis media with effusion (OME), 495
 and acute otitis media, distinguishing between, 493
 otolaryngologist, 496, 596, 597
 otosclerosis, 711
 ovarian cancer, 265, 266, 267
 ovarian carcinoma, 72
 Ovid MEDLINE, 47
 Oxfordshire Classification of Subtypes of Cerebral Infarction
 stroke, symptoms, 634
 oxygen free radicals, 616
 oxymetazoline hydrochloride, 595

P

pachydermoperiostosis, 163
 PACS. *See* partial anterior circulation infarction syndrome
 pain
 sensitivity, specificity, or likelihood ratio
 in acute cholecystitis
 guarding, 140t
 rebound, 140t
 rectal, 140t
 right upper quadrant, 147t
 rigidity, 140t
 in appendicitis, adult
 guarding, 57t
 migration, 57t
 pain before vomiting, 57t
 rebound tenderness, 57t
 rectal tenderness, 57t
 right lower quadrant, 57t
 rigidity, 57t
 in back pain
 duration, 76t
 positional, 76t
 in cancer-induced back pain,
 nocturnal, 86t
 in coma
 motor response, 225t
 withdrawal response, 225t
 in influenza, 348t-349t
 nasal congestion, 348t
 pharyngitis, 349t
 in knee injury, joint line tenderness, 370t

in meningitis, adult, headache, 406t
 in myocardial infarction
 chest wall, 467t
 pleurisy, 467t
 positional, 467t
 radiation to the arms, 576t
 in otitis media, child, ear, 503t
 in temporal arteritis
 headache, 654t
 jaw, 656t
 scalp, 654t
 in thoracic aortic dissection, 673t
 migratory, 673t
 sudden onset, 673t
 “tearing” or “ripping,” 673t
 in urinary tract infection, women, 679t
 back pain, 680t
 flank pain, 679t
 lower abdominal pain, 679t
 pain provocation test of Mimori
 for labral tears, 582f, 586t
 palpable expansile tumor, 18
 palpation, 291, 293, 294, 295
 for airflow limitation, 151, 153, 154
 of clubbing, 165-166
 sensitivity, specificity, or likelihood ratio
 abdomen, for abdominal aortic aneurysm, 27t
 of liver, for hepatomegaly, 299t
 of spleen, for splenomegaly, 612t, 613t
 of temporal artery, for temporal arteritis, 649t
 of thyroid, for goiter, 286t, 287t
 of spleen, 607-608, 609, 609t
 paradoxic ptosis. *See* curtain sign
 parainfluenza, 344
 paranasal sinuses
 coronal view of, 595f
 sagittal view of, 594f
 transillumination of, 596-597
 Waters view of, 594
 parental suspicion of otitis media, 503t
 paresthesia
 sensitivity, specificity, or likelihood ratio
 in carpal tunnel syndrome, 115t
 Parkinson disease (PD), 505-511, 513-514
 accuracy of, 510
 findings of, 513
 guidelines, evidence from, 513
 likelihood ratio of, 514
 literature review, results of, 513
 literature search, 513

