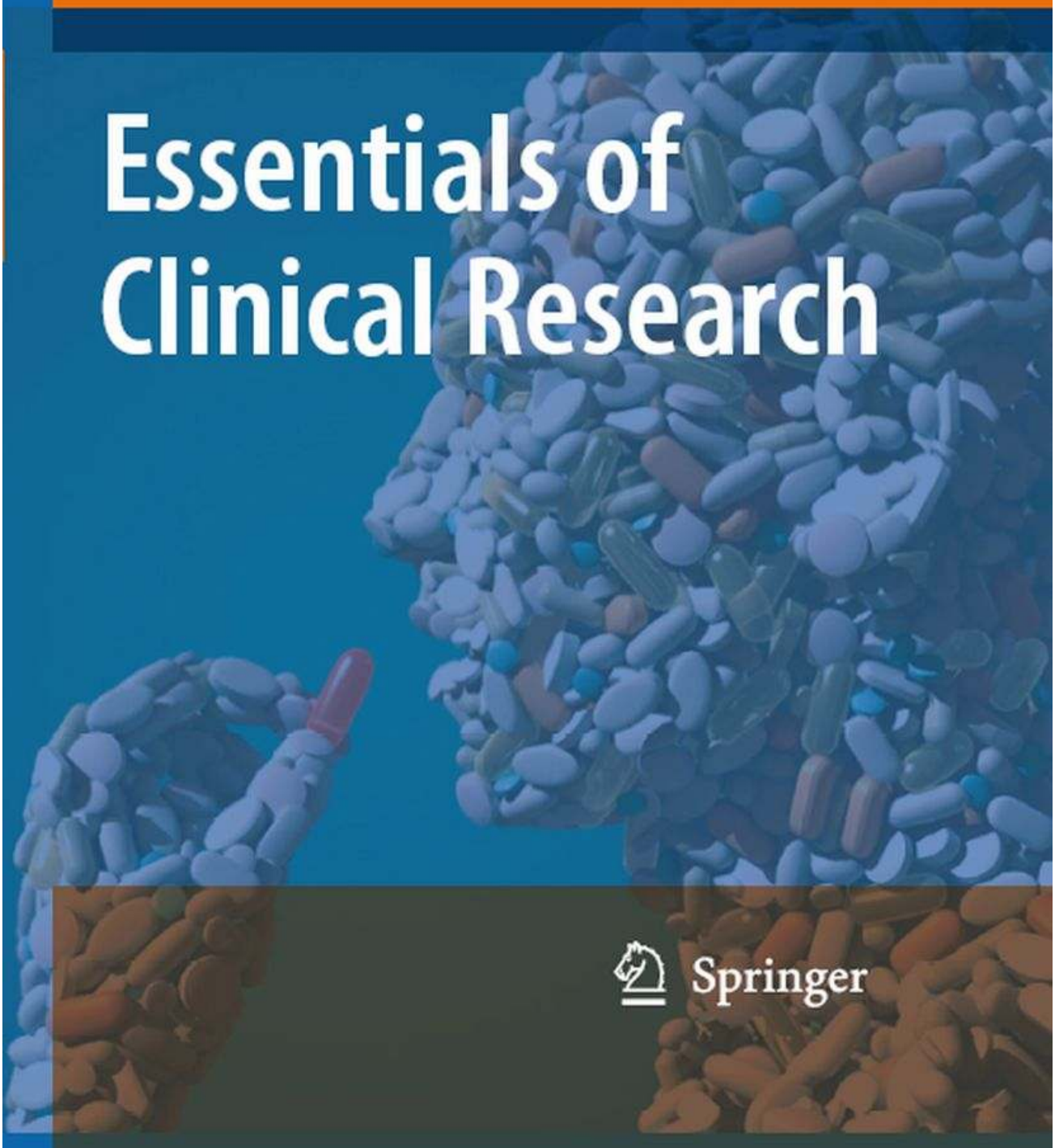
Stephen P. Glasser
*Editor*

# Essentials of Clinical Research



Springer

# Essentials of Clinical Research

Stephen P. Glasser
Editor

# Essentials of Clinical Research

Springer

*Editor*
Stephen P. Glasser
University of Alabama at Birmingham
AL, USA

ISBN 978-1-4020-8485-0 e-ISBN 978-1-4020-8486-7

Library of Congress Control Number: 2008927238



Printed on acid-free paper

9 8 7 6 5 4 3 2 1

springer.com

# Acknowledgements

The Fall Class of 2007 was asked to vet the majority of the chapters in this book, and did an excellent job. Two students went above and beyond, and for that I would like to acknowledge their contributions: John N. Booth III and Nina C. Dykes.

# Contents

"Goals in writing are dreams with deadlines." Brian Tracy http://www.briantracytickets.com/brian-tracy-quotes.php

**Part IV**

# Contributors

Jeroan J. Allison, MD, M.Sc.
Deep South Center on Effectiveness at the Birmingham VA Medical Center, Birmingham, AL; Professor of Medicine, Assistant Dean for Continuing Medical Education, UAB, University of Alabama at Birmingham, AL

Donna K. Arnett, Ph.D., MS, MPH
Professor and Chair of Epidemiology, Department of Epidemiology, School of Public Health, University of Alabama at Birmingham, Birmingham, AL

Carol M. Ashton, MD, MPH
Professor of Medicine, Division of Preventive Medicine, Department of Internal Medicine University of Alabama at Birmingham, Birmingham, AL

Gary Cutter, Ph.D.
Professor of Biostatistics, School of Public Health, University of Alabama at Birmingham, Birmingham, AL

Elizabeth Delzell, Ph.D., D.Sc.
Professor of Epidemiology, Department of Epidemiology School of Public Health, University of Alabama at Birmingham, Birmingham, AL

Sue Duval, Ph.D.
Assistant Professor Division of Epidemiology & Community Health, University of Minnesota School of Public Health, Minneapolis, MN

Carlos A. Estrada, MD, MS
Veterans' Administration National Quality Scholars Program, Birmingham VA Medical Center, Birmingham, AL; Associate Professor of Medicine, University of Alabama at Birmingham, AL; Deep South Center on Effectiveness at the Birmingham VA Medical Center, Birmingham, AL

William Frishman, MD, M.A.C.P.
The Barbara and William Rosenthal Professor and Chairman, The Department of Medicine, New York Medical College, New York City, NY

Stephen P. Glasser,
Professor of Medicine and Epidemiology, Univesity
of Alabama at Birmingham, Birmingham, Alabama
1717 11th Ave. South MT 638, Birmingham AL

Edward W. Hook III, MD
Professor of Medicine, University of Alabama at Birmingham School
of Medicine and Medical Director, STD Control Program, Jefferson County
Department of Health, Birmingham, AL

George Howard, DrPH
Professor and Chair Department of Biostatistics School of Public Health,
University of Alabama at Birmingham, Birmingham, AL

J. Michael Oakes, Ph.D.
McKnight Presidential Fellow, Associate Professor of Epidemiology &
Community Health, University of Minnesota, School of Public Health,
Minneapolis, MN

Amanda H. Salanitro, MD, MS
Veterans' Administration National Quality Scholars Program, Birmingham
VA Medical Center, Birmingham, AL

Maribel Salas, MD, D.Sc., M.Sc.
Assistant Professor at the Division of Preventive Medicine and Professor
of Pharmacoepidemiology, Department of Medicine and School of Public
Health, University of Alabama at Birmingham, Birmingham, AL.

Sadeep Shrestha, Ph.D., MHS, MS
Assistant Professor of Epidemiology, Department of Epidemiology, School
of Public Health, University of Alabama at Birmingham, Birmingham, AL

Bruno Stricker, MD, Ph.D.
Professor of Pharmacoepidemiology, Department of Epidemiology &
Biostatistics, Erasmus University Medical School, Rotterdam, and Drug Safety
Unit, Inspectorate for Health Care, The Hague, The Netherlands

O. Dale Williams, Ph.D., MPH
Professor of Medicine, Division of Preventive Medicine, University of Alabama
at Birmingham, Birmingham, AL

Nelda P. Wray, MD, MPH
Professor of Medicine, Division of Preventive Medicine, Department of Internal
Medicine, University of Alabama at Birmingham, Birmingham, AL

# List of Abbreviations

A Diabetes Outcome Prevention Trial = ADOPT
Absolute Risk Reduction = ARR
Acid Citrate Dextrose = ACD
Acute Myocardial Infarction Study = AMIS
Acute Respiratory Infections = ARI
Analysis of Variance = ANOVA
Area Under Curve = AUC
Attributable Risk = AR
Biological License Applications = BLA
Calcium Channel Blocker = CCB
Canadian Implantable Defibrillator Study = CIDS
Cardiac Arrhythmia Suppression Trial = CAST
Case-Control Study = CCS
Cholesterol and Recurrent Events = CARE
Clinical Trial of Reviparin and Metabolic Modulation of Acute Myocardial Infarction = CREATE
Computerized Provider Order Entry = CPOE
Consolidated Standards of Reporting Trials = CONSORT
Continuing Medical Education = CME
Controlled Onset Verapamil INvestigation of Cardiovascular Endpoints = CONVINCE
Coronary Heart Disease = CHD
Cross Sectional = X-sectional
Data Safety and Monitoring Board = DSMB
Data Safety and Monitoring Plan = DSMP
Deep Venous Thrombosis = DVT
Department of Health, Education and Welfare = HEW
Diltiazem LA = DLA
Division of Drug Marketing and Communications (DDMAC)
Drug Efficacy Study Implementation = DESI
Electromagnetic Energy = EME
Electron Beam Computed Tomography = EBCT
Emergency Medical Technician = EMT

Ethical, Legal, and Social Implications = ELSI
Ethylenediaminetetraacetic = EDTA
European Medicines Agency = EMEA
Evidence based Medicine = EBM
False Positive = FP
False Negative = FN
Food and Drug Administration = FDA
Good Clinical Practice = GCP
Good Medical Practice = GMP
Health Insurance Portability and Accountability Act = HIPPA
Health Maintenance Organizations = HMO
Hormone Replacement Therapy = HRT
Individual Patient Data = IPD
Institute of Medicine = IOM
Institutional Review Board = IRB
Intention to Treat = ITT
International Committee of Medical Journal Editors = ICMJE
International Conference on Harmonization = ICH
Intra-class Correlation Coefficient = ICC
Investigational New Drug = IND
Large Simple Trials = LST
Left Ventricular = LV
Linkage Disequilibrium = LD
Lung Volume Reduction Surgery = LVRS
Manual of Operations = MOOP
Multicenter Investigation of Limitation of Infarct Size = MILIS
Multicenter Isradispine Diuretic Atherosclerosis Study = MIDAS
Multiple Risk Factor Intervention Trial = MRFIT
Myocardial Infarction = MI
National Institutes of Health = NIH
Needed to Harm = NNH
Needed to Treat = NNT
New Drug Application = NDA
Nonsteroidal Antiinflamatory Drugs = NSAID
Number Needed to Treat = NNT
Odds Ratio = OR
Patient Oriented Research = POR
Pay for Performance = P4P
Pharmacoepidemiology = PE
Pharmacokinetics = PK
Physician Experience Studies = PES
Post Marketing Commitment Studies = PMCs
Premature Ventricular Contractions = PVCs
Principal Investigator = PI
Prospective, Randomized, Open-label, Blinded End-point = PROBE trial

Protected Health Information = PHI
Randomized Clinical Trial = RCT
Receiver Operator Characteristic Curves = ROC curve
Regression Towards Mediocrity = RTM
Relative Risk Reduction = RRR
Relative Risk = RR
Risk Difference = RD
Rural Diabetes Online Care = RDOC
Specific, Measurable, Appropriate, Realistic Time Bound = SMART
Stroke Prevention by Aggressive Reduction in Cholesterol Levels = SPARCL
Sudden Cardiac Death = SCD
The International Conference on Harmonization = ICH
The Myocardial Ischemia Reduction with Aggressive Cholesterol Lowering = MIRACL
The Pharmaceutical Research and Manufacturers of America = PhRMA
The Prescription Drug User Fee Act = PDUFA
The Strengthening and Reporting of Observational Studies in Epidemiology = STROBE
True Positive = TP
True Negitive = TN
U.S. National Health and Nutrition Examination Survey = NHANES
Unintended Adverse Events = UAEs
United States Federal Drug Administration = USFDA
Valsartin/Hydrochlorthiazide = VAL/HCTZ
Ventricular Premature Complexes = VPCs
White Blood Count = WBC
Woman's Health Initiative = WHI

# Part I

This Part addresses traditional clinical research, beginning with the history of the development of clinical research, to traditional clinical research designs, with a focus on clinical trials. It includes a discussion of the role of the USFDA in clinical trials and the placebo response, data safety and monitoring boards, and meta-analysis.

*When I re-read, I blush, for even I perceive enough that ought to be erased, though it was I who wrote the stuff.*

Ovid, Roma Poet as cite in Breslin JE etc. p 444

# Chapter 1
# Clinical Research: Definitions, "Anatomy and Physiology," and the Quest for "Universal Truth"[1]

**Stephen P. Glasser**

*Scientific inquiry is seeing what everyone else is seeing, but thinking of what no one else has thought.*

A. Szentgyorgyi, 1873 (he won the Nobel Prize for isolating Vitamin C)[2]

**Abstract** To answer many of their clinical questions, physicians need access to reports of original research. This requires the reader to critically appraise the design, conduct, and analysis of each study and subsequently interpret the results. This first chapter reviews some of the key historical developments that have led to the current paradigms used in clinical research, such as the concept of randomization, blinding (masking) and, placebo-controls.

## Introduction

As a former director of a National Institutes of Health (NIH)-funded K30 program, it was my responsibility to provide a foundation for young researchers to become independent principal investigators. A part of our curriculum was a course entitled 'The Fundamentals of Clinical Research.' This course, in addition to guiding students, was also designed to aid 'students' who wanted to read the medical literature more critically. This latter point is exemplified by the study of Windish et al.[3] They note that "physicians must keep current with the clinical information to practice evidence-based medicine…. To answer many of their clinical questions, physicians need access to reports of original research. This requires the reader to critically appraise the design, conduct, and analysis of each study and subsequently interpret the results."[3] Although aimed at physicians, this observation can and should be applied to all health scientists who must read the literature in order to place the results in context. The Windish study surveyed 277 completed questionnaires that assessed knowledge about biostatistics, and study design. The overall mean percent correct on statistical knowledge and interpretation of results was 41.4%.

It is my belief that the textbooks currently available are epidemiologically "slanted". There is nothing inherently wrong with that slant, but I have written this book to be more specifically geared to the clinical researcher interested in conducting

S.P. Glasser (ed.), *Essentials of Clinical Research*,

Patient Oriented Research (POR). In this first chapter I will provide a brief overview of the history of clinical research. The chapter will also address the question of why we do clinical research; define 'clinical research'; discuss our quest for 'universal truth' as the reason for doing clinical research; outline the approach taken to answer clinical questions; and describe (as Hulley an colleagues so aptly put it) 'the anatomy and physiology of clinical research.'[1]

Future chapters will examine such issues as causality (i.e., causal inference or cause and effect); the strengths and weaknesses of the most popular clinical research designs; regression to the mean; clinical decision making; meta-analysis; and the role of the Food and Drug Administration (FDA) in the clinical trial process. We will also focus on issues related to randomized clinical trials, such as the intention-to-treat analysis, the use and ethics of placebo-controlled trials, and surrogate and composite endpoints.

## Definition of Clinical Research

The definition of clinical research might appear to be self-evident; however, some researchers have narrowly defined clinical research to refer to clinical trials (i.e., intervention studies in human patients), while others have broadly defined it as any research design that studies humans (patients or subjects) or any materials taken from humans. This latter definition may even include animal studies, the results of which more or less directly apply to humans. For example, in 1991, Ahrens included the following in the definition of clinical research: studies on the mechanisms of human disease; studies on the management of disease; in vitro studies on materials of human origin; animal models of human health and disease; the development of new technologies; the assessment of health care delivery; and field surveys.[4] In an attempt to simplify the definition, some wits have opined that clinical research occurs when the individual performing the research is required to have malpractice insurance, or when the investigator and the human subject are, at some point in the study, in the same room, and both are alive and warm. So, there is a wide range of definitions of clinical research, some valid, some not. I have chosen to adopt a 'middle of the road' definition that encompasses the term 'patient-oriented-research,' which is defined as research conducted with human subjects (or on material of human origin) for which the investigator directly interacts with the human subjects at some point during the study. It is worth noting that this definition excludes in vitro studies that use human tissue that may or may not be linked to a living individual unless the investigator during the conduct of the trial has significant interaction with a living breathing human.

## History of Clinical Research

Perhaps the first clinical trial results were those of Galen (circa 250 BC) who concluded that 'some patients that have taken this herbivore have recovered, while some have died; thus, it is obvious that this medication fails only in incurable diseases.'

Galen's observations underline the fact that even if we have carefully and appropriately gathered data, there are still subjective components to its interpretation, indicating our quest for 'universal truth' is bedeviled more by the interpretation of data than by its accumulation.

James Lind is generally given credit for performing and reporting the first placebo-controlled interventional trial in the treatment and prevention of scurvy. In the 1700s, scurvy was a particularly vexing problem on the long voyages across the Atlantic Ocean. The research question that presented itself to Lind was how to prevent the condition. To arrive at an answer, Lind did what every good researcher should do as the first step in converting a research question into a testable hypothesis – he reviewed the existent literature of the time. In so doing, he found a report from 1600 that stated '1 of 4 ships that sailed on February 13th, 1600, was supplied with lemon juice, and almost all of the sailors aboard the one ship were free of scurvy, while most of the sailors of the other ships developed the disease.' On the one hand, Lind's job was easy-there was not a great deal of prior published works. On the other hand, Lind did not have computerized searches via Med Line, Pub Med etc. available.