- Parkinson disease (PD) (*Continued*)
 - methods, 509
 - original publication data,
 - improvements in, 513
 - and parkinsonism, distinguishing between, 506
 - pathophysiologic characteristics of, 506
 - precision of, 510
 - prior probability, 514
 - quality of evidence, 508*t*, 509-510
 - reference standard, changes in, 513
 - signs of, 506-507, 510*t*
 - elicitation of, 507-508
 - symptoms of, 506-507, 509*t*
 - parkinsonian facies, 509
 - paroxysmal nocturnal dyspnea. *See* dyspnea
 - parsimonious clinical examination, 13-14
 - parsimony, 13
 - partial anterior circulation infarction syndrome (PACS), 634
 - Pastia sign, 617
 - patella, 359-360
 - patella reflex, 79
 - pathophysiology
 - of appendicitis, 54
 - of community-acquired pneumonia, 528
 - patient, 1-2
 - alcoholic, 1
 - ascites, 1
 - patient-generated subjective global assessment (PG-SGA), 380
 - Patient Health Questionnaire (PHQ-9), 259, 260, 261, 263. *See also* clinical prediction rules and scores
 - patient's medical history, information in, 11
 - PCL. *See* posterior cruciate ligament
 - PD. *See* Parkinson disease
 - PDR. *See* phalangeal depth ratio
 - PE. *See* pulmonary embolism
 - peak expiratory flow, 155
 - pediatric pneumonia
 - likelihood ratio for, 550
 - multivariate findings for, 548
 - univariate findings for, 548, 549*t*
 - Pedigree Standardization Task Force, 267
 - peek sign
 - sensitivity, specificity, or likelihood ratio
 - myasthenia gravis, 460*t*
 - pelvic appendicitis, 54
 - penicillin, 174
 - penicillin allergy, 515-521, 523-525
 - β -lactam antibiotics, cross-reactivity with, 518
 - clinical history, 518-519
 - accuracy of, 519
 - findings of, 523
 - guidelines, evidence from, 524
 - how to take a history for, 518
 - hypersensitivity reactions,
 - classification of, 516-518, 517*t*
 - immediate reactions, 516-517
 - late reactions, 517-518
 - likelihood ratio of, 525
 - literature review, results of
 - univariate findings for, 524
 - literature search, 523
 - methods, 516
 - original publication data,
 - improvements in, 524
 - prior probability, 525
 - reference standard, changes in, 524, 525
 - sensitivity, specificity, or likelihood ratio
 - history of, 525*t*
 - skin testing, 519-520
 - limitations of, 520-521
 - penicillin G, 520
 - penicillin skin test, 519-520, 523, 524, 525*t*
 - limitations of, 520-521
 - percussion, 290, 291, 294, 295
 - advantages of, 295
 - for airflow limitation, 151, 153, 154
 - measurement of, 294-295
 - sensitivity, specificity, or likelihood ratio
 - in ascites
 - flank dullness, 69*t*
 - fluid wave, 69*t*, 73*t*
 - shifting dullness, 69*t*, 73*t*
 - in chronic obstructive airways disease
 - cardiac dullness, 154*t*
 - chest hyperresonance, 154*t*
 - for splenomegaly, of spleen, 612*t*, 613*t*
 - in pneumonia, adult
 - dullness of the lungs, 531*t*
 - percussion methods, for splenomegaly
 - Castell method, 607, 607*f*, 608
 - Nixon method, 606-607, 607*f*, 608
 - Traube space, 607, 607*f*
 - percussive span technique, 295
 - perimenopause, 407, 416*t*
 - definition, 408
 - estimate pretest probability of, 408
 - evaluation
 - family and medical history
 - age of mother's menopause, 409
 - cigarette use, 409
 - hysterectomy status, 410
 - laboratory tests
 - estradiol, 410
 - follicle-stimulating hormone, 410
 - inhibins, 410
 - physical signs
 - maturation index, 410
 - skin thickness, 410
 - vaginal pH, 410
 - self-assessment, 408-413
 - symptoms, 409
 - depressed mood, 409
 - hot flashes, 409
 - nervous tension and irritability, 409
 - night sweats, 409
 - urinary incontinence, 409
 - vaginal dryness, 409
 - variable sexual interest, 409
 - evidence from guidelines, 416
 - findings of, 415
 - likelihood ratio, 416*t*, 417*t*
 - literature review, 415-416
 - literature search, 415
 - methods
 - search strategy and quality review, 410-411
 - statistical methods, 412
 - original publication data,
 - improvements in, 415
 - physiology, 408
 - prior probability, 417
 - reference standard, changes in, 415
 - peripheral edema, 72
 - peripheral hemodynamic signs, 421-422
 - peritoneal fluid, 71, 72
 - perumbilical bruits, 30
 - petechiae, 617
 - PG-SGA. *See* patient-generated subjective global assessment
 - phalangeal depth ratio (PDR), 164-165, 166
 - Phalen sign, 112*t*, 114
 - likelihood ratio, 124*t*
 - sensitivity, specificity, or likelihood ratio
 - in carpal tunnel syndrome, 124*t*
 - pharyngeal exudate, in streptococcal pharyngitis, 618*t*
 - pharyngeal weakness, 451
 - pharyngitis, 615, 623. *See also* pain differential diagnosis of, 616*t*
 - streptococcal, 625

PHQ. *See PRIME-MD Patient Health Questionnaire*

PHQ-9. *See Patient Health Questionnaire*

physiologic origin, of abdominal bruit, 29

physiologic tremor, 506, 514

pigmented skin lesion. *See ABCD(E) criteria*

pill count, in medication adherence, 176^t

pill-rolling tremor, 506

PIOPED. *See Prospective Investigation of Pulmonary Embolism Diagnosis study*

PISA-PED. *See Prospective Investigative Study of Acute Pulmonary Embolism Diagnosis*

plain abdominal radiographs
for appendicitis, 54

plain film radiographs
for sinusitis
likelihood ratio, 603

plaster casts, of fingers, 171

pleuritic chest pain, 215, 225

pneumatic otoscopy, 493, 494, 495, 496, 499

pneumococcal pneumonia, 527

pneumonia. *See also acute respiratory illness; lower respiratory tract illness*
community-acquired, 527-533

pneumonia, in infant and child, 539-545, 547-550. *See also tachypnea*
anatomy and pathophysiology of, 540-541

bacterial, 540

findings of, 547

guidelines, evidence from, 549

literature search, 547

methods, 539-540

original publication data, improvements in, 548

pediatric pneumonia
likelihood ratio test for, 550

multivariate findings for, 548

univariate findings for, 548, 549^t

prior probability, 550

reference standard, changes in, 540, 548, 550

symptoms and signs
accuracy of, 543-545

elicitation of, 541-542

precision of, 542-543

pneumonia score, 548

Pneumonia Severity Index, 536

point-of-care testing, 623-624, 625^t

polyps, 594, 595

popliteal-brachial gradient, in aortic regurgitation, 425^t

positive likelihood ratio (LR+), 19, 56, 58, 197, 236, 454, 513, 554, 617
in clinical examination, 9, 12
median, 253
for test of Speed, 589
for test of Yergason, 589

positive predictive value, 4
calculation of, 8^f

posterior circulation infarction syndrome (POCS), 634

posterior cruciate ligament (PCL), 358, 359, 361, 362
physical examination
accuracy of, 364^t