As a result of the above, in 1747, Lind set up the following trial. He took 12 patients 'in the scurvy' on board the HMS *Salisbury*. 'These cases were as similar as I could have them….they lay together in one place … and had one diet common to all. The consequence was that the most sudden and visible good effects were perceived from the use of oranges and lemons.' Indeed, Lind evaluated six treatment groups: 'One group of two was given oranges and lemons. One of the two recovered quickly and was fit for duty after 6 days, while the second was the best recovered and was assigned the role of nurse for the remaining patients.' The other groups were each treated differently and served as controls. If we examine Lind's 'study' we find a number of insights important to the conduct of clinical trials as follows. For example, Lind noted that 'on the 20th May, 1747, I took twelve patients in the scurvy on board the Salisbury at sea… Their cases were as similar as I could have them. They all in general had putrid gums, the spots and lassitude, with weakness of their knees…' here Lind was describing eligibility criteria for his study. He continues, '…They lay together in one place, being a proper apartment for the sick in the fore-hold; and had one diet in common to all…' '… Two of these were ordered each a quart of cyder a day. Two others took twenty five gutts of elixir vitriol three times a day upon an empty stomach,

… Two others took two spoonfuls of vinegar three times a day
… Two … were put under a course of sea water.
… Two others had each two oranges and one lemon given them every day.
… The two remaining patients took the bigness of a nutmeg three times a day.'

By this description, Lind described the interventions and controls. To continue, '… The consequence was that the most sudden and visible good effects were perceived from the use of the oranges and lemons; one of those who had taken them being at the end of six days fit four duty. The spots were not indeed at that time quite off his body, nor his gums sound; but without any other medicine than a gargarism or elixir of vitriol he became quite healthy before we came into Plymouth, which was on the

16th June.' This latter description represents the outcome parameters and interpretation of his study. In summary, Lind addressed the issues of parallel-group design and the use of control groups, and he attempted to assure similarity between the groups except for the intervention.

Clearly, sample size considerations and randomization were not used in Lind's trial, but this small study was amazingly insightful for its time. Other selected milestones in the history of clinical research include:

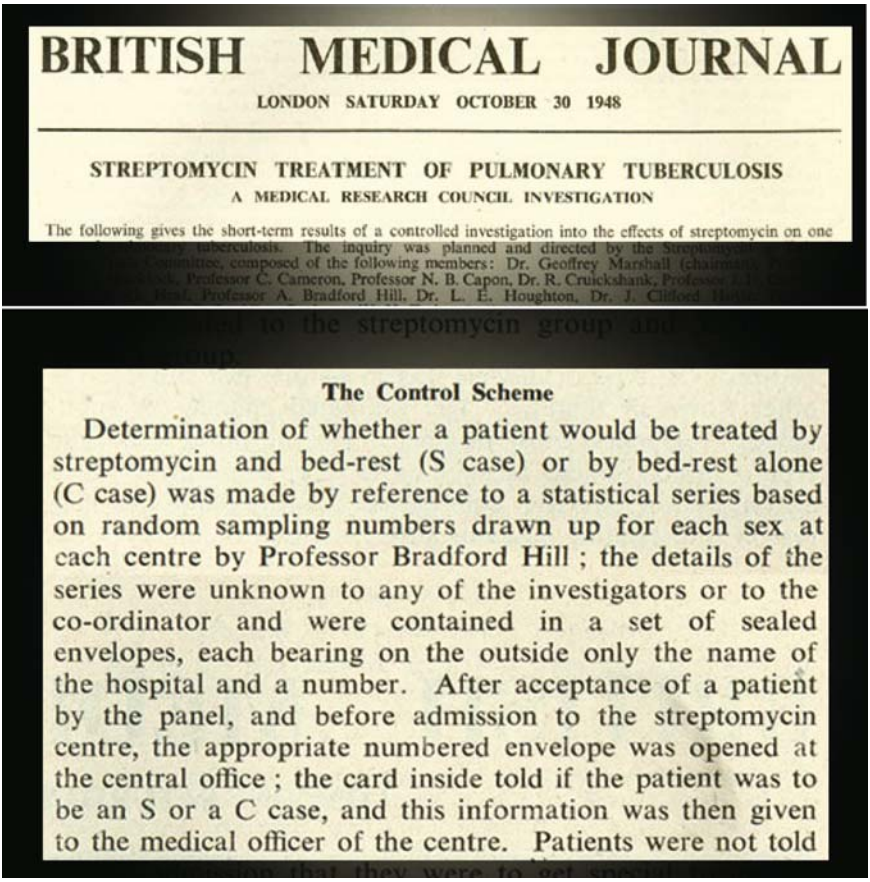
- Fisher's introduction of the concept of randomization in 1926.[5]
- The announcement in 1931 by the Medical Research Council that they had appointed 'a therapeutics trials committee…to advise and assist them in arranging for properly controlled clinical tests of new products that seem likely on experimental grounds to have value in the treatment of disease'.[6]
- Amberson and colleagues' introduction of the concept of 'blindness' in clinical trials[6] and their study of tuberculosis patients where the process of randomization was applied.[7] They noted that after careful matching of 24 patients with pulmonary tuberculosis, the flip of a coin determined which group received the study drug.[7]

Further analysis of the tuberculosis streptomycin study of 1948 is regarded as the beginning of the modern era of clinical research and is instructive in this regard. In the 1940s tuberculosis was a major public health concern, and randomization was being recognized as a pivotal component to reduce bias in clinical trials.[8] As a result the Medical Research Council launched a clinical trial in which 55 patients were randomized to treatment with bed rest (the standard of care treatment at that time) and 52 were treated with bed rest alone. In Fig. 1.1 one can read Professor Bradford Hill's insightful description of the randomization process.[9]

Other significant developments include reference to the use of saline solution in control subjects as a placebo, and the requirement in 1933 that animal toxicity studies be performed before human use.[8] In the 1940s, the Nuremberg Code, the Declaration of Helsinki, the Belmont Report, and the doctrine of Good Clinical Practice (GCP) were developed, which will be discussed in more detail later. As mentioned above, In 1948, the Medical Research Council undertook a streptomycin study[9] which was perhaps the first large-scale clinical trial using a properly designed randomized schema. This was followed by an antihistamine trial that used a placebo arm and double-blind (masked) design.[10]

In 1954, there were large-scale polio studies – field trials of 1.8 million school-age children. A controversy regarding the best design resulted in two trials, one design in which some school districts' second graders received the dead virus vaccine while first and third graders acted as the controls; and another design in which second graders randomly received either the vaccine or a saline injection. Both studies showed a favorable outcome for the vaccine (Fig. 1.2).

In 1962, the thalidomide tragedy became widely known and resulted in the tightening of government regulations. The story behind this tragedy is instructive. By 1960, thalidomide worldwide was being sold, but not in the United States. At the



BRITISH MEDICAL JOURNAL

LONDON SATURDAY OCTOBER 30 1948

STREPTOMYCIN TREATMENT OF PULMONARY TUBERCULOSIS

A MEDICAL RESEARCH COUNCIL INVESTIGATION

The following gives the short-term results of a controlled investigation into the effects of streptomycin on one

The Control Scheme

Determination of whether a patient would be treated by streptomycin and bed-rest (S case) or by bed-rest alone (C case) was made by reference to a statistical series based on random sampling numbers drawn up for each sex at each centre by Professor Bradford Hill ; the details of the series were unknown to any of the investigators or to the co-ordinator and were contained in a set of sealed envelopes, each bearing on the outside only the name of the hospital and a number. After acceptance of a patient by the panel, and before admission to the streptomycin centre, the appropriate numbered envelope was opened at the central office ; the card inside told if the patient was to be an S or a C case, and this information was then given to the medical officer of the centre. Patients were not told

**Fig. 1.1** From the British Medical Journal, Sir Bradford Hill's description of the randomization process

time, the prevailing US law was the 1938 Federal Food, Drug, and Cosmetic Act, which required proof of safety be sent to the FDA before a medication could be approved for sale in the United States. The law did not require demonstration of efficacy for approval. It also allowed "investigational" or "experimental" use of a drug while approval for its sale was being sought, allowing a medication to be widely distributed prior to approval. The application for the USA use was given to Frances Kelsey who noted a lack of teratogenicity data, and she also had other worries about thalidomide. As a result, Kelsey rejected the application and requested additional data from the company, but the company complained to her superiors that she was nit-picking and unreasonable. Kelsey continued to refuse to approve thalidomide for sale in the United States, and in total, the company resubmitted its application to the FDA six times, but with no new evidence in those applications Kelsey refused approval.

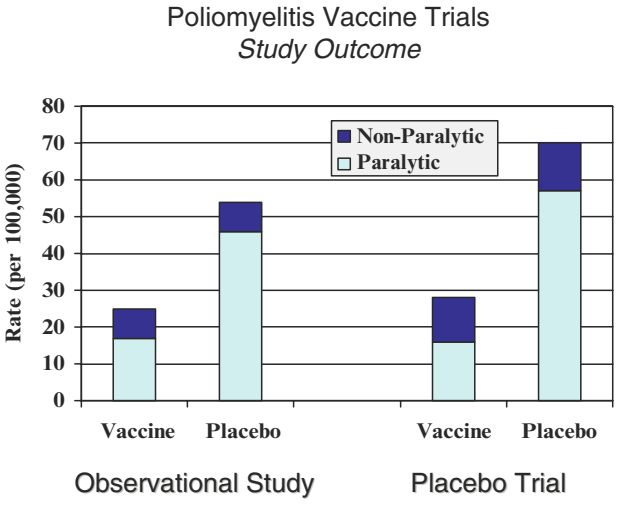



**Fig. 1.2** Results form the use of polio vaccines used in both an observational trial and a placebo controlled clinical trial

Subsequently, reports regarding a number of birth defects were reported and the drug was removed worldwide, thereafter.[11]

As mentioned prior, at the time of the thalidomide disaster, trials of new drugs were required to prove safety but not efficacy as described under the FDA's 1938 Act. As a result of the disaster, tightening of the regulations was instituted and trials were to have an 'adequate and well-controlled design' before approval of any new drug. This was followed by the Drug Efficacy Study Implementation (DESI) review and the FDA's development of the four stages of clinical trials necessary for new drug approval, which set the stage for today's process of drug approval.

In the 1970s and 1980s, clinical research was prospering, but by the 1990s there began a decline in the number of new clinical investigators. This trend caught the eye of a number of academicians and the NIH, which then commissioned the Institute of Medicine (IOM) to address ways to stimulate individuals to pursue careers in clinical investigation, to define appropriate curricula for training, and to ensure adequate support mechanisms for retaining clinical researchers.

The NIH also developed granting mechanisms for supporting individual clinical investigators at various levels of their careers (K23 and K24 grants) and for programmatic support of institutions that developed clinical research training programs (K30 grants). The IOM report documented the decline in clinical investigators (particularly MD investigators), and noted that the time commitment necessary to do clinical research was underappreciated.[12]

Recently, DeMets and Califf noted, 'we are entering an era in which the imperative to understand the rational basis for diagnostic and therapeutic options has become a major force in medical care.' Medical products (drugs, devices, and

biologics) are proliferating simultaneously with substantial restructuring of the delivery of health care, with a focus on evidence to support medical intervention.[13]

Today, we are left with the 'good, the bad, and the ugly' regarding clinical research. The 'good' is that many experts think that sound comprehension of the scientific method and exposure to biomedical research comprise the essential core of medical education, and that the very essence of the American academic model is a balance between education, patient care, and research. The 'bad' is the increasing number of voices questioning the relevancy of research in academic health centers, as well as those concerned about the commitment to other components of training and the cost of research in a setting where the 'triple threat' (i.e., excelling in teaching, patient care, and research) may no longer be tenable given the increasing complexity of each area. The 'ugly' is that in 2003 only about 3 cents of every health care dollar was spent on medical research; and, it was estimated that only 5% of Congress could be counted on to take the initiative and be leaders in the support of clinical research; and few potential investigators were being supported to pursue careers or were given enough time to conduct research. By and large, these same issues persist today.

With the above background, how do we begin our quest for knowledge? In general, research questions are generated in a variety of settings (e.g., during journal reading, hospital rounds, discussions with colleagues, seminars, and lectures). The resultant questions can then be refined into a research idea and, after further review of the literature, ultimately developed into a hypothesis. Based on a number of factors (to be discussed in subsequent issues), a study design is chosen, and the study is then preformed and analyzed, the results of which are then interpreted and synthesized. These results add to the body of knowledge, and this may raise additional questions that will invariably generate further research (Fig. 1.3).

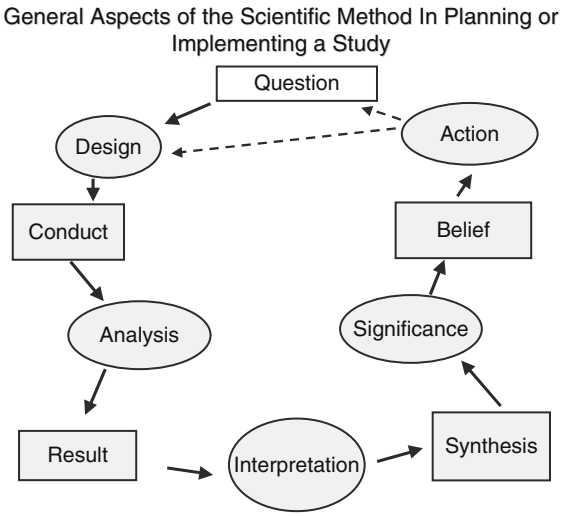



**Fig. 1.3** General aspects of the scientific method in planning or implementing a study

Of course, the primary goal of clinical research is to minimize presumption and to seek universal truth. In fact, in science, little if anything is obvious, and the interpretation of results does not mean truth, but is really an opinion about what the results mean. Nonetheless, in our quest for universal truth, Hully and colleagues have diagrammed the steps that are generally taken to seek this 'truth' (Fig. 1.4).[1] These latter concepts will be discussed in subsequent chapters. Finally, it should be realized that clinical research can encompass a broad range of investigation as portrayed in Fig. 1.5.

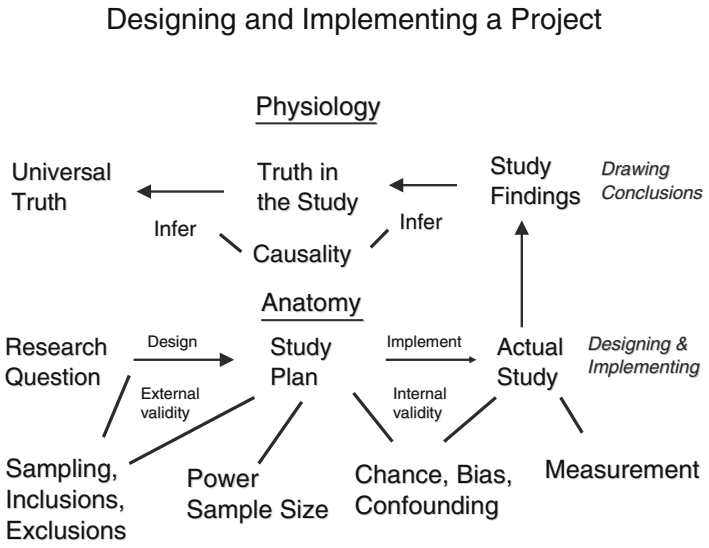



**Fig. 1.4** Designing and implementing a project

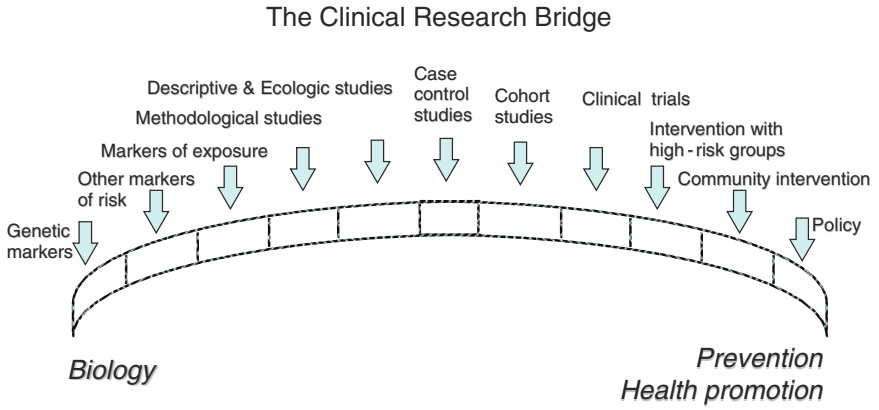



**Fig. 1.5** Portrays the broad range that encompasses the term "clinical research"

## References


1. Hulley S, Cummings S, Browner Wea. *Designing Clinical Research*. 2nd ed. Philidelphia, PA: Lippincott, Williams & Wilkins; 2000.
2. http://www.brainyquote.com/quotes/authors/a/albert szentgyorgyi.html
3. Windish DM, Huot SJ, Green ML. Medicine residents’ understanding of the biostatistics and results in the medical literature. *JAMA*. Sept 5, 2007; 298(9):1010–1022.
4. Ahrens E. *The Crisis in Clinical Research: Overcoming Institutional Obstacles*. New York: Oxford University Press; 1992.
5. Fisher R. *The Design of Experiments*. Edinburgh: Oliver & Boyd; 1935.
6. Hart PD. Randomised controlled clinical trials. *BMJ*. May 25, 1991; 302(6787):1271–1272.
7. Amberson JB, MacMahon BT, Pinner M. A clinical trial of sanocrysin in pulmonary tuberculosis. *Am Rev Tuber*. 1931; 24:401–435.
8. Hill AB. The clinical trial. *Br Med Bull*. 1951; 7(4):278–282.
9. White L, Tursky B, Schwartz G. *Placebo: Theory, Research, and Mechanisms*. New York: Guilford Press; 1985.
10. Medical Research Council. Streptomycin treatment of pulmonary tuberculosis. *BMJ* 1948; ii:769–782.
11. Thalidomide. http://en.wikipedia.org/wiki/Thalidomide.
12. Institute of Medicine. *Careers in Clinical Research: Obstacles and Opportunities*. Washington, DC: National Academy Press; 1994.
13. DeMets DL, Califf RM. Lessons learned from recent cardiovascular clinical trials: Part I. *Circulation*. Aug 6, 2002; 106(6):746–751.

# Chapter 2
# Introduction to Clinical Research and Study Designs

**Stephen P. Glasser**

*To educate is to guide students on an inner journey toward more truthful ways of seeing and being in the world.*

Parker J. Palmer[1]

**Abstract** This chapter addresses some of the central concepts related to clinical research and what is meant by the strength of scientific evidence. We also begin to discuss the different clinical research designs along with their respective strengths and weaknesses.

## Sampling

An essential characteristic and the goal of any clinical research are to make inferences from the population under study (the sample or study population) and apply those inferences to a broader population (the target population i.e. the population about which we want to draw conclusions). Imagine if the investigator could only learn about and apply the results in the sample population? Rather we must be able to extrapolate the results of the findings in the sample population to a broader group of patients-otherwise the results would have no utility at all. Thus, one of the most important weaknesses of any study is that inferences drawn from a study are based on a limited sample (again, a sample is a select subset of a population that the investigator hopes represents the general population, but which is unlikely to do so). This limitation is further compounded by the fact that disease is not distributed randomly, whereas samples tend to be, and that the causes of disease are multifactorial. Ideally, when performing clinical research, we would like to include everyone in our study who has the disease of interest. Because this is impossible we settle for a sample of the diseased population, however, the researcher now has to deal with a degree of uncertainty (see Chapter 18). Because different samples contain different people with different co-morbidities, and differing experiences, we end up with different data. The question now facing the researcher is which data from which sample is most representative of the entire population? Sampling errors commonly result in Type I and II errors. For example, if the researcher finds a certain effect of

S.P. Glasser (ed.), *Essentials of Clinical Research*,

an interventional therapy, the question to be asked is ‘how likely is it that this therapy observation that was made from this sample is falsely representing the total population (in which there was in fact no therapy effect)? This potential false result is Type I error and is addressed by the p value. The reverse situation is a total population that in fact has a therapy effect, but the sample studied shows no such effect. This is the Type II error.

## The Linear-Semilinear Relationship of Biological Variables

Another important concept of clinical research is the fact that most, if not all biological variables have a linear–semilinear relationship in terms of exposure and outcomes, whereas clinical medicine is replete with the use of ‘cut-points’ to separate normal and abnormal or effect and no effect (Fig. 2.1). A cut-point presumes that there is some value or range of values that separates normal form abnormal rather than considering that the relationships tend to be on a continuum.

## Strength of Relationships

Another important issue in clinical research relates to what we mean when we talk about ‘the strength of evidence.’ The greatest strength of evidence is often attributed to the randomized clinical trial (RCT). In fact, in response to the question of what is the best clinical research design, the answer generally given is ‘the RCT,’ when in fact the correct answer should be ‘it depends,’ an answer which will be

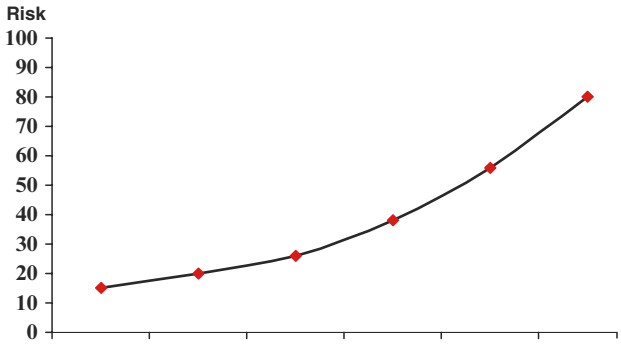



**Fig. 2.1a** Epidemiological view of determining risk

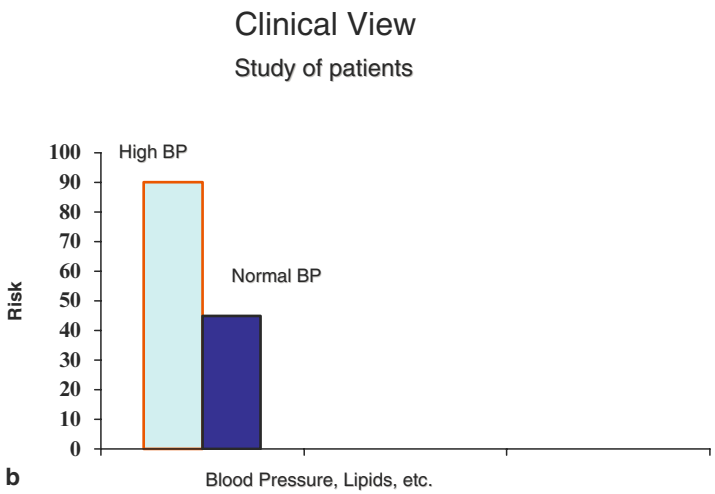



**Fig. 2.1b** Study of patients: clinical view

further discussed later in this book. What is actually meant by 'the highest level of evidence' is how certain we are that an exposure and outcome are causally related, that is, how certain we are that an effect is the result of a cause, and that the observations are not just an association that exists; but, which are not causally related.

## The Hypothesis

Let's return to the question: 'What is the best study design?' This is a different question from 'What is the best study design for a given question and given the specific question, which study design leads to the highest level of evidence?'; which may finally be different from asking 'What is the study design for a given question that will result in the greatest certainty that the results reflect cause and effect?' This latter question is really the one that is most often sought, and is the most difficult to come by. (See Chapter 16 on Causation.) Other important factors in considering the most appropriate study design, besides the most important factor – ethics – include the natural history of the disease being studied, the prevalence of the exposure, disease frequency, the characteristics and availability of the study population, measurement issues, and cost.

Let us now return to our quest for 'universal truth.' What are the steps we need to take in order to achieve 'truth'? The fact is that truth is at best elusive and is not actually achievable since truth is more a function of our interpretation of data,

which is mostly dictated by our past experiences, than any finite information that is absolute. The steps needed to achieve this uncertain quest for truth begins with a research question, perhaps the result of a question asked during teaching rounds, or stimulated by contact with a patient, or provoked during the reading of a book or journal, and so on. The research question is usually some general statement such as 'Is there an association between coffee drinking and myocardial infarction (MI)?' or 'Is passive smoke harmful to a fetus?' Let us examine this last research question and consider its limitations in terms of a testable hypothesis. In addressing a question such as 'Is passive smoke harmful to a fetus?' one needs first to ask a few questions such as: 'what is the definition of 'harmful'; how will passive smoke be measured and what do we mean by the term i.e. how is it to be defined in the study to be proposed?' Answering these questions comes nearer to something that is testable and begins to define the clinical research design that would have the greatest level of evidence with that specific question in mind. For the question proposed above, for example, it would be best, from a research design perspective, to randomize exposure of pregnant women to both passive smoke and 'placebo passive smoke.' But considering the ethics issue alone, this would not be acceptable; thus, an RCT would not be the 'best study design' for this research question, even if it would lead to the 'highest level of evidence'.

The hypothesis is generally (for the traditional approach of superiority testing) stated in the null (Ho). The alternative hypothesis (HA) i.e. the one you are really interested in is, for example, that a new drug is better than placebo. That is, if one wants to compare a new investigational drug to placebo, the hypothesis would be constructed in the null, i.e. that there is no difference between the two interventions. If one rejects the null, one can then say that the new drug is either better (or worse-depending on the results of the study) than placebo. By the way, if the null is not rejected one cannot say that the new drug is the same as placebo, one can only claim that no difference between the two is evident from these data (this is more than a nuance as will be discussed later).

In order to understand why the hypothesis is stated in the null and why one cannot accept the null but only reject it, consider the following three examples (taking a trip with your family, shooting baskets with Michael Jordon, and contemplating the US legal system). Consider the scenario outlined by Vickers[2] where you have just finished packing up your SUV (a hybrid SUV no doubt) with all of your luggage, the two kids, and your dog, and just as you are ready to depart; your wife says 'honey, did you pack the camera?' At least two approaches present themselves; one that the camera is in the automobile, or two that the camera is in the house. Given the prospect of unpacking the entire SUV, you decide to approach the question with, 'the camera is not in the house (Ho) i.e. it is in the car'. If you in fact do not find the camera in the house you have rejected your null and your assumption is that it is in the car. Of course, one can easily see that the camera could be in the house (you just did not find it), and even if you did such a thorough job of searching the house that you can be almost certain that it is not there, it still may not be in the car (you might have left it elsewhere (the office, a prior vacation, etc.) Another way to look at this issue is to envision that you are out on the basketball court when Michael Jordon

comes in. You challenge him to a free throw shooting contest and he makes 7 of 7 while you make 3 of 7. It turns out the p value for this difference is 0.07 i.e. there is no "statistically significant difference between the shooting skills of MJ and your shooting skills-you can draw your own conclusions about this likelihood.[2] In the Woman's Health Initiative (WHI), women eating a low fat diet had a 10% reduction in breast cancer c/w controls P = 0.07. This was widely interpreted as low fat diets don't work. In fact, the NY Times trumpeted that 'low fat diets flub a test' and that the study provided 'strong evidence that the war against all fats was mostly in vain'. This is what we call accepting the null hypothesis (i.e. it was not rejected so it was accepted) and is to be avoided i.e. failure to reject it does not mean you accept it, rather it means that these data do not provide enough evidence to reject it. By the way, guess what happens when the next study does reject the null – 'but they said it did not work!'. Finally, consider our Anglo-American legal system.

It is no mere coincidence that the logic of hypotheses testing in scientific inquiry is identical to that which evolved in the Anglo-American legal system and most of the following descriptions are taken from The Null Logic of Hypothesis Testing found on the World Wide Web.[3] Much of the pioneering work in the logic of hypothesis testing and inferential statistics was done by English mathematicians and refined by their American counterparts. For instance consider the contributions made by W.S. Gossett, R.A. Fisher, and Karl Pearson to the logic of hypothesis testing and statistical inference. The concept of the null hypothesis can be compared to the legal concept of guilty vs non guilty, the latter of which does not mean innocence. What is interesting is that the guilt vs innocent scenario involves two diametrically apposed logics, one affirmative and the other null. From the time a crime is reported to the police an affirmative, accusatory, and inductive logic is followed. Detective X gathers the evidence, follows the evidentiary trail, and based upon the standard of probable cause, hypothesizes that the accused is guilty and charges him accordingly. The District Attorney reviews the case for probable cause and quality of evidence and affirms the accusation. The case is argued affirmatively before the grand jury, and they concur. But relative to the jury, at the point the trial begins, the logic is reversed. It is no longer affirmative, it becomes null. The jury, the trier of the facts, is required to assume that the defendant is not guilty unless the facts established otherwise. Let's abstract this two part logical process and represent it symbolically. The police, the prosecutor, and the grand jury hypothesized ($H_1$) that the accused (X) committed the crime (Y).

The jury on the other hand hypothesizes ($H_0$) that the accused (X) was not guilty of the crime (Y) unless the evidence reached the standard of "beyond a reasonable doubt".

Formulating the logic in this manner, one can be certain of two things. Either:

$H_0$ is true, the accused is not guilty, or
$H_1$ is true, accused is guilty, and
$H_0$ and $H_1$ cannot both be true

The logic of establishing someone's guilt is not the simple converse of the logic of establishing his/her innocence. For instance, accusing someone of a crime and

requiring them to prove their innocence requires proving a negative, something that is not logically tenable. However, assuming that someone is not guilty and then assessing the evidence to the contrary is logically tenable.

The decision matrix in Table 2.1 shows the possible outcomes and consequences of this legal logic as applied to the case of the accused, our hypothetical defendant. Assume $H_0$: the accused is not guilty unless the evidence is convincing beyond a reasonable doubt. Notice that in terms of verdicts and outcomes, there are two kinds of errors the jury might have made, identified as (I) and (II).

Type I Error The jury finds the accused guilty when in fact he is not guilty.
Type II Error The jury finds the accused not guilty when in fact he is guilty.
Compare this with the Table 18.2.

In the Anglo-American legal tradition, the consequences of these two possible errors are not considered equivalent. On the contrary, considerable safeguards have been incorporated into the criminal law to minimize the probability ($\alpha$) of making a Type I error (convicting an innocent person), even at the risk of increasing the probability ($\beta$) of making a Type II error (releasing a guilty person). Indeed, this is where the concept of innocent until proven guilty comes from, and the quote: as the noted 18th Century English jurist Sir William Blackstone that justice is better served if made ten guilty persons escape than that one innocent suffer."[4]

It is logical and critical to distinguish between the concepts of not guilty and innocent in the decision paradigm used in criminal law, i.e.:

If $H_1$ = guilty, then does …
$H_0$ = not guilty, or does …
$H_0$ = innocent?

Here, guilty does not mean the same thing as innocent A not guilty verdict means that the evidence failed to convince the jury of the defendant's guilt beyond a reasonable doubt (i.e. the scientific corollary is that data in this study was insufficient to determine if a difference exists, rather than there is no difference"). By this logic it is quite conceivable that a defendant can be found legally not guilty and yet not be innocent of having committed the crime in question.

**Table 2.1** Decision Matrix for Determining Guilt or Innocence

| | The Verdict | |
|---|---|---|
| The Truth | Accused *is not* guilty: $H_0$ accepted | Accused *is* guilty: $H_0$ rejected |
| Accused *is not* guilty: $H_0$ true | Justice is served | (I) An innocent man is convicted<br>Probability = $\alpha$ |
| Accused *is* guilty: $H_0$ false | (II) A guilty man is set free<br>Probability = $\beta$ | Justice is served |

The evaluation of a hypothesis involves both deductive and inductive logic. The process both begins and ends with the research hypothesis.

Step 1 Beginning with a theory about the phenomenon of interest, a research hypothesis is deducted.
This hypothesis is then refined into a statistical hypothesis about the parameters in the population.
The statistical hypothesis may concern population means, variances, medians, correlations, proportions, or other statistical measures.
The statistical hypothesis is then reduced to two mutually exclusive and collectively exhaustive hypotheses that are called the null ($H_0$) and alternative hypothesis ($H_1$).

Step 2 If the population is too large to study in its entirety (the usual case), a representative sample is drawn from the population with the expectation that the sample statistics will be representative of the population parameters of interest.

Step 3 The data gathered on the sample are subjected to an appropriate statistical test to determine if the sample with its statistical characteristics could have come from the associated population if the null hypothesis is true.

Step 4 Assuming that the null hypothesis ($H_0$) is true in the population, and that the probability that the sample came from such a population is very small ($p \leq 0.05$), the null hypothesis is rejected.

Step 5 Having rejected the null hypothesis, the alternative hypothesis ($H_1$) is accepted, and, by inductive inference is generalized to the population from whence the sample came.

These five steps are illustrated in Fig. 2.2, that is, the conduct of research involves a progressive generation of four kinds of hypotheses: Research hypothesis, Statistical hypothesis Null hypothesis; and, Alternative hypothesis.

A research hypothesis is an affirmative statement about the relationship between two variables. For instance, consider the following example of a research hypotheses: “there is a positive correlation between the level of educational achievement of citizens and their support of rehabilitation programs for criminal offenders”. From the research hypotheses three other kinds of hypotheses can be formulated:

A statistical hypothesis
A null hypothesis
An alternative hypothesis

Again, a statistical hypothesis is a statement about the parameters of a population. The null hypothesis, which is symbolized $H_0$, is the negative statement of the statistical hypothesis; and, the alternative hypothesis, symbolized $H_1$ (or Ha), is the obverse of the null hypothesis and by custom, is stated to correspond to the research hypothesis being tested. Statements that are mutually exclusive are such that one or the other statement must be true. They cannot both be true at the same time. For instance:

Something is either “A” or “not A”. It cannot be both “A” and “not A” at the same time.

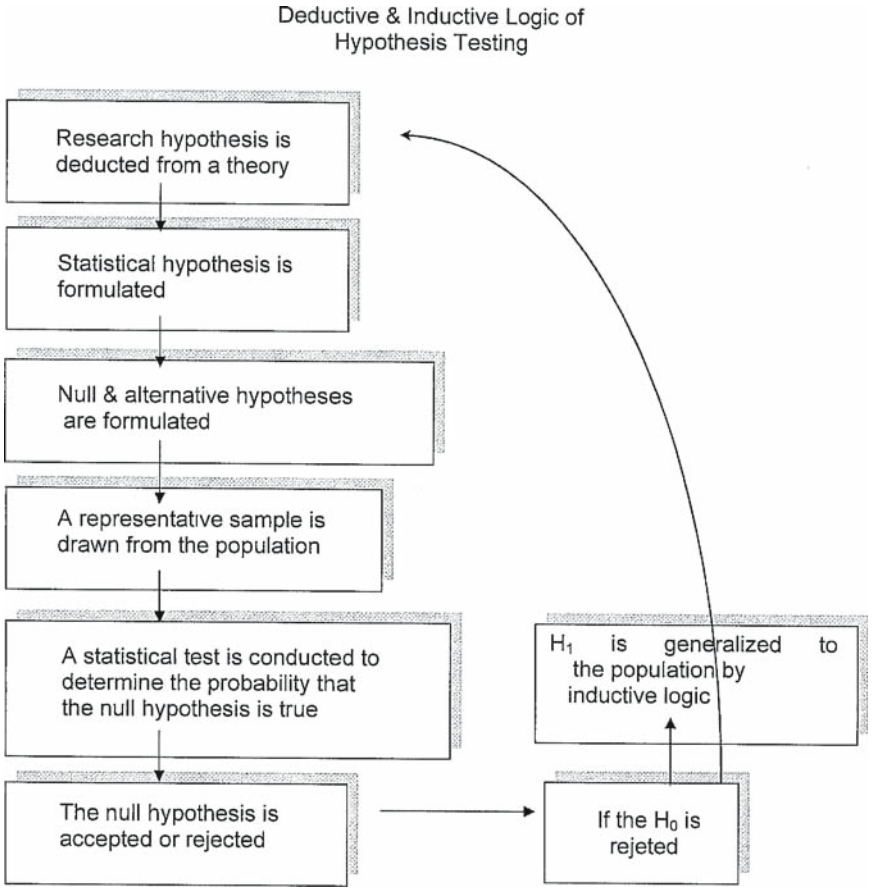



**Fig. 2.2** Deductive and inductive logic of hypothesis testing

For instance, the object on the kitchen table is either an apple or a non-apple. Saying the object on the kitchen table is either an "apple" or a "non-apple" covers every possible thing that the object could be.