posterior drawer test, 361

posterior probability, 9

postmyocardial infarction, 210

postphlebitic syndrome, 235

posttest probability, 4, 7
calculation of, 8^f

postural tachycardia. *See tachycardia*

postural vital signs, 316

PR. *See pulmonic regurgitation*

precision. *See also κ*
calculation of, 8^f
of clinical examination, 1, 3-4, 9
“good” symptom or sign, 11-12
for left-sided heart failure, 189
likelihood ratio, 9-11
meta-analysis, 12-13
pretest probability, 11
“sensitivity-only” studies, 13

pregnancy, 551-557, 559-560
guidelines, evidence from, 559
home pregnancy tests, accuracy of, 555-556
likelihood ratio test for, 560
literature review, results of, 559
literature search, 559
methods
search strategy, 553-554

original publication data, improvements in, 559

patient history, accuracy of, 554-555

physical examination, accuracy of, 556-557

prior probability, 560

reference standard, changes in, 559, 560

signs and symptoms, 554-555
elicitation of
medical history, 552

physical examination, 552-553
reference standard for, 553

during first trimester
anatomic and physiologic origins, 552

uterine height, 553^f

pregnant women
problems in
at-risk drinking, 44-45

preoperative carotid bruit, 105

pressure provocation test, 112^t

pretest probability, 5, 7, 11
calculation of, 8^f

prevalence, calculation of, 8^f

Primary Care Evaluation of Mental Disorders (PRIME-MD), 250, 251, 252, 259, 260, 261, 263

PRIME-MD Patient Health Questionnaire (PHQ), 250, 251, 252.

PRIME-MD. *See Primary Care Evaluation of Mental Disorders; clinical prediction rules and scores*

problem drinking, 47

Prospective Investigation of Pulmonary Embolism Diagnosis (PIOPED) study, 562, 563, 564

Prospective Investigative Study of Acute Pulmonary Embolism Diagnosis (PISA-PED) study, 563, 564, 565, 566, 572

prostaglandins, 616

prostate cancer, 265

protein-energy malnutrition, 372

provocation test
for labral tears, 579, 580^t
for shoulder instability, 579, 580^t

pseudohypertension, 304

Psoas sign, of appendicitis, 55
sensitivity, specificity, or likelihood ratio
for appendicitis, 57^t

psychogenic dizziness, 711

ptosis, 451, 455

puddle sign, 66
sensitivity, specificity, or likelihood ratio
for ascites, 69^t

pulmonary crackles. *See rales*

pulmonary embolism (PE), 227, 235, 561-569, 571-575
clinical examination, precision of, 566

clinical gestalt, 563-564
negative result for, 563
positive result for, 563

clinical prediction rules, 564-565
accuracy of, 565^t, 568^t

pulmonary embolism (PE) (*Continued*)
 components of, 566
 validation of, 565–566
 findings of, 571
 guidelines, evidence from, 574
 likelihood ratio test for, 572*t*, 575
 literature review, results of, 572–574
 literature search, 571
 methods
 data analysis, 563
 data sources, 562
 study selection and data extraction,
 562–563
 original publication data,
 improvements in, 571–572
 pretest probability, accuracy of, 564*t*
 prior probability, 575
 reference standard, changes in
 computed tomography (CT)
 angiography, 572
 D-dimer assay, 572
 enzyme-linked immunosorbent
 assay, 572
 pulmonary fibrosis, 528
 pulmonic regurgitation (PR), 421
 and mitral stenosis, 423, 425
 pulse deficit (arms), in thoracic aortic
 dissection, 673*t*
 pulse pressure, 421
 in aortic regurgitation, 425*t*
 pulsus paradoxus
 sensitivity, specificity, or likelihood
 ratio
 in chronic obstructive airways
 disease, 154*t*
 pupillary response. *See* reflexes
 pyridostigmine bromide, 452

Q

quadriceps weakness, 79
 sensitivity, specificity, or likelihood
 ratio
 in sciatica, 79*t*
 QUADAS. *See* Quality Assessment of
 Diagnostic Accuracy Studies
 checklist
 quality, 15–16
 Quality Assessment of Diagnostic
 Accuracy Studies (QUADAS)
 checklist, 583
 Quebec Task Force on Spinal Disorders,
 80
 questionnaire
 for depression
 Beck Depression Inventory (BDI),
 250

Center for Epidemiologic Studies
 Depression (CES-D), 250
 Duke Anxiety and Depression Scale
 (DADS), 250
 Geriatric Depression Scale (GDS),
 250
 Patient Health Questionnaire
 (PHQ-9), 263*t*
 Primary Care Evaluation of Mental
 Disorders (PRIME-MD), 250
 PRIME-MD, 263*t*
 PRIME-MD Patient Health
 Questionnaire (PHQ), 250
 Zung Self-Rating Depression Scale
 (SDS), 250
 for malnutrition, adult, 380*t*
 for medication adherence
 Morisky questions, 182*t*
 for problem alcohol drinking
 Alcohol Use Disorders
 Identification Test (AUDIT),
 51*t*
 Alcohol Use Disorders
 Identification Test,
 Consumption Questions
 (AUDIT-C), 51*t*
 CAGE questions, 52*t*
 T-ACE questions, 52*t*
 TWEAK questions, 52*t*
 quiver eye movements, in myasthenia
 gravis, 454*t*

R

radiographic cardiomegaly, 187, 189,
 190
 radiographic findings
 sensitivity, specificity, or likelihood
 ratio
 in the breathless emergency patient,
 chest radiograph, 213*t*
 in left ventricular dysfunction,
 chest radiograph, 213*t*
 for sinusitis, sinus films, 603*t*
 radiographic redistribution, 186, 187,
 190
 radiographic techniques
 for appendicitis, 54
 radioisotopic scintiscan
 for splenomegaly, 606
 radionuclide scanning
 for acute cholecystitis, 138
 rales
 sensitivity, specificity, or likelihood
 ratio
 in the breathless emergency patient,
 213*t*

in myocardial infarctions, 467*t*
 in pneumonia, adult, 536*t*
 in pneumonia, infant and child,
 550*t*
 Ramsay Hunt syndrome, 712
 random-effects measure, 13, 19
 random-effects model, 316
 randomized controlled trials, 88
 range of motion, 358
 rapid influenza test, 355
 likelihood ratio for, 356*t*
 reactive airway disease, 452
 readers' guides
 for diagnostic test, 3*t*
 rebound tenderness, 55. *See also* pain;
 guarding
 receiver operating characteristic (ROC)
 curve, 617
 recent weight gain. *See* weight gain
 rectal examination, for appendicitis, 55
 recurrent vestibulopathy, 711
 reduced muscle power, 451, 455
 reflexes
 sensitivity, specificity, or likelihood
 ratio
 Achilles tendon, in normal patients,
 84*t*
 corneal, in coma, 220*t*
 cough, in coma, 220*t*
 eye movements, in coma, 220*t*
 gag, in coma, 220*t*
 glabella tap, in Parkinsonism, 514*t*
 oculocephalic, in coma, 220*t*
 pupillary, in coma, 220*t*
 relocation test
 for labral tears, 586*t*
 for shoulder instability or labral tear,
 581*f*, 583, 585*t*, 589, 590, 591*t*
 renal artery stenosis, 35
 abdominal bruits in, 30
 likelihood ratio, 37
 multivariate findings for, 36
 prior probability, 37
 reference standard tests, 37
 univariate findings for, 36*t*
 renovascular hypertension
 abdominal auscultation in, accuracy
 of, 31
 evaluation of, abdominal bruits in,
 29–32
 prognosis of, 32
 reserpine, 249
 respiratory distress, 547
 respiratory illness. *See* community-
 acquired pneumonia;
 pneumonia
 respiratory rate, in children, 541
 respiratory syncytial virus (RSV), 549

rest test, 451, 455
 rest tremor, 506, 507
 retractions, chest
 sensitivity, specificity, or likelihood ratio
 in pneumonia, infant and child, 550t
 reverse workup bias, 520
 rheumatoid arthritis, 164
 rhinitis, 593. *See also* nasal congestion
 allergic, 594
 viral, 594
 rhinorrhea
 differential diagnosis of, 594t
 rhinosinusitis, 603. *See also* sinusitis
 rhinoviruses, 344
 rhonchi, 155
 sensitivity, specificity, or likelihood ratio
 in chronic obstructive airways disease, 154t
 rib-pelvis distance
 for occult lumbar vertebral fractures, test, 479, 485
 in osteoporosis, 483t
 rigidity, 55, 506, 507
 sensitivity, specificity, or likelihood ratio
 cog wheeling
 in Parkinson disease, 506, 514t
 ROC curve. *See* receiver operating characteristic curve
 Rovsing sign, of appendicitis, 55
 RSV. *See* respiratory syncytial virus
 ruptured abdominal aortic aneurysm
 abdominal palpation for, 19

S

S2 (second heart sound). *See* heart sounds
 S3 (third heart sound). *See* heart sounds
 S4 (fourth heart sound). *See* heart sounds
 SBP. *See* systolic blood pressure
 Scandinavian Neurological Stroke Scale, 634
 scarlet fever, 617
 scattergram, 250
 Schamroth sign, 165
 for clubbing, 165f
 Schober test, 77
 sciatica, 77-78
 physical examination accuracy
 for lumbar disk herniation among patients, 79t