It is critical to understand that it is the null hypothesis ($H_0$) that is actually tested when the data are statistically analyzed, not the alternative hypothesis ($H_1$). Since $H_0$ and $H_1$ are mutually exclusive, if the analysis of the data leads to the rejection of the null hypothesis ($H_0$), the only tenable alternative is to accept the alternative hypothesis ($H_1$). But, this does not mean that the alternative hypothesis is true, it may or may not be true. When we reject the null hypothesis it is because there is only a remote possibility that the sample could have come from a population in which the null hypothesis is true. Could we be wrong? Yes, and that probability is called alpha ($\alpha$), and the error associated with alpha is called a Type I error.

What about the converse situation, accepting the null hypothesis? If the null hypothesis is accepted, the alternative hypothesis may or may not be false. For example, the null hypothesis may be accepted because the sample size was too small to achieve the required degrees of freedom for statistical significance; or, an uncontrolled extraneous variable or spurious variable has masked the true relationship between the variables; or, that the measures of the variables involved are grossly unreliable, etc. The issue is the same as a not guilty verdict in a criminal trial. That is, a verdict of not guilty does not necessarily mean that the defendant is innocent, it only means that the evidence was not sufficient enough to establish guilt beyond a reasonable doubt. There is a further discussion about the null hypothesis in Chapter 18.

## An Overview of the Common Clinical Research Designs

The common clinical research designs are listed in Table 2.2. There are many ways to classify study designs but two general ways are to separate them into descriptive and analytic studies and observational and experimental studies. These designations are fairly straight-forward. In descriptive studies one characterizes a group of subjects; for example 'we describe the characteristics of 100 subjects taking prophylactic aspirin in the stroke belt.' In contrast, with analytic studies there is a comparator group. In experimental studies the investigator is controlling the intervention in contrast to observational studies where the exposure (intervention) of interest is occurring in nature and as the investigator you are observing the subjects with and without the exposure. Basically, experimental trials are clinical trials, and if subjects are randomized into the intervention and control (comparator) groups it is a RCT.

### *Ecologic Studies*

Ecologic studies use available population data to determine associations. For example, to determine an association between coronary heart disease (CHD) and the intake of saturated fat, one could access public records of beef sales in different

**Table 2.2** Overview – study types

| Observational | Experimental |
|---|---|
| • Ecological studies | • Clinical trials |
| • Case reports | • Group trials |
| • Case series | |
| • Cross-sectional studies | |
| • Case-control studies | |
| • Cohort studies | |

states (or counties or regions of the country) and determine if an association existed between sales and the prevalence of CHD.

## *Case Reports and Case Series*

Case reports and case series are potential ways to suggest an association, but, although limited in this regard, should not be deemed unimportant. For example, the recognition of the association of the diet drug combination of Fen-phen was the result of a case series.[5]

## *Cross-Sectional Studies*

In cross-sectional studies, one defines and describes disease status (or outcome), exposure(s), and other characteristics at a point in time (point in time is the operative phrase), in order to evaluate associations between them.

Cross-sectional studies are different from cohort studies in that the latter observe the association between a naturally occurring exposure and outcome (e.g., between health and a disease or between disease and an event) over a period of time rather than at a point in time). With cross-sectional studies, the exposure and outcome are evaluated at a point in time – i.e. there is no follow-up period. Indeed that is both the strength and weakness of the cross-sectional (X-sectional) study design. Lack of a follow-up period means the study can be performed more rapidly and less expensively than a cohort study, but one sacrifices temporality (an important component for determining causality). In addition, because X-sectional studies are evaluating cases (disease, outcomes) at a point in time, one is dealing with prevalent cases (not incident cases as is true of a cohort study). There are a number of factors that must be considered when using prevalence (rather than incidence) and these are summarized in Fig. 2.3.

## *Case-Control Study*

In a case-control study (CCS), the investigator identifies a certain outcome in the population, then matches the 'diseased group' to a 'healthy group,' and finally identifies differences in exposures between the two groups.

With a CCS one approaches the study design the opposite of a cohort design (in fact some have suggested the use of the term 'trohoc design' – cohort spelled backwards). The term case-control study was coined by Sartwell to overcome the implication that the retrospective nature of the design was an essential feature.[6] That is, patients with the outcome of interest are identified, a control group is selected, and

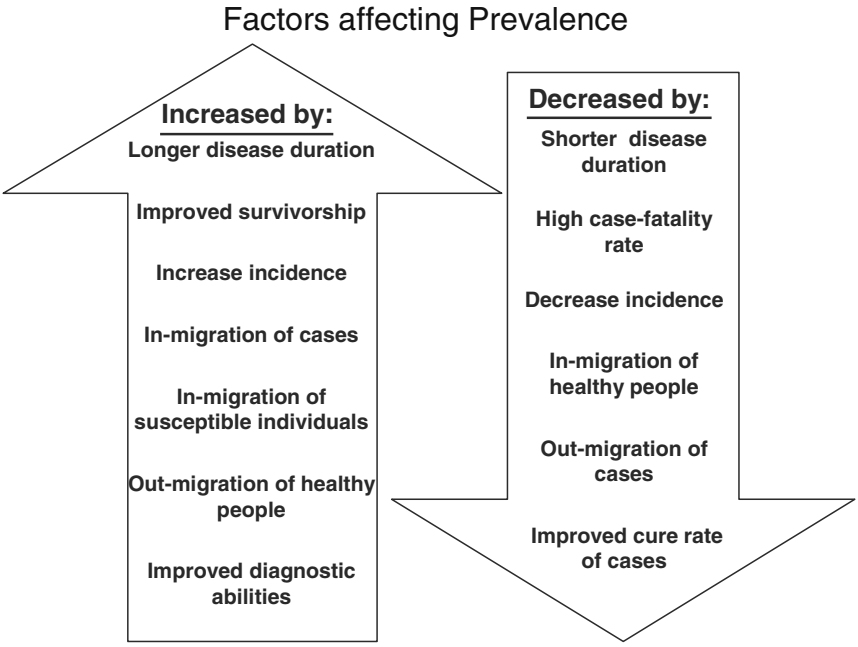



**Fig. 2.3** Factors affecting prevalence

one then looks back for exposures that differ between the two. Two major biases exist with the CCS; first the selection of the control group is problematic, and second, one is looking back in time (i.e. it is a retrospective study in that sense). Selecting the control group for a CCS is problematic because if one selects too many matching criteria it becomes difficult to find an adequate control group, while if one has too few matching criteria, the two groups can differ in important variables. For CCS designs, recall bias is also an issue (this is even a greater issue if death is an outcome, in which case one not only has to deal with recall bias, but the recall is obtained from family members, caregivers, etc.).

One of the strengths of the CCS design is that if one is interested in a rare disease, one can search the area for those cases, in contrast to randomly selecting a cohort population which will develop this rare disease infrequently, even over a long follow-up time period. Also, in contrast to a cohort study in which the sample population is followed for a time period, a CCS obviates this need so one can complete the study much sooner (and therefore less expensively).

There are several variations of the case-control design that overcome some of the shortcomings of a typical CCS (although they have their own limitations): a prospective CCS and a nested CCS. In the prospective CCS, one accrues the cases over time (i.e. in a prospective fashion) so that recall bias is less of an issue. However, one then has to wait until enough cases are accrued (problematic again for rare diseases); and, the selection of an appropriate control group still exists. A nested case-control

study is a type of study design where outcomes that occurred during the course of a cohort study or RCT are compared to controls selected from the cohort population who did not have he outcome. Compared with the typical case-control study, a nested case-control study can reduce 'recall bias' and temporal ambiguity, and compared with a cohort study, it can reduce cost and save time. One drawback of a nested case-control study is that the non-diseased persons from whom the controls are selected may not be fully representative of the original cohort as a result of death or failure to follow-up cases. As mentioned, the nested CCS design can be placed within a cohort study or RCT. An example is taken from the Cholesterol and Recurrent Events (CARE) Study.[7] The primary study was aimed at the prevention of recurrent MI when patients with a prior MI and 'normal' cholesterol levels were further treated with pravastatin. As part of the original study plasma was stored and after the report of the primary study was published the following nested CCS was designed: Patients with recurrent MI were identified and age and sex matched with subjects in the study without recurrent MI. The plasma was then analyzed for components of large and small LDL-C and associations with recurrent MI were determined.

## *Cohort Study*

A cohort study is much like a RCT except that the intervention in an RCT is investigator controlled, while in a cohort study the interventions is a naturally occurring phenomenon. A cohort design is a study in which two or more groups of people that are free of disease and that differ according to the extent of exposure (e.g. exposed and unexposed) are compared with respect to disease incidence. A cohort study assembles a group of subjects and follows them over time. One follows these subjects to the development of an outcome of interest and then compares the characteristics of the subjects with and without the outcome in order to identify risk factors (exposures) for that outcome. A major assumption made in cohort studies is that the subject is disease free at the beginning of the study (disease free means for the outcome of interest). For example, if the outcome of interest is a *recurrent* myocardial infarction, the subject would have had the first infarction (so in that sense he is not disease free) but in terms of the outcome of interest (a second infarction) we assume that at study onset, he is not having a second infarction. This example may seem obvious, but let us use colon cancer as another example. At study onset, one assumes that the subject is disease free (cancer-free or 'normal') at the time of enrollment, while in fact he or she may already have colon cancer that is as yet undiagnosed. This could bias the results of the study since the exposure of interest may have nothing to do with the outcome of interest (colon cancer) since the subject already has the outcome irrespective of the exposure (say a high fat diet). This also raises the issue as to what is 'normal'. One whit suggested that a normal subject is one that has been insufficiently tested! The cohort assumption mentioned above is diagrammed in Fig. 2.4. Of course, one also assumes that the

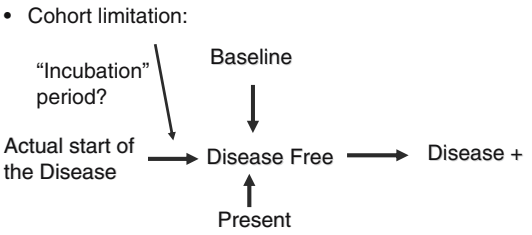



**Fig. 2.4** Cohort limitation

incorrect assumption of no disease at onset is equally balanced in the two groups under study, and that is indeed the hope, but not always the realization. Cohort studies are considered the best way to study prognosis, but one can also do this by using a case-control design.

Cohort studies are generally prospective; however, retrospective cohort studies do exist. The key to the study design is identifying 'normal' subjects without disease (i.e. the outcome of interest), evaluate for that outcome after a period of time has elapsed, and determining factors that are different in those with and without the outcome. Retrospective cohort studies are particularly well suited to the study of long-term occupational hazards. An example of a retrospective cohort study is the study of nickel refinery workers where about 1,000 nickel refinery workers were identified from company records and their outcomes identified over a prior 10 year period. Sixteen were found to have died from lung cancer (expected rate was 1 from National data), 11 died from nasal cancer (1 expected) and 67 from other causes (72 expected).[8]

Another modification of cohort studies is the case-cohort design. With the case-cohort design, a 'subcohort' is randomly selected from the cohort sample, a separate exposure of interest from the total cohort is identified, and cases (outcomes) are then determined in the same manner as the primary design. An example might be a cohort study of 10,000 subjects that is assessing some outcome – let's say a CVD outcome – in relation to dietary fat. The investigator decides that she would also like to know the association of CVD with a measure of coronary artery calcium, so electron beam computed tomography (EBCT – a relatively expensive procedure to perform on the all of the original cohort) is measured in a random sample of 100 of the cohort subjects (the 'subcohort'). The association of EBCT to CVD outcome is then ultimately determined.

## *Randomized Control Trial (RCT)*

In the randomized-controlled trial (RCT), the exposure is controlled by the investigator, which contrasts it to all the other study designs. A detailed discussion of the RCT will be presented in Chapter 3. However, it should be noted that RCTs cannot be used to address all important questions. For example, observational studies are

more appropriate when studies are used to detect rare or late consequences of interventions.

## Discussion

One should now be able to begin to understand the key differences, and therefore limitations, of each study design; and, circumstances where one design might be preferable to another. Let's, for example, use the exposure of electromagnetic energy (EME) and cancer outcome (e.g. leukemia). With a cross-sectional study, a population is identified (target population), cancer rates determined, and exposure and lack of exposure to EME is ascertained in a sample. One then analyzes the exposure rates in subjects with cancer and those that are cancer free. If the cancer rate is higher in those who were exposed, an association is implied. This would be a relatively inexpensive way to begin to look at the possible association of these variables, but limitations should be obvious. For example, since there is no temporality in this type of design, and since biologically, exposure to EME if it did cause cancer would likely have to occur over a long period of time, one could easily miss an association.

In summary, it should be evident that observational studies (e.g. cross-sectional, case-control, and cohort studies) have a major role in research. However, despite their important role, von Elm et al. discussed the lack of important information that was either missing or unclear in prior published observational studies; and why this lack of information lead to a guideline document for reporting observational studies (the STROBE statement – the Strengthening and Reporting of Observational Studies in Epidemiology). The STROBE statement was designed after the CONSORT – the Consolidated Standards of Reporting Trials –; this statement outlines the guidelines for reporting RCTs. The STROBE statement is a checklist of 22 items that are to be considered essential for good reporting of observational studies.[9]

## References

1. Parker_Palmer. http://en.wikipedia.org/wiki/Parker_Palmer
2. Vickers AJ. Michael Jordan won't accept the null hypothesis: notes on interpreting high P values. *Medscape*. 2006; 7(1).
3. The Null Logic of Hypothesis Testing. http://www.shsu.edu/~icc_cmf/cj_787/research6.doc
4. Blackstone. Cited in The Null Logic of Hypothesis Testing. 2 Bl. Com. C. 27, margin page 358, ad finem. Available at: http://www.shsu.edu/~icc_cmf/cj_787/research6.doc
5. Connolly HM, Crary JL, McGoon MD, et al. Valvular heart disease associated with fenfluramine-phentermine. *N Engl J Med*. Aug 28, 1997; 337(9):581–588.
6. Cited in Sartwell P and Nathanson N. Epidemiologic Reviews. 1993.
7. Sacks FM, Pfeffer MA, Moye LA, et al. The effect of pravastatin on coronary events after myocardial infarction in patients with average cholesterol levels. Cholesterol and recurrent events trial investigators. *N Engl J Med*. Oct 3, 1996; 335(14):1001–1009.

8. Doll R. Cohort studies: history of the method. II. Retrospective cohort studies. *Soz Praventivmed* 2001; 46(3):152–160.
9. von Elm E, Altman DG, Egger M, Pocock SJ, Gotzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *Ann Intern Med*. Oct 16, 2007; 147(8):573–577.

# Chapter 3
# Clinical Trials*

**Stephen P. Glasser**

**Abstract** The spectrum of evidence imparted by the different clinical research designs ranges from ecological studies through observational epidemiological studies to randomized control trials (RCTs). This chapter addresses the definition of clinical research, the major aspects of clinical trials eg ethics, randomization, masking, recruitment and retention of subjects enrolled in a clinical trial, patients/subjects lost to follow-up during the trial etc. Although this chapter focuses on the weaknesses of clinical trials, it is emphasized that the randomized, placebo-controlled, double blind clinical trial is the design that yields the greatest level of scientific evidence.

*A researcher is in a gondola of a balloon that loses lift and lands in the middle of a field near a road. Of course, it looks like the balloon landed in the middle of nowhere. As the researcher ponders appropriate courses of action, another person wanders by. The researcher asks, 'Where am I?' The other person responds, 'You are in the gondola of a balloon in the middle of a field.' The researcher comments, 'You must design clinical trials.' 'Well, that's amazing, how did you know?' 'Your answer was correct and precise and totally useless.'*

## Introduction

The spectrum of evidence imparted by the different clinical research designs ranges from ecological studies through observational epidemiological studies to randomized control trials (RCTs). The differences in clinical research designs and the different weights of evidence are exemplified by the post-menopausal hormone replacement therapy (HRT) controversy. Multiple observational epidemiological studies had shown that HRT was strongly associated with the reduction of atherosclerosis, myocardial infarction risk, and stroke risk.[1-3] subsequently, 3 clinical trials suggested that HRT was not beneficial, and might even be harmful.[4-6] This latter observation raises a number of questions, including: why can this paradox occur? what can contribute to this disagreement?; and, why do we believe these 3 RCT's more than so many well-done observational trials?

* Over 50% of this chapter is taken from "Clinical trial design issues: at least 10 things you should look for in clinical trials"[7] with permission of the publisher.