SCID. *See* Structured Clinical Interview for *DSM-III-R*; Structured Clinical Interview for *DSM-IV-TR*
 scintigraphy
 for liver examination, 292, 295
 SCORE. *See* simple calculated osteoporosis risk estimate
 Scottish International Guidelines Network, 134
 scratch test
 sensitivity, specificity, or likelihood ratio
 for hepatomegaly, 300t
 SDDDS-PC. *See* Symptom Driven Diagnostic System for Primary Care
 SDS. *See* Zung Self-Rating Depression Scale
 seizure, 217, 221-222
 sensitivity, specificity, or likelihood ratio
 in coma, 226t
 self-administered medication therapy, 173
 self-diagnosis
 sensitivity, specificity, or likelihood ratio
 otitis media, parental suspicion, 503t
 pregnancy, suspicion of, 560
 urinary tract infection, women, 680t
 vaginal candidiasis, 696t
 sensitivity, calculation of, 8f
 "sensitivity-only" studies, 13
 sensory change
 sensitivity, specificity, or likelihood ratio
 in carpal tunnel syndrome, 115t
 in sciatica, 79t
 sex
 sensitivity, specificity, or likelihood ratio
 in ventricular dysfunction, 211t
 sexually transmitted diseases, 676
 SGA. *See* subjective global assessment shadowgrams, 171
 shadowgraph method, 166
 shifting dullness, 66, 67. *See also* percussion
 Short Michigan Alcoholism Screening Test (SMAST), 41
 shoulder instability, 577-587, 589-591
 anatomy of, 578-579, 578f
 clinical tests for, 579-581, 580t, 581f
 findings of, 589
 guidelines, evidence from, 590
 labral tears, 579
 clinical tests for, 579-581, 580t, 582f
 limitation of, 583
 physical examination, diagnostic accuracy of, 583, 586t
 likelihood ratio for, 591
 limitation of, 583, 583
 literature search, 589
 original publication data, improvements in, 590
 physical examination
 diagnostic accuracy of, 583, 585t
 tests for, 580t
 precision of
 laxity maneuvers, 590t
 provocation maneuvers, 590t
 prior probability, 591
 reference standard, changes in, 590, 591
 signed rank test, 455
 silicone models, 92, 94, 95
 simple calculated osteoporosis risk estimate (SCORE)
 questionnaire, 490, 491
 Simplified Medication Adherence Questionnaire, 180
 Simplified Wells scoring system, 571, 572, 574, 575
 for pulmonary embolus, 575t
 SimpliRed assay, 230
 Single Question (SQ), for depression 250, 251
 single-fiber electromyography
 for myasthenia gravis, 450
 single-leg sit-to-stand test, 84
 sinusitis, 593-598, 601-603. *See also* rhinosinusitis
 anatomy and pathophysiology of, 594
 differential diagnosis of, 594t
 findings of, 601
 guidelines, evidence from, 602
 likelihood ratio test for, 603
 literature search, 601
 original publication data, improvements in, 602
 paranasal sinuses
 coronal view of, 595f
 sagittal view of, 594f
 transillumination of, 596-597
 Waters view of, 594
 prior probability, 603
 reference standard for, 594
 changes in, 602, 603
 symptoms and signs
 accuracy of, 596-598
 elicitation of, 594-595
 precision of, 595-596
 univariate findings for literature review results, 602

- sit-to-stand test, 84
 sensitivity, specificity, or likelihood ratio
 back pain
 upper lumbar herniation, 86t
 disk herniation, 86t
 skin examination, for malignant melanoma
 accuracy of
 ABCD(E) checklist, 385t, 385-387
 for detecting presence or absence, 387t, 387-388
 revised 7-point checklist, 386, 386t
 checklists as diagnostic aid, 384
 criterion standard for diagnosis, 385
 historical feature assessment, 384
 physical examination technique, 384
 precision of, 385
 signs and symptoms, 383-384
 skin turgor, 319, 331
 sensitivity, specificity, or likelihood ratio
 in hypovolemia, child, 334t
 skinfold thickness test
 for occult vertebral fracture, 479, 485
 SLAP lesion. *See* superior labrum anterior posterior lesion
 sleep test, 451, 455
 sensitivity, specificity, or likelihood ratio
 for myasthenia gravis, 460t
 SLR. *See* straight-leg raising sign
 SMAST. *See* Short Michigan Alcoholism Screening Test
 smoking. *See* tobacco use
 sneezing
 sensitivity, specificity, or likelihood ratio
 in influenza, 356t
 spasticity, 506
 sensitivity, specificity, or likelihood ratio
 in Parkinsonism, 506
 specificity, calculation of, 8f
 spectrum bias, 141, 497
 speech
 sensitivity, specificity, or likelihood ratio
 in myasthenia gravis, unintelligible, 460t
 in Parkinsonism, soft voice, 510t
 in stroke, abnormal, 641t
 sphenoid sinuses, 594f, 595
 sphygmomanometers, 306
 spider nevi
 clinical examination
 precision of, 3f
 Spiegel criteria, 694t
 sensitivity, specificity, or likelihood ratio
 for vaginitis, 707t, 694t
 spinal compression fractures, 77
 spinal infections, 77
 spinal stenosis, 80
 spine range-of-motion measures, 77
 spiral computed tomography (CT)
 scanning
 for pulmonary embolism, 564
 spirometry, 149, 150, 154, 160
 spleen. *See also* splenomegaly
 palpation of, 607-608, 609, 609t
 size of, 605-606, 606f
 splenomegaly, 605-610, 611-613. *See also* spleen
 anatomic landmarks, 605
 clinical examination for consequences of, 606
 guidelines, 609t
 inspection, 606
 palpation, 607-608, 609, 609t
 percussion, 606-607, 607f, 608
 findings of, 611
 guidelines, evidence from, 612
 likelihood ratio test for, 613
 literature review, results for, 612
 literature search, 611
 original publication data, improvements in, 612
 prior probability, 613
 reference standard, changes in, 612, 613
 signs of
 accuracy, 608-609
 precision, 608
 splenic size, 605-606, 606f
 sputum production, 151, 152
 sensitivity, specificity, or likelihood ratio
 in obstructive airways disease, 152t
 SQ. *See* Single Question for depression
 square wrist sign, 112t
 stadiometer, 478-479
Staphylococcus aureus endocarditis, 519
 sterile stethoscope, 32
 straight-leg raising (SLR) sign, 78
 sensitivity, specificity, or likelihood ratio
 for disk herniation, 86t
 strawberry tongue, 617
 strength testing
 sensitivity, specificity, or likelihood ratio
 in carpal tunnel syndrome
 thumb, 115t
 in Parkinsonism
 difficulty rising from a chair, 514t
 in sciatica
 ankle, 79t
 great toe, 79t
 quadriceps, 79t
 strep throat, 615-621, 623-625
 clinical prediction rules for, 618-620, 619t
 Centor clinical prediction rule, 619, 619f
 McIsaac clinical prediction rule, 620, 620f
 Walsh algorithm, 621f
 findings of, 623
 guidelines, evidence from, 624
 likelihood of, 625
 literature review, results for, 624
 literature search, 623
 methods
 search strategy and quality review, 616
 statistical methods, 617
 original publication data, improvements in, 624
 pathophysiology of, 616
 pretest probability estimation of, 618
 prior probability, 625
 reference standard, changes in, 624, 625
 symptoms and signs, 617
 diagnostic accuracy of, 617-618
 precision of, 617
Streptococcus pneumoniae, 344, 400, 494, 501
 stroke, 627-638
 classification of, 635
 diagnosis of
 accuracy, 633
 flow, 628f
 reliability, 633
 ischemic stroke subtype analysis, 635-636
 likelihood ratio test for, 641
 methods, 629-630
 prehospital assessment, 630-631
 prior probability, 641
 prognosis of, 636-637, 636t
 reference standard tests, 641
 severity, assessment of, 634-635
 symptoms
 Oxfordshire Classification of Subtypes of Cerebral Infarction, 634
 transient ischemic attack, 627, 631-633
 vascular distribution of, 633-634

- Structured Clinical Interview for *DSM-III-R* (SCID), 40, 250, 254
- Structured Clinical Interview for *DSM-IV-TR* (SCID), 259
- students' aneurysm, 18
- subcutaneous tissue loss, subjective global assessment (SGA), 374
- subjective global assessment (SGA), of nutritional status
- in adult malnutrition, 373*t*, 376*t*
 - dietary intake change, 373
 - functional capacity, 374-375
 - loss of fluid from intravascular to extravascular space, 374-375
 - loss of subcutaneous fat, 374
 - muscle wasting, 374
- gastrointestinal symptoms, 374
- and postoperative complications, relationship between, 376*t*
- weight change, 372, 373
- sublingual nitroglycerin, 183
- sulcus sign, 585*t*
- sunken eyes
- sensitivity, specificity, or likelihood ratio
 - in hypovolemia, child 334*t*
- superior labrum anterior posterior (SLAP) lesion, 578, 586*t*
- swallowing, in myasthenia gravis, 454*t*
- Swan-Ganz catheterization, 203
- sweating
- sensitivity, specificity, or likelihood ratio
 - in pneumonia, adult, night sweats, 536*t*
- Swedish Two-County Trial, 89
- Symptom Driven Diagnostic System for Primary Care (SDDS-PC), 250, 251, 252
- symptomatic carotid bruit, 105
- systemic disease, low back pain
- ankylosing spondylitis, 77
 - cancer, 76
 - compression fractures, 77
 - spinal infections, 77
 - spine range-of-motion measures, 77
- systemic glucocorticoids, 149
- systolic blood pressure (SBP), 301, 302, 303
- systolic bruits, 35, 36
- systolic click, with mitral valve prolapse, 440*t*
- systolic dysfunction, 184, 186, 187
- diagnosis of, 210-211
 - and diastolic dysfunction, difference between, 211
- echocardiograms, 210-211
- postmyocardial infarction, 210
- and diastolic dysfunction, difference between, 189
- systolic murmurs, abnormal, 420, 436-437
- anatomic and physiologic origins of, 433-434
- aortic stenosis, 437
- causes, 434*t*
- clinical examination
- accuracy of, 436, 437*t*
 - precision of, 435-436, 436*t*
- evidence from guidelines, 445
- examination, 434-435, 440
- features, 435*f*
- findings of, 443
- hypertrophic cardiomyopathy, 439
- literature review
- accuracy, 444-445
 - precision, 444
- literature search, 443
- mitral regurgitation, 438-439
- mitral valve prolapse, 439-440
- original publication data, improvements in, 443
- prior probability, 446
- reference standard, changes in, 444, 447
- tricuspid regurgitation, 439
- systolic-diastolic abdominal bruits, 30, 31, 32
-
- ## T
- TA. *See* temporal arteritis
- T-ACE questionnaire, 44, 45, 49, 52*t*
- sensitivity, specificity, or likelihood ratio
 - for alcohol abuse, 50*t*
- tachycardia
- sensitivity, specificity, or likelihood ratio
 - in hypovolemia, adult, postural, 327*t*
 - in pneumonia, adult, 536*t*
 - for ventricular dysfunction, 211*t*
- supine, 320
- tachypnea, 543, 547, 548, 549, 561, 571, 574. *See also* pneumonia, in infant and child
- World Health Organization criteria for, 548, 548*t*
- definition, age based for children, 548
- sensitivity, specificity, or likelihood ratio
- in pneumonia, infant and child, 550*t*
- in pneumonia, adult, 536*t*
- in pneumonia, child, 340*t*
- telephone diagnosis, urinary tract infection, women, 687
- temporal arteritis (TA), 643-644
- accuracy, 646
 - of laboratory evaluation, 649-650
 - of physical examination, 646-649
 - of symptoms, 646
 - elicit signs and symptoms, 644-645
 - evidence from guidelines, 655
 - findings of, 653
 - likelihood ratio, 648*t*, 649*t*, 654*t*, 656*t*
 - literature review, 654
 - literature search, 653
 - methods
 - search strategy and quality review, 645
 - statistical methods, 645-646
 - multivariate findings for, 654-655
 - original publication data, improvements in, 654
 - pathophysiology of, 644
 - precision, 646
 - of medical history and physical examination, 464
 - prior probability, 656
 - reference standard, changes in, 654, 656
 - sensitivity, 648*t*, 649*t*
- temporal artery, in temporal arteritis, 649*t*
- tenderness of bicipital groove, 586
- tenting, 331
- test of Speed, 586*t*
- negative likelihood ratio, 589
 - positive likelihood ratio, 589
- test of Yergason, 586*t*
- negative likelihood ratio, 589
 - positive likelihood ratio, 589
- test of Zaslav. *See* internal rotation resistance strength test
- thenar atrophy, 112*t*
- third heart sound, 190
- Third National Health and Nutrition Examination Survey, 478
- thoracic aortic dissection, acute, 659-660
- chest radiograph
 - accuracy of, 666
 - sensitivity of, 667*t*
 - clinical examination
 - accuracy of, 662*t*, 673*t*, 665*t*