S.P. Glasser (ed.), *Essentials of Clinical Research*,

Before addressing these above questions it is appropriate to point out that frequently, there is confusion about the difference between clinical research and clinical trials. A clinical trial is one type of clinical research. A clinical trial is a type of experimental study undertaken to assess the response of an individual (or in the case of group clinical trials-a population) to interventions introduced by an investigator. Clinical trials can be randomized or non-randomized, un-blinded, single-blinded, or double-blinded; comparator groups can be placebo, active controls, or no treatment controls, and RCTs can have a variety of designs (eg parallel group, crossover, etc.). That being said, the RCT remains the 'gold-standard' study design and its results are appropriately credited as yielding the highest level of scientific evidence (greatest likelihood of causation). However, recognition of the limitations of the RCT is also important so that results from RCTs are not blindly accepted. As Grimes and Schultz point out, in this era of increasing demands on a clinician's time it is 'difficult to stay abreast of the literature, much less read it critically. In our view, this has led to the somewhat uncritical acceptance of the results of a randomized clinical trial'.[8] Also, Loscalzo, has pointed out that 'errors in clinical trial design and statistical assessment are, unfortunately, more common that a careful student of the art should accept'.[9]

What leads the RCT to the highest level of evidence and what are the features of the RCT that renders it so useful? Arguably, one of the most important issues in clinical trials is having matched groups in the interventional and control arms; and, this is best accomplished by randomization. That is, to the degree that the 2 groups under study are different, results can be confounded, while when the 2 groups are similar, confounding is reduced (See chapter 17 for a discussion of confounding). It is true that when potential confounding variables are known, one can relatively easily adjust for them in the design or analysis phase of the study. For example, if smoking might confound the results of the success of treatment for hypertension, one can build into the design a stratification scheme that separates smokers form non-smokers, before the intervention is administered and in that way determine if there are differential effects in the success of treatment (e.g. smokers and non-smokers are randomized equally to the intervention and control). Conversely, one can adjust after data collection in the analysis phase by separating the smokers from the non-smokers and again analyze them separately in terms of the success of the intervention compared to the control. The real challenge of clinical research, is not how to adjust for known confounders, but how to have matched (similar groups- how to adjust) in the intervention and control arms, when potential confounders are **not** known. Optimal matching is accomplished with randomization, and this is why randomization is so important. More about randomization later, but in the meanwhile one can begin to ponder how un-matching might occur even in a RCT. In addition to randomization, there are a number of important considerations that exist regarding the conduct of a clinical trial, such as: is it ethical? what type of comparator group should be used? what type of design and analysis technique will be utilized? how many subjects are needed and how will they be recruited and retained? etc.

Finally, there are issues unique to RCTs (eg intention-to-treat analysis, placebo control groups, randomization, equivalence testing) and issues common to all clinical research (eg ethical issues, blinding, selection of the control group, choice

**Table 3.1** Issues of importance for RCTs

| |
|---|
| Ethical considerations |
| Randomization |
| Eligibility criteria |
| Efficacy vs effectiveness |
| Compliance |
| Run-in periods |
| Recruitment and retention |
| Masking |
| Comparison groups |
| Placebo |
| 'Normals' |
| Analytical issues |
| ITT |
| Subgroup analysis |
| Losses to follow-up |
| Equivalence vs traditional testing |
| Outcome selection |
| Surrogate endpoints |
| Composite endpoints |
| Trial duration |
| Interpretation of results |
| Causal inference |
| The media |

of the outcome/endpoint, trial duration, etc) that must be considered. Each of these issues will be reviewed in this chapter (Table 3.1). To this end, both the positive and problematic areas of RCTs will be highlighted.

## Ethical Issues

Consideration of ethical issues is key to the selection of the study design chosen for a given research question/hypothesis. For RCTs ethical considerations can be particularly problematic, mostly (but by no means solely) as it relates to using a placebo control. A full discussion of the ethics of clinical research is beyond the scope of this book, and for further discussion one should review the references noted here.[10-12] (There is also further discussion of this issue under the section entitled Traditional vs. Equivalence Testing and Chapters 4 and 7). The opinions about when it is ethical to use placebo controls is quite broad. For example, Rothman and Michaels are of the opinion that the use of placebo is in direct violation of the Nuremberg Code and the Declaration of Helsinki,[12] while others would argue that placebo controls are ethical as long as withholding effective treatment leads to no serious harm and if patients are fully informed. Most would agree that placebo is unethical if effective life-saving or life-prolonging therapy is available or if it is likely that the placebo group could suffer serious harm. For ailments that are not likely to be of harm or cause severe discomfort, some would argue that placebo is justifiable.[11] However, in the majority of scenarios, the use of a placebo control

is not a clear-cut issue, and decisions need to be made on a case-by-case basis. One prevailing standard that provides a guideline for when to study an intervention against placebo is when one has enough confidence in the intervention that one is comfortable that the additional risk of exposing a subject to the intervention is low relative to no therapy or the 'standard' treatment; but, that there is sufficient doubt about the intervention that use of a placebo or active control ('standard treatment') is justified. This balance, commonly referred to as *equipoise*, can be difficult to come by and is likewise almost always controversial. Importantly, equipoise needs to be present not only for the field of study (i.e. there is agreement that there is not sufficient evidence of the superiority of an alternative treatments), but equipoise also has to be present for individual investigators (permitting individual investigators to ethically assign their patients to treatment at random).

Another development in the continued efforts to protect patient safety is the Data Safety and Monitoring Board (DSMB-see chapter 9). The DSMB is now almost universally used in any long-term intervention trial. First a data and safety monitoring plan (DSMP) becomes part of the protocol, and then the DSMB meets at regular and at 'as needed' intervals during the study in order to address whether the study requires early discontinuation. As part of the DSMP, stopping rules for the RCT will have been delineated. Thus, if during the study, either the intervention or control group demonstrates a worsening outcome, or the intervention group is showing a clear benefit, or adverse events are greater in one group vs the other (as defined within the DSMP) the DSMB can recommend that the study be stopped. But, the early stopping of studies can also be a problem. For example, in a recent systematic review by Montori et al, the question was posed about what was known regarding the epidemiology and reporting quality of RCTs involving interventions stopped for early benefit.[13] Their conclusions were that prematurely stopped RCTs often fail to adequately report relevant information about the decision to stop early, and that one should view the results of trials that are stopped early with skepticism. [13]

## Randomization

Arguably, it is randomization that results in the RCT yielding the highest level of scientific evidence (i.e. resulting in the greatest likelihood that the intervention is causally related to the outcome). Randomization is a method of treatment allocation that is a distribution of study subjects at random (i.e. by chance). As a result, randomization results in all randomized units (e.g. subjects) having the same and independent chance of being allocated to any of the treatment groups, and it is impossible to know in advance to which group a subject will be assigned. The introduction of randomization to clinical trials in the modern era can probably be credited to the 1948 trial of streptomycin for the treatment of tuberculosis (Fig. 1.1).[14] In this trial, 55 patients were randomized to either streptomycin with bed rest, or to treatment with bed rest alone (the standard treatment at that time). To quote from that paper, 'determination *of whether a patient would be treated by streptomycin and bed rest (S case)*

*or bed rest alone (C case), was made by reference to a statistical series based on random sampling numbers drawn up for each sex at each center by Professor Bradford Hill; the details of the series were unknown to any of the investigators or to the co-coordinator and were contained in a set of sealed envelopes each bearing on the outside only the name of the hospital and a number. After acceptance of a patient by the panel and before admission to the streptomycin centre, the appropriate numbered envelope was opened at the central office; the card inside told if the patient was to be an S or C cases, and this information was then given to the medical officer at the centre'.* Bradford Hill was later knighted for his contributions to science including the contribution of randomization.

With randomization the allocation ratio (number of units-subjects- randomized to the investigational arm versus the number randomized to the control arm) is usually 1:1. But a 1:1 ratio is not required, and there may be advantages to unequal allocation (e.g. 2:1 or even 3:1). The advantages of unequal allocation are: one exposes fewer patients to placebo, and one gains more information regarding the safety of the intervention. The main disadvantage of higher allocation ratios is the loss of power.

There are 3 general types of randomization: simple, blocked, and stratified. Simple randomization can be likened to the toss of an unbiased coin-ie heads group A, tails group B. This is easy to implement, but particularly with small sample sizes, could result in substantial imbalance (for example if one tosses a coin 10 times, it is not improbable that one could get 8 heads and 2 tails. If one tosses the coin 1000 times it is likely that the distribution of heads to tails would be close to

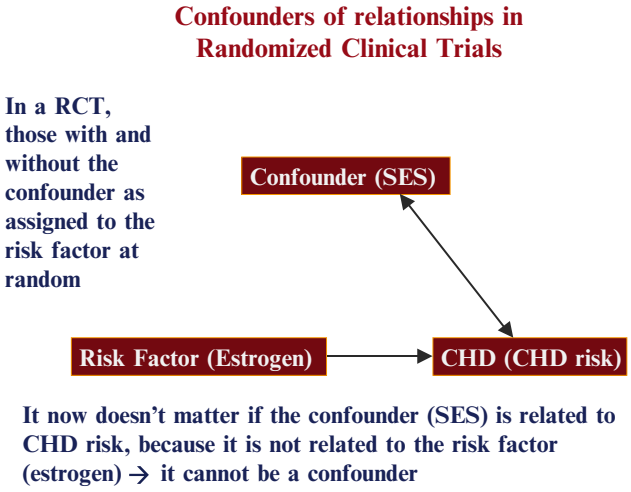



**Fig. 3.1** The relationship of confounders to outcome and how they are eliminated in a RCT

500 heads and 500 tails). Blocked randomization (sometimes called permuted block randomization) is a technique common to multi-center studies. Whereas the entire trial might intend to enroll 1000 patients, each center might only contribute 10 patients to the total. To prevent between center bias (recall each sample population has differences even if there is matching to known confounders) blocked randomization can be utilized. Blocked randomization means that randomization occurs within each center ensuring that about 5 patients in each center will be randomized to the intervention and 5 to the control. If this approach was not used, one center might enroll 10 patients to the intervention and another center, 10 patients to the control group. Recall that the main objective of randomization is to produce between-group comparability. If one knows prior to the study implementation that there might be differences that are not equally distributed between groups (again particularly more likely with small sample sizes) stratified randomization can be used. For example, if age might be an important indicator of drug efficacy, one can randomize within strata of age groups (e.g. 50–59, 60–69 etc.). Within each stratum, randomization can be simple or blocked.

In review, simple randomization is the individual allocation of subjects into the intervention and control groups, block randomization creates small groups (blocks) in which there are equal numbers in each treatment arm so that there are balanced numbers throughout a multi-center trial, and stratified randomization addresses the ability to separate known confounders into strata so that they can no longer confound the study results. Again, randomization is likely the most important key to valid study results because (if the sample size is large enough), it distributes known, and ***more importantly unknown***, confounders equally to the intervention and control groups.

Now, as to the problems associated with randomization. As prior discussed, the issue of confounders of relationships is inherent in all clinical research. A confounder is a factor that is associated with both the risk factor and the outcome, and leads to a false apparent association between the risk factor and outcome (See Fig. 3.2). In observational studies, there are two alternative approaches to remove the effect of confounders:.

- Most commonly used in case/control studies, one can match the case and control populations on the levels of potential confounders. Through this matching the investigator is assured that both those with a positive outcome (cases) and a negative outcome (controls) have similar levels of the confounder. Since, by definition, a confounder has to be associated with both the risk factor and the outcome; and, since through matching the suspected confounder is not associated with the outcome – then the factor cannot affect the observed differences in the outcome. For example, in a study of stroke, one may match age and race for stroke cases and community controls, with the result that both those with and without strokes will have similar distributions for these variables, and differences in associations with other potential predictors are not likely to be confounded, for example, by higher rates in older or African American populations.

- In all types of observational epidemiological studies, one can statistically/mathematically 'adjust' for the confounders. Such an adjustment allows for the comparison between those with and without the risk factor at a 'fixed level' of the confounding factor. That is, the association between the exposure and the potential confounding factor is removed (those with and without the exposure are assessed at a common level of the confounder), and as such the potential confounder cannot bias the association between the exposure and the outcome. For example, in a longitudinal study assessing the potential impact of hypertension on stroke risk, the analysis can 'adjust' for race and other factors. This adjustment implies that those with and without the exposure (hypertension) are assessed as if race were not associated with both the exposure and outcome.

The major shortcoming with either of these approaches is that one must know what the potential confounders are in order to match or adjust for them; and, it is the ***unknown confounders*** that represent a bigger problem. Another issue is that even if one suspects a confounder, one must be able to appropriately measure it. For example, a commonly addressed confounder is socio-economic status (usually a combination of education and income); but, clearly this is an issue in which there is disagreement and, which measure or cut-point is appropriate. The bottom line is that one can never perfectly measure all known confounders and certainly one cannot measure or match for unknown confounders. As mentioned, the strength of the RCT is that randomization (performed properly and with a large enough sample size) balances both the known and unknown confounders between the interventional and control groups. But even with an RCT, randomization can be further compromised as will be discussed in some of the following chapters, and by the following example from "Student's" Collected Papers regarding the Lanarkshire Milk Experiment:[15]

> "Student" (ie, the great William Sealy Gosset) criticized the experiment for it's loss of control over treatment assignment. As quoted: ... Student's " contributions to statistics, in spite of a unity of purpose, ranged over a wide field from spurious correlation to Spearman's correlation coefficient. Always kindly and unassuming, he was capable of a generous rage, an instance of which is shown in his criticism of the Lancashire Milk Experiment. This was a nutritional experiment on a very large scale. For four months 5,000 school children received three-quarters of a pint of raw milk a day, 5,000 children the same quantity of pasteurized milk and 10,000 other children were selected as controls. The experiment, in Gosset's view, was inconclusive in determining whether pasteurized milk was superior in nutritional value to raw milk.
>
> This was due to failure to preserve the random selection of controls as originally planned. "In any particular school where there was any group to which these methods (i.e., of random selection) had given an undue proportion of well-fed or ill-nourished children, others were substituted to obtain a more level selection." The teachers were kind-hearted and tended to select ill-nourished as feeders and well-nourished as controls. Student thought that among 20,000 children some 200–300 pairs of twins would be available of which some 50 pairs would be identical-of the same sex and half the remainder nonidentical of the same sex. The 50 pairs of identicals would give more

> reliable results than the 20,000 dealt with in the experiment, and great expense would be saved. It may be wondered, however, whether Student's suggestion would have proved free from snags. Mothers can be as kind-hearted as teachers, and if one of a pair of identical twins seemed to his mother to be putting on weight...

## Implications of Eligibility Criteria

In every study there are substantial gains in statistical power by focusing the intervention in a homogenous patient population likely to respond to treatment, and to exclude patients that could introduce 'noise' by their inconsistent responses to treatment. Conversely, at the end of a trial there is a need to generalize the findings to a broad spectrum of patients who could potentially benefit from the superior treatment. These conflicting demands introduce an issue of balancing the inclusion/exclusion (eligibility criteria) such that the enrolled patients are as much alike as possible; but, on the other hand to be as diverse as possible in order to be able to apply the results to the more general population (i.e. generalizability). Fig. 3.2 outlines this balance. What is the correct way of achieving this balance? There really is no correct answer, there is always a tradeoff between homogeneity and generalizability; and each study has to address this, given the availability of subjects, along with other considerations. This process of sampling represents one of the reasons that scientific inquiry requires reproducibility of results, that is, one study generally cannot be relied upon to portray 'truth' even if it is a RCT. The process of sampling embraces the concept of generalizability. The issue of generalizability is nicely portrayed in a video entitled 'A Village of 100'.[16] If one

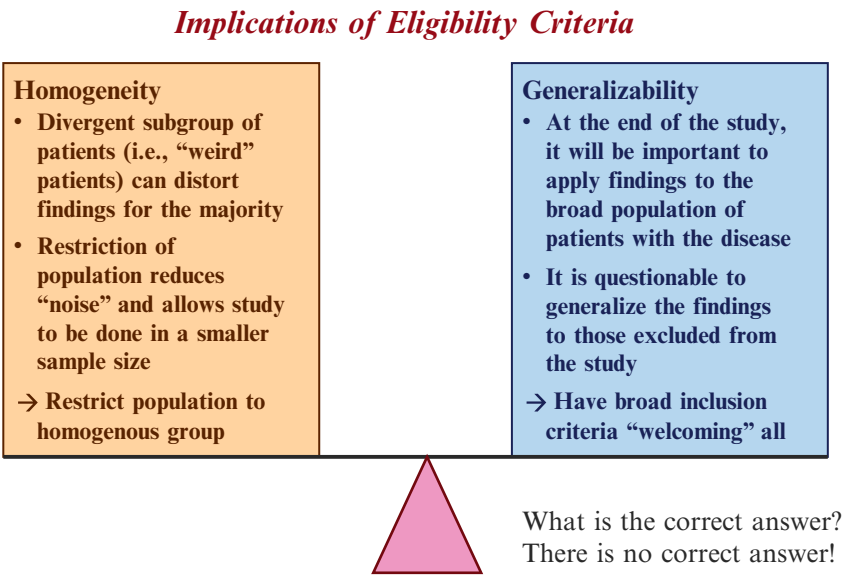



**Fig. 3.2** The balance of conflicting issues involved with patient selection

wanted to have a representative sample of the world for a study, this video (although predominately focused upon tolerance and understanding), is an excellent way of understanding the issue of generalizability. The central theme of the video asks the question 'if we shrunk the earth's population to a village of precisely 100 people, with all existing ratios remaining the same, what would it look like?' To paraphrase, if we maintained the existing ratios of the earth's population in a study of 100 people, what would our sample look like? The answer–there would be 57 Asians, 21 Europeans, 14 from the Western Hemisphere, 51 females and 49 males, 70 non-white and 30 white, 70 non Christians and 30 Christians, 89 heterosexuals, 50% of the worlds wealth would belong to 6 citizens of the USA, 80 would live in sub-standard housing, 70 would be unable to read (a potential problem with IRB approval), 50 would be malnourished, 1 would have a college education, and 4 would own a computer. When is the last time a study had a population representative of the Village of 100?