- thoracic aortic dissection, acute
(Continued)
- clinical history
 accuracy of, 663, 665
 sensitivity of, 664t
- combinations of findings
 accuracy of, 666-667
- diagnosis, 668t
- evidence from guidelines, 672
- findings of, 671
- likelihood ratio, 667t, 673, 673t
- literature review, 672
- literature search, 671
- methods
 data analysis, 663
 literature search and selection, 661
 study characteristics, 663
- original publication data,
 improvements in, 671
- pathophysiology of, 660
- physical examination
 accuracy of, 665-666
 sensitivity of, 666t
- prior probability, 673
- reference standards, 673
- sensitivity of, 672t
- sensitivity, specificity, or likelihood ratio
 in congestive heart failure, 666t
 signs and symptoms of, 660-661
- thumb abduction, testing, 113
- thyroid size
 sensitivity, specificity, or likelihood ratio
 for goiter, 282t
- thyroid-stimulating hormone (TSH), 286
- thyrotoxicosis, 277
- TIA. *See* transient ischemic attack
- tibia, 358, 359, 361
- Tinaquant D-dimer, 573
- Tinel sign, 112t, 114, 116
 sensitivity, specificity, or likelihood ratio
 in carpal tunnel syndrome, 124t
- tissue necrosis factor, 616
- tobacco use
 sensitivity, specificity, or likelihood ratio
 in myocardial infarction, 476t
 in obstructive airways disease, 161t, 162t
 in perimenopause, 416t
- tolerance, worry, eye opener, amnesia, *kut down* (TWEAK questionnaire), 49, 52t
 sensitivity, specificity, or likelihood ratio
 for alcohol abuse, 50t
- tongue weakness, 451
- tonsillar enlargement, in streptococcal pharyngitis, 618t
- tonsils, 617
- toothache
 sensitivity, specificity, or likelihood ratio
 in sinusitis, 603t
- tooth count, in osteoporosis, 481t
- total anterior circulation infarction syndrome (TACS), 634
- tourniquet test, 112t
- TR. *See* tricuspid regurgitation
- transient ischemic attack (TIA), 627, 631-633
 diagnosis of
 accuracy, 632
 reliability, 632-633
- transillumination
 sensitivity, specificity, or likelihood ratio
 in maxillary sinusitis, 603t
- Traube space percussion, 607, 607f
 sensitivity, specificity, or likelihood ratio
 for splenomegaly, 613t
- tremor, 507, 514
 action, 506, 507
 classic essential, 506, 514
 of Parkinson disease, 506
 physiologic, 506, 514
 rest, 506, 507
 sensitivity, specificity, or likelihood ratio
 in Parkinsonism, 514t
 tremor syndromes, 506
- tricuspid regurgitation (TR), 439
- tricuspid valvular dysfunction, 293
- TSH. *See* thyroid-stimulating hormone
- TWEAK questionnaire. *See* tolerance, *worry, eye opener, amnesia, kut down*
- 2-point discrimination, 112t
- 2-tailed χ^2 test, 529
- tympanic membrane, in otitis media, child, 503t
- tympanocentesis, 495, 496, 498, 501, 503
- tympanometry, 495
- type I collagen, 478
- U**
- ultrashort questionnaire. *See* Primary Care Evaluation of Mental Disorders
- ultrasonography, 17-18
 for abdominal aortic aneurysm, 26
- for acute cholecystitis, 138, 142
- for appendicitis, 54
- for deep vein thrombosis, 235, 240-241
- for liver examination, 292, 295
- for paranasal sinuses, 594
- unguisometer, 164, 171
- Unified Neurological Stroke Scale, 634
- unstable angina, 462
- upper lumbar disk herniation
 sit-to-stand test, accuracy, 84t
- urinary incontinence, perimenopausal, 409
- urinary tract infection (UTI), 675
 accuracy
 of dipstick urinalysis, 678
 of physical examination, 678
 of self-diagnosis, 678
 of signs and symptoms, 678, 679-680t
 of symptoms combinations, 678
- algorithm for evaluate patients with symptoms, 681f, 682-683
- definition, 675-676
- differential diagnoses, 676
- evidence from guidelines, 688
- findings of, 687
- likelihood ratio
 for symptoms combinations, 681t
- literature review, 688
- literature search, 687
- methods
 data analysis, 676-677
 quality assessment of included articles, 676
- multivariate approach, 689t
- original publication data,
 improvements in, 688
- precision of, 678
- pretest probability of, 680-682
- prior probability, 689
- reference standard, changes in, 688, 689
- refining probability
 dipstick urinalysis, 682
 with medical history and physical examination, 682
- rule out complicate, 680
- sensitivity analysis, 678, 680
- study characteristics, 677
- univariate findings, 689t
- urine specific gravity. *See* laboratory findings
- US Department of Agriculture, 48
- US Department of Health and Human Services, 48