For an example of sampling issues, most of the major studies assessing the efficacy of the treatment of extracranial atherosclerosis with endarterectomy had excluded octogenarians on the basis that this patient population may have a response to the challenges of surgery that is different than their younger counterparts.[17, 18] Exclusion of these patients may have contributed to the successful completion of 'positive' trials (finding a benefit for the new treatment – endarterectomy). However, now that the trials are complete, there is not 'level 5' evidence (data that is a result from RCTs) to guide the management of octogenarians with extracranial atherosclerosis, one of the subpopulations where the need for this information is important. In the absence of this information, thousands of endarterectomies are performed in this older patient population each year under the assumption that the findings from a younger cohort are generalizable to those at older ages. For another example, let's presume that in a multicenter trial that included Framingham Mass., and Birmingham, AL, that a representative sample of each was recruited into a study. The makeup of the sample from each is illustrated in Table 3.2. As one can see, there are significant

**Table 3.2** Birmingham vs Framingham: comparison of key variables

| | Birmingham | Framingham |
|---|---|---|
| Population | 242,800 | 62,910 |
| % African-American | 73.5 | 5.1 |
| Age | | |
| 25–44 | 30 | 35 |
| 45–64 | 20 | 22 |
| 65-> | 14 | 13 |
| Median Income $ | 26,700 | 55,300 |
| Education | | |
| <High School | 25 | 13 |
| High School | 28 | 23 |
| >High School | 48 | 64 |
| CVD | 528–582 | 336–451 |

differences in the representative sample populations, and these differences could affect not only the success of the intervention or confound its relationship.

## Efficacy vs Effectiveness

Another limitation of RCTs is that they are designed to test safety and efficacy (i.e. does the drug work under optimal circumstances?) and not to answer questions about the effectiveness of a drug, the more relevant question for clinicians and economic analysts (i.e. does the drug work under ordinary circumstances of use?). Thus, the increased use of effectiveness trials has been suggested, to more closely reflect routine clinical practice. Effectiveness trials use a more flexible dosage regimen, and a 'usual care' comparator instead of a placebo comparator. (Two approaches to this more 'real world trial' is the phase 4 trial- see Chapter 5) or the prospective, randomized, open-label, blinded end-point –PROBE-Trial. The PROBE Trial is further discussed in the next section entitled Degree of Masking). As to phase 4 trials, they are surrounded by some controversy as well. Fig. 3.3 compares efficacy and effectiveness trials in terms of some of their more important variables.

## Patient Compliance

### *Run-in Periods*

Another issue surrounding RCTs, and one which is almost unique to clinical trials, is the use of run-in periods and their impact on who is eligible to be randomized.

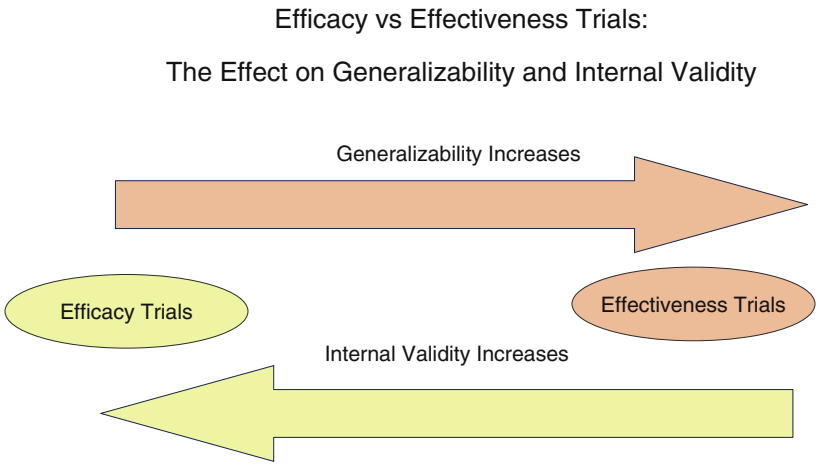



**Fig. 3.3** Efficacy vs Effectiveness

Pre-randomization run-in periods are frequently used to select or exclude patients in clinical trials, but the impact of run-in periods on clinical trial interpretation and generalization has not been systematically studied. The controversy regarding run-in periods also addresses the issue of efficacy vs. effectiveness, as the run-in period allows one to exclude patients that are less compliant, or do not tolerate placebo (or whatever other intervention is used in the active comparison group). Although this issue has not been systematically studied, intuitively one can see that the potential for over-estimating the impact of an investigational drug is present when run-in periods are utilized, as the run-in period will likely exclude patients from the study who would not have ideally responded.

A study can achieve high compliance in at least 3 general ways: designing a simple protocol (complexity makes compliance more difficult); the use of compliance aids such as automatic reminders, telephone calls, calendars, etc; or by selecting subjects based upon pre-study or pre-randomization compliance. Of course, high compliance is a desirable characteristic of any research. High compliance attenuates the argument of whether to use intention to treat vs. compliance only as the primary analysis. Also, high compliance will optimize the studies power as the diluting effect of non-compliers will not be manifest (all other things being equal). While the run-in period increases the proportion of compliers in the trial, it may introduce important differences in the outcomes, particularly if compliers and non-compliers are inherently different in the way they would respond to the intervention of interest. Thus, the effect of run-in periods on generalizability should be considered carefully before implementation. Lang[19] has listed some recommendations for helping to decide whether to use a run-in as part of a clinical trial, including:

1. Consider a run-in whenever the contact between study staff and participants is low
2. Consider a run-in period for a primary prevention trial because compliance is likely to be harder compared to therapeutic trials
3. For any trial, list the key features of the study protocol and see which features compliance could be directly tested prior to randomization
4. Before using active agents during a run-in, consider both the expected frequency of occurrence of side effects and the postulated effect of the agent on the outcome of interest
5. All trials can use any available pre-randomization period for the simultaneous purpose of characterizing patients and evaluating compliance, whether of not the compliance information will be used for exclusions

In fairness, as Franciosa points out, clinicians use variants of run-in periods to treat their patients, such as dose titration, or challenge dosing (such as using small doses of ACE Inhibitors to rule out excessive responders). Pablos-Mendez et al, analyzed illustrative examples of reports of clinical trials in which run-in periods were used to exclude non-compliant patients, placebo responders, or patients that could not tolerate or did not respond to active drug.

Thus, the use of run-in periods is another reason that the results of RCTs may not accurately portray what the drugs overall effectiveness will be. What can be said is that there does need to be more focus on the details of run-in periods, and as

is true of most things the researcher does in designing and implementing a clinical trial, judgments have to be made regarding the best approach to use regarding inclusions and exclusions, as well as judging what the impact of the run-in period is on the ultimate interpretation of a clinical trial.

## Recruitment and Retention

Nothing is more critical to the success of a clinical trial than the recruitment and retention of subjects. As will be discussed in more detail in Chapter 8, there are a number of reasons for failure of the recruitment process including: delayed start-up, and inadequate planning, In terms of patient/subject retention, there are arguably differences in the handling of clinical patients in contrast to research subjects (although this could and perhaps should be challenged). Losses-to-follow-up need to be kept to a minimum and is discussed later in this chapter.

## Degree of Masking (Blinding)

Although the basic concept of clinical trials is to be at equipoise, this does not change the often pre-conceived 'suspicion' that there is a differential benefit of the investigational therapy (e.g. the investigational drug is better than placebo). Thus, if study personnel know the treatment assignment, there may be differential vigilance where the supposed 'inferior group' is more intensively monitored (e.g. 'are you certain you have not had a problem?' they might ask). In this case, unequal evaluations can provide unequal opportunities to differentially 'discover' events. This is why the concept of double-blinding (masking) is an important component of RCTs. There is an argument about which term-blinding or masking is most appropriate, and Fig. 3. 4 portrays a humorous example of this argument. But, one cannot always have a double-blind trial, and some would argue that double-blinding distances the trial from a 'real-world' approach. An example where blinding is difficult to achieve might be a surgical vs. medical intervention study where post operative patients may require additional follow-up visits, and each visit imparts an additional opportunity to elicit events. That is, it has been said that 'the patient cannot have a fever if the temperature is not taken,'[20] and for RCTs, events cannot be detected without patient contact to assess outcomes.

In order to realize a more 'real-world' principal to clinical trials, the prospective randomized open-label blinded endpoint design (PROBE design) was developed. Randomization is used so that important component of study design is retained. By using open-label therapy, the drug intervention and its comparator can be clinically titrated as would occur in a doctor's office. Of course, blinding is lost here, but only

**Fig. 3.4** A humerous example of blinding

as to the therapy. In a PROBE design, blinding is maintained as to the outcome. To test whether the use of open-label vs. double-blind therapy affected outcomes differentially, a meta analysis of PROBE trials and double-blind trials in hypertension was reported by Smith et al.[21] They found that changes in mean ambulatory blood pressure from double-blind controlled studies and PROBE trials were statistically equivalent.

## Selection of Comparison Groups

Sometimes studies assess a new active (investigational) treatment versus an approved (standard) active treatment (i.e. to assess if the old 'standard' treatment should be replaced with the new treatment), in other cases, studies are assessing if a new treatment should be added (not replacing, but rather supplementing), current treatment. In this latter case, the comparison of interest is the outcome of patients with and without the new treatment. In this instance, masking can only be accomplished by the use of a double-blind technique. Traditionally, placebo treatment has been used as the comparator to active treatment, and has been one of the standards of clinical trials.

The use of the placebo comparator has more and more been the subject of ethical concerns. In addition to ethical issues involved with the use of placebos, there are other considerations raised by the use of placebo-controls. For

example, an important lesson was learned from the Multiple Risk Factor Intervention Trial (MRFIT) regarding the use and analysis of the placebo control group, which might best be summed up as 'why it is important to watch the placebo group'.[22] MRFIT screened 361,662 patients to randomize high risk participants (using the Framingham criteria existent at that time) to special intervention (n = 6428) and usual care (n = 6438) with coronary heart disease mortality as the endpoint. The design of this well-conducted study assumed that the risk factor profile of those receiving 'special treatment interventions' would improve, while those patients in the 'usual care' group would continue their current treatments and remain largely unaffected as far as additional benefit. The special intervention approaches in MRFIT were quite successful, and all risk factor levels were reduced. However, there were also substantial and significant reductions observed the control group. That both treatment groups experienced substantial improvements in their risk factor profile translated to almost identical CHD deaths during the course of the study. Why did the control group fare so well? Several phenomena may have contributed to the improvement in the placebo-control group. First, is the Hawthorne effect, which suggests that just participating in a study is associated with increased health awareness and changes in risk factor profile, irrespective of the intervention. In addition, for the longer-term trials, there are changes in the general population that might alter events. For example, randomization in MRFIT was conducted during the 1980's, a period when health awareness was becoming more widely accepted in the USA, and likely beneficially affected the control group.

Although the ethics of placebo controls is under scrutiny, another principal regarding the placebo-control group is that sometimes being in the placebo group isn't all that bad. The Alpha-Tocopherol, Beta Carotene Cancer Prevention Study was launched in 1994.[23] By the early 1990s there was mounting clinical epidemiologic evidence of reduced cancer risk associated with higher intake of antioxidants. Treatment with vitamin E and beta carotene were considered unlikely to be harmful, and likely to be helpful; and, the question was asked whether antioxidants could reduce lung cancer-even in smokers. A double-blind, placebo-controlled RCT was launched with a 2 x 2 factorial design (see Chapter 4), and over 7000 patients in each cell. No benefit was seen with either therapy, but compared to placebo; a disturbing worsening trend was observed in the beta-carotene treated group.

Frequently, the comparison group or control group is a so called 'normal' population. Inherent to this concept is 'what is normal?'. A wit once opined that 'a normal person is one who is insufficiently tested'. Interestingly, there are a number of scientific definitions of normal (See Table 3.3). One definition of normal might be someone who fits into 97% of a Gaussian distribution, another that they lay within a preset percentile of a laboratory value or values. Other definitions exist, suffice it to say, whatever definition is used it needs to be clearly identified.

**Table 3.3** What is normal?

| Property | Term | Consequences of application |
|---|---|---|
| Distribution shape | Gaussian | Minus values |
| Lies w/in preset percentile | Percentile | Normal until workup |
| Carries no additional risk of morbidity/ mortality | Risk factor | Assumes altering risk factor alters risk |
| Socially/politically aspired | Culturally desirable | Role of society in medicine |
| Range before test suggests D- | Diagnostic | Need to know PV in your practice |
| Therapy does more good than harm | Therapeutic | New therapies continually alter this |

## Analytic Approach

### *Intention to Treat and Per-Protocol Analysis*

There are 3 general analytic approaches to clinical trials; intention-to-treat (ITT) analysis (or analysis as randomized), compliers only (or per-protocol) analysis, and analysis by treatment received. Probably the least intuitive and the one that causes most students a problem is ITT. ITT was derived from a principle called the pragmatic attitude.[24] The concept was that one was to compare the effectiveness of the intention to administer treatment A vs. the intention to administer treatment B, i.e the comparison of two treatment policies rather than a comparison of two specific treatments. With ITT, everyone assigned to an intervention or control arm is counted in their respective assigned group, whether they ultimately receive none of the treatment, or somewhat less than the trial directed. For example, if in a 1 year trial, a patient is randomized to receive an intervention, but before the intervention is administered, they drop out (for what ever reason) they are analyzed as if they received the treatment for the entire year. The same applies if the patient drops out at any time during the course of the study. Likewise, if it is determined that the patient is not fully compliant with treatment, they are still counted as if they were. In fact whether there is compliance, administrative, or protocol deviation issues, patients once randomized are counted as if they completed the trial. Most students initially feel that this is counter-intuitive. Rather the argument would be that one is really interested in what would happen if a patient is randomized to a treatment arm and they take that treatment for the full trial duration and are fully compliant-this, one would argue, gives one the real information needed about the optimal effect of an intervention (this, by the way, is a description of the compliers only analysis). So why is ITT the scientifically accepted primary analysis for clinical trials? As mentioned before, randomization is arguably one of the most important aspects of a clinical trial design. If patients once randomized to a treatment are not included in the analysis, the process of randomization is compromised. It is not a leap of

faith to wonder if patients dropping out of the intervention arm might be different than the patients dropping out of a control arm. Thus, if ITT is not used, one loses the assurance of equal distribution of unknown confounders between the treatment groups. One example of the loss of randomization if ITT is not used might be differential dropouts between the intervention and control arm for adverse events. Also, if patients with more severe disease are more likely to dropout from the placebo arm; or conversely patients who are older dropout more frequently from the placebo arm thereby removing them from the analysis, this could result in an imbalance between the two groups. Another argument for ITT is that it provides for the most conservative estimate of the intervention effect (if the analysis includes patients that did not get the entire treatment regimen and the regimen is beneficial, clearly the treatment effect will have been diluted). Thus if using ITT analysis reveals a benefit, it adds to the credibility of the effect measure. Of course, one could argue that one could miss a potentially beneficial effect if the intervention is diluted.

With the compliers only analysis, only the patients that complete the trial and comply fully with that treatment are analyzed. The problem is that if a beneficial effect is seen, one can wonder what the loss of randomization (and thereby equality of confounders between groups) means to that outcome, particularly if ITT does not demonstrate a difference. The loss of randomization and the loss of balanced confounders between the treatment and control groups is exemplified by an analysis of the Coronary Drug Project, where it was determined that poor compliers to placebo had a worse outcome than good compliers to placebo.[25] This would suggest that there are inherent differences in patients who comply vs. those who do not. The Coronary Drug Project was a trial aimed at comparing clofibrate with placebo in patients with previous myocardial infarction with the outcome of interest being mortality. Initially reported as a favorable intervention (there was a 15% 5 year mortality in the compliers only analysis clofibrate group, compared to a 19.4% mortality in the placebo group-$p<.01$), with ITT analysis there was essentially no difference in outcome (18.2 vs. 19.4%-$p<.25$). Given the differences in outcome between placebo compliers and placebo non compliers, one can only assume the same for the investigational drug group. Likewise, the Anturane Reinfarction Trial was designed to compare anturane with placebo in patients with a prior MI and in whom mortality was the outcome of interest.[26] 1629 patients were randomized to placebo and 812 to anturane (71 patients were later excluded because it was determined that they did not meet eligibility criteria). The study initially reported anturane as a favorable intervention (although the $p<.07$), but when the 71 ineligible randomized patients were included in the analysis the $p=.20$. Again further analysis demonstrated that in the anturane ineligible patients, overall mortality was 26% compared to the mortality in the anturane eligible patients which was 9%.

If one considers the common reasons for patient withdrawal from a study, ineligibility is certainly one. In addition, patients may be dropped from a trial for poor compliance, and adverse drug events; and patients may be excluded from analysis due to protocol deviations or patients lost to follow up. Some of the reasons for ineligibility are protocol misinterpretations, clerical error, or wrong diagnosis at the time of randomization. Sometimes the determination of ineligibility is above question

(eg the patient fell outside of the studies predetermined age limit) but frequently ineligibility requires judgment. The MILIS study is an example of this latter concept. MILIS compared propranolol, hyaluronidase, and placebo in patients with early acute MI, in order to observe effects on mortality. Subsequently, some patients were deemed ineligible because the early diagnosis of MI was not substantiated. But, what if the active therapy actually had an effect on preventing or ameliorating the MI? The problem with not including patients in this instance is that more patients could have been withdrawn from the placebo group compared to the active therapy group and as a result interpretation of the data would be compromised.

Of course, as is true of most things in clinical research there is not just one answer, one has to carefully assess the trial specifics. For example, Sackett and Gent cite a study comparing heparin to streptokinase in the treatment of acute myocardial infarction.[27] The ITT analysis showed that streptokinase reduced the risk of in-hospital death by 31% ($p=0.01$). However, 8 patients randomized to the heparin group died after randomization, but before they received the heparin. Analysis restricted to only those who received study drug decreased the benefit of streptokinase (and the p value).

In summary, ITT is the most accepted (by most scientists and the FDA) as the analysis of choice for clinical trials. This is because it assures statistical balance (as long as randomization was properly performed), it 'forces' disclosure of all patients randomized in a trial, and most of the arguments against ITT can be rationally addressed.

Analysis as treated is another analytic approach that addresses not the group to which the patient was randomized and not compliers only, but what the patient actually received. This analytic approach is utilized most often when patients cross over from one treatment to the other; and, this occurs most often in surgical vs. medical treatment comparisons. For example, patient's randomized to medical treatment (vs. coronary artery bypass surgery) might, at sometime during the study, be deemed to need the surgery, and are thus crossed over to the surgical arm and are then assessed as to the treatment they received ie surgery. Like compliers only analysis, this might be an interesting secondary analytic technique, but shares many of the same criticisms discussed earlier for compliers only analysis. In fact, because such trials cannot easily be double-blind, even greater criticism can be leveled against this analytic approach than compliers only analysis. In addition, statistical testing with this analytic approach, is more complicated, not only by the crossovers, but by the inherent nature of the comparison groups. In comparison trials of 1 drug and placebo, for example, it is reasonable to assume that if the drug is superior to placebo (or an active control) patients in the drug group will average fewer events in the follow-up period. When this is displayed as survival curves, the survival curves will increasingly separate. In trials comparing surgical to medical therapy, the aforementioned approach may not be reasonable. For example, if patients randomized to surgery have a high early risk (compared to the non-surgical group) and a lower risk later, these risks may cancel and be similar to the number of events under the null hypothesis of no difference between groups. The issue of comparing surgical and non-surgical therapies in clinical trials has been nicely summarized by Howard et al.[28]

## *Subgroup Analysis*

As pointed out by Assmann et al, most clinical trials collect substantial baseline information on each patient in the study.[29] The collection of baseline data has at least 4 main purposes: 1) to characterize the patients included in the trial, ie to determine how successful randomization was 2) to allow assessment of how well the different treatment groups are balanced, 3) to allow for analysis per treatment group, 4) to allow for subgroup analysis in order to assess whether treatment differences depend on certain patient characteristics. It is this 4$^{th}$ purpose that is perhaps the most controversial because it can lead to 'data dredging' or has some wits have opined, 'if you interrogate the data enough, you can have it admit to anything'. For example, Sleight and colleagues, in order to demonstrate the limitations of subgroup analysis, performed subgroup analysis in the ISIS-2 trial by analyzing treatment responses according to the astrological birth sign of the subject.[30] This analysis suggested that the treatment was quite effective and statistically significant for all patients except those born under the sign of Gemini or Libra. The validity of any subgroup observation tends to be inversely proportional to the number of subgroups analyzed. For example, for testing at the 5% significance level (p = .05) an erroneous statistically significant difference will be reported (on average) 5% of the time (i.e. false + rate of 5%). But, if 20 subgroups are analyzed, the false positive rate would approach 64% (Table 3.4, Fig. 3.5).

It is true, that meaningful information from subgroup analysis is restricted by multiplicity of testing and low statistical power and that surveys on the adequacy of the reporting of clinical trials consistently find the reporting of subgroup analyses to be wanting. Most studies enroll just enough participants to ensure that the primary efficacy hypothesis can be adequately tested, and this limits the statistical ability to find a difference in subgroup analyses; and, the numbers of subjects available for subgroup analysis is further compounded by loss of compliance, the need

**Table 3.4** Probability of at least one significant result at the 5% significance level given no true differences

| Number of tests | Probability |
|---|---|
| 1 | 0.05 |
| 2 | 0.10 |
| 3 | 0.14 |
| 5 | 0.23 |
| 10 | 0.40 |
| 20 | 0.64 |

Cook D I et al. **Subgroup analysis in clinical trials**. MJA 2004; 180: 289–291. © 2004. *The Medical Journal of Australia*. Reproduced with permission.

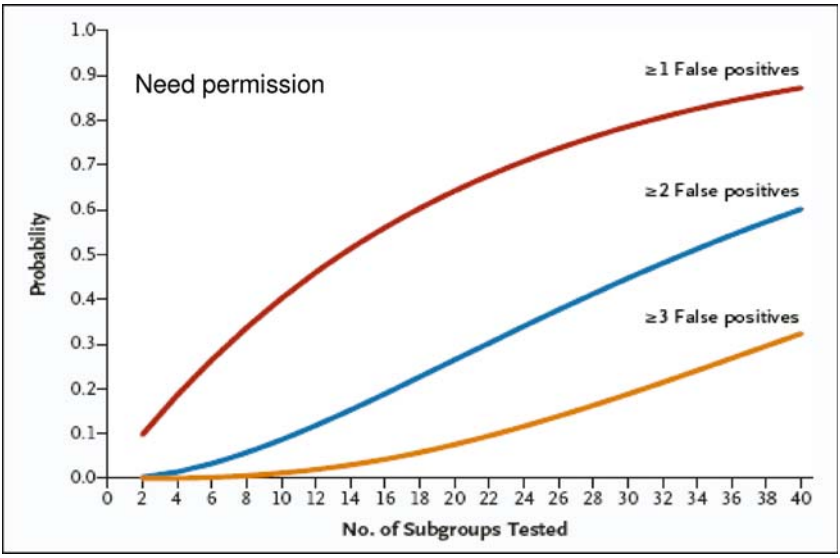



**Fig. 3.5** Probability that multiple subgroup analyses will yield at least one (Red), two (Blue), or three (Yellow) false positive results

for adjustments for multiple testing, etc. Some have taken this to mean that subgroup analyses are useless. When results from a subgroups analysis are at variance from the overall group outcome, the results are still likely to be true if the subgroup is large, they are pre-specified rather than *post hoc* (i.e. 'after the fact') and they are of limited number (not all post hoc analyses are subgroup analyses, but arguably most are). At the least, whether pre-specified or *post hoc* subgroup analyses serve as hypothesis generating questions for subsequent trials. An example where a post-hoc analysis was not ignored is the Stroke Prevention by Aggressive Reduction in Cholesterol Levels (SPARCL) study where LIPITOR 80 mg vs placebo was administered in 4,731 subjects without CHD who had a stroke or TIA within the preceding 6 months.[31] A higher incidence of hemorrhagic stroke was seen in the LIPITOR 80 mg group compared to placebo. Subjects with hemorrhagic stroke on study entry appeared to be at increased risk for hemorrhagic stroke. As a result, Pfizer revised the US Prescribing Information for atorvastatin to include a precaution for its use of 80 mg in patients with prior history of stroke.

What can be said is that if subgroup analysis is used and interpreted carefully, it can be useful. Even among experts, opinions range from only accepting pre-specified subgroup analyses supported by a very strong *a priori* biological rationale, to a more liberal view in which subgroup analyses, if properly carried out and interpreted, are permitted to play a role in assisting doctors and their patients to choose between treatment options. In reviewing a report that includes subgroup analyses, Cook et al suggest addressing the following issues (Table 3.5): 1) were the

**Table 3.5** Checklist for subgroup analyses

Design
- Are the subgroups based on pre-randomisation characteristics?
- What is the impact of patient misallocation on the subgroup analysis?
- Is the intention-to-treat population being used in the subgroup analysis?
- Were the subgroups planned a *priori*?
- Were they planned in response to existing trial or biological data?
- Was the expected direction of the subgroup effect stated a *priori*?
- Was the trial designed to have adequate power for the proposed subgroup analysis?

Reporting
- Is the total number of subgroup analyses undertaken declared?
- Are relevant summary data, including event numbers and denominators, tabulated?
- Are analyses decided on a *priori* clearly distinguished from those decided on a *posteriori*?

Statistical analysis
- Are the statistical tests appropriate for the underlying hypotheses?
- Are tests for heterogeneity (i.e., interaction) statistically significant?
- Are there appropriate adjustments for multiple testing?

Interpretation
- Is appropriate emphasis is being placed on the primary outcome of the study?
- Is the validity of the findings of the subgroup analysis discussed in the light of current biological knowledge and the findings from similar trials?

Cook D I et al. **Subgroup analysis in clinical trials**. MJA 2004; 180: 289–291. © 2004. *The Medical Journal of Australia*. Reproduced with permission.

subgroups appropriately defined, (that is be careful about subgroups that are based upon characteristics measured after randomization e.g. adverse drug events may be more common as reasons for withdrawal from the active treatment arm whereas lack of efficacy may be more common in the placebo arm); 2) were the subgroup analyses planned before the implementation of the study (in contrast to after the study completion or during the conduct of the study); 3) does the study report include enough information to assess the validity of the analysis eg the number of subgroup analyses; 4) does the statistical analyses use multiplicity and interaction testing; 5) were the results of subgroup analyses interpreted with caution; 6) is there replication of the subgroup analysis in another independent study; 7) was a dose-response relationship demonstrated; 8) was there reproducibility of the observation within individual sites; and 9) is there a biological explanation.

## *Traditional versus Equivalence testing (Table 3.6)*

Most clinical trials have been designed to assess if there is a difference in the efficacy to two (or more) alternative treatment approaches (with placebo usually being the comparator treatment). There are reasons why placebo-controls are preferable to active controls, not the least of which is the ability to distinguish an effective treatment from a less effective treatment. However, if a new treatment is

**Table 3.6** The types of RCTs and there relationship to hypothesis testing[7]

| RCT type | Null hypothesis | Alternative hypothesis |
|---|---|---|
| Traditional | New = Old | New ≠ Old (i.e., New < Old or New > Old) |
| Equivalence | New < Old + δ (where δ is a "cushion," that is that the new is at least δ worse than the old) | New ≥ Old + δ |
| Non-inferiority | New < Old | New = Old |

considered to be equally effective but perhaps less expensive and/or invasive, or a placebo-control is considered unethical, then the new treatment needs to be compared to an established therapy and the new treatment would be considered preferable to the established therapy, even if it is just as good (not necessarily better) as the old. The ethical issues surrounding the use of a placebo-control and the need to show a new treatment to only be as 'good as' (rather than better) has given rise to the recent interest in equivalence testing. With traditional (superiority) hypothesis testing, the null hypothesis states that 'there is no difference between treatment groups (i.e. New = Old or placebo or standard therapy). Rejecting the null, then allows one to definitively state if one treatment is better (or worse) than another (i.e. New > or < Old). The disadvantage is if at the conclusion of an RCT there is not evidence of a difference, one cannot state that the treatments are the same, or as good as one to the other, only that the data are insufficient to show a difference. That is, when the null hypothesis is not accepted, it is simply the case where it cannot be rejected. The appropriate statement when the null hypothesis is not rejected is 'there is not sufficient evidence in these data to establish if a difference exists.'

Equivalence testing in essence 'flips' the traditional null and alternative hypotheses. Using this approach, the null hypothesis is that the new treatment is worse than the old treatment (i.e. New < Old); that is, rather than assuming that there is no difference, the null hypothesis is that a difference exists and the new treatment is inferior. Just as in traditional testing, the two actions available resulting from the statistical test are 1) reject the null hypothesis, or 2) failure to reject the null hypothesis. However, with equivalence testing rejecting the null hypothesis is making the statement that the new treatment is not worse than old treatment, implying the alternative, that is 'that the new treatment is **as good** as or better than the old' (i.e. New ≥ Old). Hence, this approach allows a definitive conclusion that the new treatment is as good as the old.

One caveat is the definition of 'as good as,' which is defined as being in the 'neighborhood' or having a difference that is so small that it is to be considered clinically unimportant (generally, event rates within +/− 2% – this is known as the equivalence or noninferiority margin usually indicted by the symbol δ). The need for this 'neighborhood' that is considered 'as good as' exposes the first shortcoming of equivalence testing – having to make a statement that 'I reject the null

hypothesis that the new treatment is worse than the old, and accept the alternative hypothesis that it is as good or better – *and by that I mean that it is within at least 2% of the old*' (the wording in italics are rarely included in the conclusions of a manuscript). A second disadvantage of equivalence testing is that no definitive statement can be made that there is evidence that the new treatment is worse. Just as in traditional testing, one never accepts the null hypothesis – one only fails to reject it. Hence if the null is not rejected, all one can really say is that there is ***insufficient evidence in these data*** that the new treatment is as good as or better than the old treatment. Another problem with equivalence testing is that one has to rely on the effectiveness of the active control obtained in previous trials, and on the assumption that the active control would be equally effective under the conditions of the present trial.

An example of an equivalence trial is the Controlled ONset Verapamil INvestigation of Cardiovascular Endpoints study (CONVINCE), a trial that also raised some ethical issues that are different from those usually involved in RCT's.[32] CONVINCE was a large double-blind clinical trial intended to assess the equivalence of verapamil and standard therapy in preventing cardiovascular disease-related events in hypertensive patients. The results of the study indicated that the verapamil preparation was not equivalent to standard therapy because the upper bound of the 95% confidence limit (1.18) slightly exceeded the pre-specified boundary of 1.16 for equivalence. However, the study was stopped prematurely for commercial reasons. This not only hobbled the findings in terms of inadequate power, it also meant that participants who had been in the trial for years were subjected to a 'breach in contract'. That is, they had subjected themselves to the risk of an RCT with no ultimate benefit. There was a good deal of criticism borne by the pharmaceutical company involved in the decision to discontinue the study early. Parenthetically, the company involved no longer exists.

Another variant of equivalence testing is non-inferiority testing. Here the question is again slightly different in that one is asking whether the new intervention is simply not inferior to the comparator (i.e. New $\not<$ Old). One advantage is that statistical significance could be only 'one-tailed' since there is no implication that the analysis is addressing whether the new treatment is better or as good as, only that it is not inferior. Weir et al utilized this approach in evaluating a comparison of valsartin/hydrochlorthiazide (VAL/HCTZ) with amlodipine in the reduction of mean 24-hour diastolic BP (DBP).[33] Noninferiority of the VAL/HCTZ combination to amlodipine was demonstrated, and fewer adverse events were noted with the combination treatment as well. The null hypothesis for this analysis was that the reduction in mean 24-hour DBP from baseline to the end of the study with VAL/HCTZ was $\geq 3$ mmHg less (the non-inferiority margin) than that with amlodipine. Again, a caveat has been recently raised by LeHenanff et al. and Kaul et al.[34, 35] LeHananff et al[35] reviewed studies published between 2003 and 2004 that were listed as equivalence or noninferiority, and noted a number of deficiencies, key among them being the absence of the equivalence or non inferiority margin.[35]

Equivalence/non-inferiority trials are further discussed in Chapter 4.

## Losses to Follow Up

Patients who are lost to follow-up are critical in clinical trials and are particularly problematic in long-term trials. Patients lost to follow-up might be regarded as having had poor results (that is assumed that they experienced treatment failure); so if there are sufficient numbers of them, trial results can be skewed to less of an effect, even if, in fact, they did not have poor results. If, in the different study arms, there are equal numbers lost to follow-up, and they are lost for the same reasons, lost to follow up would not matter, but this is unlikely to occur. Of course, in ITT analysis, patients lost-to-follow-up are still counted, but the argument is how to count them. Some would argue that it is appropriate to count them as poor outcomes since this will give the most conservative result, while others argue that since their outcome is not known, they should not be counted. In fact, there is little data reported on the actual impact on a study result of patients lost to follow up. In one study, Joshi et al did address this issue in a long-term follow-up (up to 16 years of follow-up) of patients who had undergone knee arthroplasty. With the concerted effort of full-time personnel and a private detective, all 123 patients initially lost to follow-up were traced. Patients cited a variety of reasons why they did not attend follow-up visits, including: change of residence, inability to travel, displeasure with the physician or staff, financial constraints, satisfaction with the results so that they did not feel follow-up was necessary, and poor results. They also found that more women than men were lost to follow-up.

## Surrogate Endpoints

In 1863, Farr said ‘death is a fact, the rest is inference’. In choosing outcomes of interest, death or a disease event is usually the event of interest. However, it is frequently necessary to use a surrogate for the endpoint of interest, such as when the disease occurrence is rare and/or far in the future. The main variable that drives sample size and Power is the difference in the outcome between the intervention and the control group. Table 3.7 summarizes the sample size necessary based upon these aforementioned differences. One can see from Table 3.7 that most studies would have to be quite large unless the treatment difference is large, and for most outcomes these days, it is not common to have treatment differences of more than 20%.