USPSTF. *See* US Preventive Services Task Force
 US Preventive Services Task Force (USPSTF), 18, 47, 99, 109, 248, 259, 261, 286, 392, 393, 416, 477, 688, 706
 UTI. *See* urinary tract infection

V

vaginal complaints, in vaginitis, 691-692
 elicit symptoms and signs, 692-693
 evidence from guidelines, 706
 findings of, 705
 likelihood ratio, 707
 literature search, 705
 methods
 criterion standards, evaluation of, 693
 data extraction, 693
 evaluation of, 693
 inclusion and exclusion criteria, 693, 694t
 search strategy, 693
 statistical analysis, 693
 microscopic examination, 692f
 office laboratory tests, accuracy of, 699t, 700t
 inflammation, microscopic evidence of, 700
 microscopy, 700
 pH level, 700-701
 whiff test, 701
 original publication data,
 improvements in, 706
 precision of, 693
 prior probability, 707
 reference standard tests, 707
 signs, accuracy of
 discharge characteristics, 697, 699
 inflammation, 699-700
 odor, 700
 symptoms, accuracy of, 693-697
 bleeding, 697
 discharge characteristics, 695
 dyspareunia, 697
 irritative symptoms, 695
 itching, 695
 odor, 695, 697
 self-diagnosis, 697
 univariate findings for, 706t
 vaginal dryness
 perimenopausal, 409
 sensitivity, specificity, or likelihood ratio
 in menopause, 412t

vaginal infections, 676, 682
 vaginal symptoms
 sensitivity, specificity, or likelihood ratio
 in vaginitis, 707t
 discharge characteristics, 695t-698t
 in urinary tract infection
 discharge, 689t
 irritation, 689t
 Valsalva maneuver, 128, 196, 420
 sensitivity, specificity, or likelihood ratio
 in heart failure, 200t
 valvular heart disease
 physical examination, 444-445
 variable sexual interest,
 perimenopausal, 409
 vascular bruits, 29
 vascular distribution, of stroke, 633-634
 accuracy of, 633
 reliability of, 633-634
 vasodilator therapy, 184
 venography, 229
 venous hums, 104
 compared to arterial bruit, 292t
 venous thromboembolism, 227, 561
 risk factors for, 562t
 venous waveforms
 abnormal, 126t
 analysis of, 126
 in central venous pressure assessment, 126t
 ventricular fibrillation cardiac arrest, 215
 verification bias, 16, 138, 141, 498, 582, 589
 vertigo, 709-710
 causes, 710t
 elicit symptoms and signs, 711-712
 finding of, 715
 likelihood ratio, 717
 literature review, 716
 literature search, 715
 origin of, 710
 original publication data,
 improvements in, 716
 prior probability, 717
 reference standard, changes in, 716, 717
 symptoms and signs
 accuracy of, 712-713
 vestibular neuritis, 711, 712
 Veterans Affairs, 179

visual analog scale, 564
 volume depletion, 127, 315, 316, 325, 326, 327, 330, 331
 vomiting. *See* nausea and vomiting

W

wall-occiput distance test
 for occult thoracic vertebral fractures, 479, 485
 in osteoporosis, 483t
 Walsh algorithm
 for sore throat, 621f
 water hammer pulse (Corrigan), 425t
 weak thumb abduction, 112t
 web resources, for alcohol screening, 49
 weight
 in osteoporosis, 484, 485, 490t
 sensitivity, specificity, or likelihood ratio
 gain in ascites, 68t, 73t
 loss in low back pain, 76t
 weighted κ statistic, 571
 Welch-Allyn-Finnoff transilluminator, 595
 Wells scoring system, 565, 566t. *See also* clinical prediction rules and scores
 sensitivity, specificity, or likelihood ratio
 for deep vein thrombosis, simplified, 246t
 for pulmonary embolus, simplified, 572t, 575t
 wheezing, 151, 154-155
 sensitivity, specificity, or likelihood ratio
 in obstructive airways disease, 153f, 154t, 161t, 162t
 in pneumonia, adult, 536t
 in pneumonia, infant and child, 550t
 white coat hypertension. *See* office blood pressure
 WHO. *See* World Health Organization
 WHO-5. *See* World Health Organization-5 Well-Being Scale
 whole blood agglutination test, 230
 whole blood assays, 239
 WinBUGS software, 218
 WISE. *See* Women's Ischemia Syndrome Evaluation study

Women's Ischemia Syndrome
Evaluation (WISE) study,
415
workup bias. *See* verification bias
World Health Organization (WHO),
41, 48, 161, 247, 330, 393,
408, 462, 478, 539, 547,
548
Flunet, 344

International Influenza Program,
344
World Health Organization-5 Well-
Being Scale (WHO-5), 259

X

x-ray. *See* radiographic findings

Y

Yale 1-question screen, 259-260

Z

Zung Self-Rating Depression Scale
(SDS), 250, 251