A surrogate endpoint is simply a laboratory value, sign, or symptom that is a substitute for the real outcome one is interested in.[36] The assumption is that changes induced in a surrogate endpoint accurately and nearly completely reflect changes in the clinically meaningful endpoint. To realize that assumption, an accurate well-documented model of the outcome of interest is a prerequisite, but it should be understood that the model is only that, and the model may be far from the truth. As is true of most definitions, there is debate about the best definition for a surrogate endpoint, and it is also important to distinguish surrogate endpoints

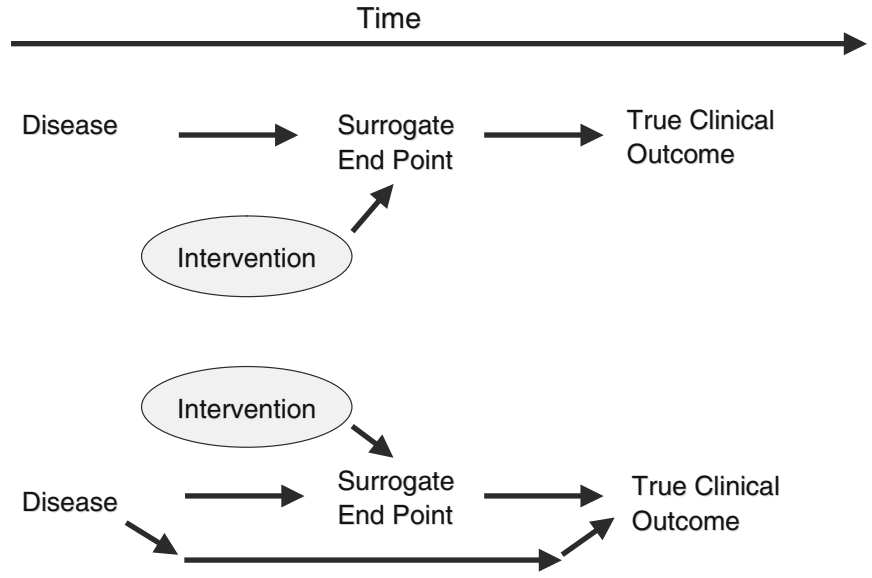
**Table 3.7** Selection of endpoints

- RCTs are expensive and time consuming
- Most focus on differences in "events"
- Generally implies large studies

| | | Treatment effect | | | |
|---|---|---|---|---|---|
| | | 10% | 20% | 30% | 50% |
| Rate in the "control" treatment | 2% | 1.8% | 1.6% | 1.4% | 1% |
| | | 97,959 | 23,223 | 9,757 | 3,104 |
| | 3% | 2.7% | 2.4% | 2.1% | 1.5% |
| | | 64,673 | 15,339 | 6,447 | 2,053 |
| | 5% | 4.5% | 4.0% | 3.5% | 2.5% |
| | | 38,044 | 9,032 | 3,800 | 1,212 |
| | 10% | 9% | 8% | 7% | 5% |
| | | 18,072 | 4,302 | 1,815 | 581 |
| | 20% | 18% | 16% | 14% | 10% |
| | | 8,086 | 1,937 | 882 | 266 |
| | 50% | 45% | 40% | 35% | 266 |
| | | 2,095 | 518 | 226 | 77 |

from intermediate endpoints and statistical correlations. Speaking statistically, Prentice[37] has offered the following definition: '*a response variable for which a test of the null hypothesis of no relationship to the treatment groups under comparison is also a valid test of the corresponding null hypothesis based on the true endpoint.*'

Examples of surrogate endpoints include blood pressure reduction in lieu of stroke (this has been termed a 'strong surrogate' by Anand et al);[38] fasting blood sugar (or hemoglobin H1C) in lieu of diabetic complications; and bone mineral density in lieu of fractures. Surrogates are also commonly used early in drug development such as dose ranging or preliminary proof of efficacy ('developmental surrogates'). 'Supportive surrogates' are those outcomes that support and strengthen clinical trial data. The reasons for choosing a surrogate endpoint predominantly revolve around the fact that it might be easier to measure than the clinical endpoint of interest, or that it occurs early in the natural history of the disease of interest (and thus long-term trials are avoided). But as is true of almost any decision one makes in conducting a clinical trial, there are assumptions and compromises one has to make when choosing a surrogate endpoint. For example, many surrogates have been inadequately validated, and many if not most surrogates have several effect pathways (See Fig. 3.6). Other considerations for using a surrogate endpoint are that it should be easier to assess than the corresponding clinical endpoint, and in general, be more frequent; and, that an estimate of the expected clinical benefit should be derivable from the interventions effect upon the surrogate. An example of the controversy regarding surrogate endpoints is highlighted by the discussion of Kelsen[39] regarding the use of tumor regression as an adequate surrogate for new drugs to treat colorectal cancer. On the basis of a meta-analysis, Buyse et al[40] proposed that surrogate endpoints of efficacy, without direct demonstration of an




*Adapted from: Fleming, T. R. et. al. Ann Intern Med 1996;125:605-613*

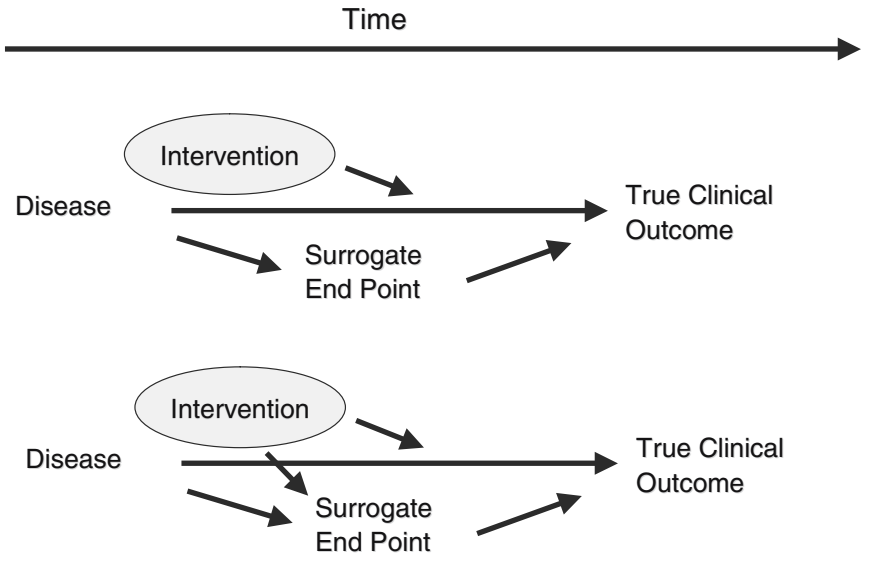



**Fig. 3.6** The different potential pathways for surrogate endpoints

improvement in survival, could be used to identify effective new agents. The FDA, however, requires that there be a survival advantage before it approves such a drug. That is, a response rate higher than standard therapy (defined as tumor regression >50%) is by itself an inadequate benefit for drug approval. As stated in the commentary by Kelsen 'the critical question in the debate over the adequacy of response rate as a surrogate endpoint for survival is whether an objective response to treatment is merely associated with a better survival, or whether the tumor regression itself lengthens survival.'

It should be understood that there are differences in an intermediate endpoint, correlate, and a surrogate endpoint, although an intermediate endpoint may serve as a surrogate. Examples of intermediate endpoints include such things as angina pectoris, or hyperglycemic symptoms i.e. these are not the ultimate outcome of interest (MI, or death etc) but are of value to the patient should they be benefited by an intervention. Another example is from the earlier CHF literature where exercise walking time was used as an intermediate endpoint as well as a surrogate marker (in lieu of survival). A number of drugs improved exercise walking time in the CHF patient; but long-term studies proved that the same agents that improved walking time resulted in earlier death. An example of surrogate 'misadventure' is characterized by a hypothetical scenario where a new drug is used in pneumonia, and it is found to lower the patients white blood count (wbc-this used as a surrogate marker for improvement in the patients pneumonia). Subsequently, this 'new drug' is found to be cytotoxic to wbc's but obviously had little effect on the pneumonia. But, perhaps the most glaring example of a surrogate 'misadventure' is represented by a real trial –the Cardiac Arrhythmia Suppression Trial (CAST).[41] At the time of CAST, premature ventricular contractions (PVC's) were thought to be a good surrogate for ventricular tachycardia or ventricular fibrillation, and thereby for sudden cardiac death (SCD). It was determined that many anti-arrhythmic agents available at the time or being developed reduced –PVC's, and it was assumed would benefit the real outcome of interest- SCD. CAST was proposed to test the hypothesis that these anti-arrhythmic agents did actually reduce SCD (in a post MI population) and this study was surrounded with some furor about the studies ethics, since a placebo control was part of the study design (it was felt strongly by many that the study was unethical since it was so likely that reduction in PVCs led to a reduction in SCD and how could one justify a placebo arm). In fact, it turned out that the anti-arrhythmic therapy not only failed to reduce SCD, but in some cases it increased its frequency. A final example occurred in 2007, when the Chairman of the FDA Advisory panel that reviewed the safety of rosiglitazone stated that the time has come to abandon surrogate endpoints for the approval of type 2 diabetes drugs. This resulted from the use of glycated hemoglobin as a surrogate for diabetes morbidity and mortality as exemplified in the ADOPT (A Diabetes Outcome Prevention Trial) study where patients taking rosiglitazone had a greater decrease in glycosolated hemoglobin than in patients taking comparator drugs, yet the risks of CHF and cardiovascular ischemia were higher with rosiglitazone.[42]

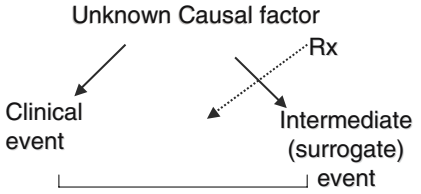



**Fig. 3.7** Depicts a correlation (statistically significant) between a causal factor and a cllnical event. However, while treatment impacted the intermediate (surrogate) event, it had no effect on the clinical event since it does not lie in the direct pathway

Correlates may or may not be good surrogates. Recall, 'that a surrogate endpoint requires that the effect of the intervention on the surrogate end-point predicts the effect on the clinical outcome-a much stronger condition than correlation.'[36] Another major point of confusion is that between statistical correlation and proof of causality as demonstrated in Fig. 3.7 as discussed by Boissel et al.[43]

In summary, it should be understood that most (many) potential surrogates markers used in clinical research have been inadequately validated and that the surrogate marker must fully (or nearly so) capture the effect of the intervention on the clinical outcome of interest. However, many if not most treatments have several effect pathways and this may not be realized, particularly early in the research of a given intervention. Table3. summarizes some of the issues that favor support in using a surrogate. Surrogate endpoints are most useful in phase 1 and 2 trials where 'proof of concept' or dose-response is being evaluated. One very important additional down-side to the use of surrogate measures is a result of its strength i.e. the ability to use smaller sample sizes and shorter trials, in order to gain insight into the benefit of an intervention. This is because smaller and shorter term studies result in the loss of important safety information.

## Selection of Endpoints

Table 3.8 makes the point that for most clinical trials, one of the key considerations is the difference in events between the investigational therapy and the control. It is this difference (along with the frequency of events) that drives the sample size and power of the study. From Table 3, one can compare the rate in the control group compared to the intervention effect. Thus, if the rate in the control group of the event of interest is high (say 20%) and the treatment effect is 20% (i.e. an expected 50% reduction compared to control), a sample size of 266 patients would be necessary. Compare that to a control rate of 2% and a treatment effect of 10% (i.e. a reduction compared to control from 2% to 1.8%), where a sample size of 97959 would be necessary.

**Table 3.8** Support for surrogates

| Factor | Favors surrogate | Does not favor surrogate |
|---|---|---|
| Biological plausibility | Epi. evidence extensive, consistent, quantitative; credible animal model; pathogenesis & drug mechanism understood; surrogate late in causal path | Inconsistent epi; no animal model; pathogenesis unclear; mechanisms not studied, surrogate earlier in causal path |
| Success in clinical trials | Effect on surrogate has predicted outcome with other drugs in class; and in several classes | Inconsistent results across classes |
| Risk/B, PubH considerations | Serious or life-threatening illness and no alternative Rx; large safety database; short term use; difficult to study clinical end point | Less serious disease; little safety data; long term use; easy to study clinical endpoint |

## *Composite endpoints*

Composite endpoints (rather than a single endpoint) are being increasingly used as effect sizes for most new interventions are becoming smaller. Effect sizes are becoming smaller because newer therapies need to be assessed when added to all clinically accepted therapies; and, thus the chance for an incremental change is reduced. For example, when the first therapies for heart failure were introduced, they were basically added to diuretics and digitalis. Now, a new therapy for heart failure would have to show benefit in patients already receiving more powerful diuretics, digitalis, angiotensin converting enzyme inhibitors and/or angiotensin receptor blockers, appropriately used beta adrenergic blocking agents, statins etc. To increase the 'yield' of events, composite endpoints are utilized (a group of individual endpoints that together form a 'single' endpoint for that trial). Thus, the rationale for composite endpoints comes from 3 basic considerations: statistical issues (sample size considerations due to the need for high event rates in the trial in order to keep the trial relatively small, of shorter duration and with less expense), the pathophysiology of the disease process being studied, and the increasing need to evaluate an overall clinical benefit. The risk associated with the use of composite endpoints is that the benefits ascribed to an intervention are assumed to relate to all the components of the composite. Consider the example of a composite endpoint that includes death, MI, and urgent revascularization. In choosing the components of the composite, one should not be driven by the least important variable just because it happens to be the most frequent (e.g. death, MI, urgent revascularization, would be a problem if revascularization turned out to be the main positive finding). Montori et al provided guidelines for interpreting composite endpoints which included asking whether the individual components of composite endpoints were of similar importance, occurred with about the same frequency, had similar relative risk reductions, and had similar biologic mechanisms.[44]

Freemantle et al. assessed the incidence and quality of reporting of composite endpoints in randomized trials and asked whether composite endpoints provide for greater precision but with greater uncertainty.[45] Their conclusion was that the reporting of composite outcomes is generally inadequate and as a result, they provided several recommendations regarding the use of composite endpoints such as following the CONSORT guidelines, interpreting the composite endpoint rather than parsing the individual endpoints, and defining the individual components of the composite as secondary outcomes. The reasons for their recommendations stemmed from their observations that in many reports they felt that there was inappropriate attribution of the treatment effects on specific endpoints when only composite endpoints yielded significant results, the effect of dilution when individual endpoints might not all react in the same direction, and the effect of excessively influential endpoints that are not associated with irreversible harm. In an accompanying editorial by Lauer and Topel they list a number of key questions that should be considered when composite endpoints are reported or when an investigator is contemplating their use.[46] First, is whether the end points themselves are of clinical interest to patients and physicians, or are they surrogates; second, how non fatal endpoints are measured (e.g. is judgment involved in the end point ascertainment, or is it a hard end point); third, how many individual endpoints make up the composite and how are they reported (ideally each component of the composite should be of equal clinical importance - in fact, this is rarely the case); and finally, how are non fatal events analyzed - that is are they subject to competing risks. As they point out, patients who die cannot later experience a non fatal event so a treatment that increases the risk of death may appear to reduce the risk of non fatal events.[46]

Kip et al[47] reviewed the problems with the use of composite endpoints in cardiovascular studies. The term "major adverse cardiac events:" or MACE is used frequently in cardiovascular studies, a term that was born with the percutaneous coronary intervention studies in the 1990's. Kip et al noted that MACE encompassed a variety of composite endpoints, the varying definitions of which could lead to different results and conclusions, leading them to the recommendation that MACE should be avoided. Fig. 3.8 from their article demonstrates this latter point rather well.

## Trial Duration

An always critical decision in performing or reading about a RCT (or any study for that matter) is the specified duration of follow-up, and how that might influence a meaningful outcome. Many examples and potential problems exist in the literature, but basically in interpreting the results of any study (positive or negative) the question should be asked 'what would have happened had a longer follow-up period been chosen?' A recent example is the Canadian Implantable Defibrillator Study (CIDS).[48] CIDS was a RCT comparing the effects of defibrillator implantation to amiodarone in preventing recurrent sudden cardiac death in 659 patients. At the end of study (a mean of 5 months) a 20% relative risk reduction occurred in all-cause mortality, and a 33%

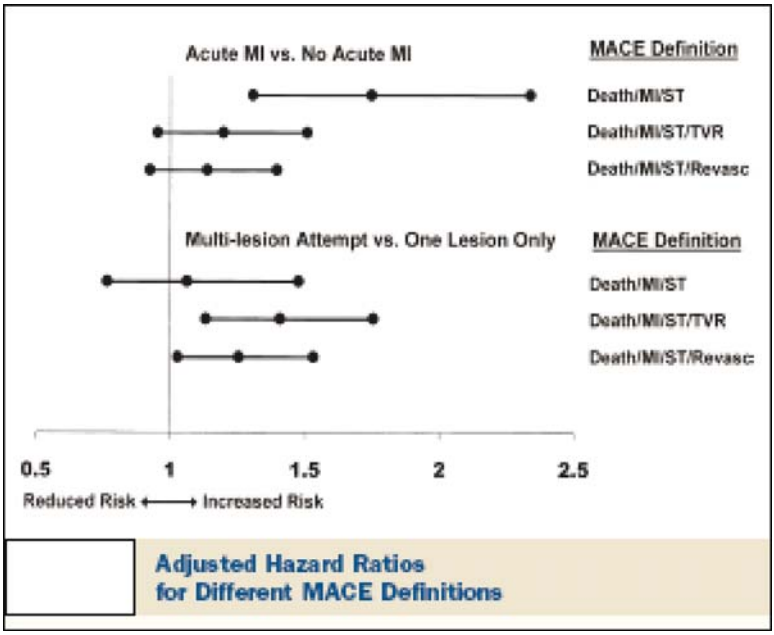



**Fig. 3.8** Adjusted hazard ratios or different definitions of major adverse cardiac events (MACE) comparing acute myocardial infarction (MI) versus nonacute MI patients (top) and patients with multilesion versus single-lesion percutaneous coronary intervention (bottom). Filled center circles depict the adjusted hazard ratios, filled circles at the left and right ends depict the lower and upper 95% confidence limits. Reverse = revascularization; ST = stent thrombosis; TVR = target vessel revascularization

reduction occurred in arrhythmic mortality, when ICD therapy was compared with amiodarone (this latter reduction did not reach statistical significance). At one center, it was decided to continue the follow-up for an additional mean of 5.6 years in 120 patients who remained on their originally assigned intervention.[49] All-cause mortality was then found to be increased in the amiodarone group. The Myocardial Ischemia Reduction with Aggressive Cholesterol Lowering (MIRACL) trial is an example of a potential problem in which study duration could have been problematic (but probably wasn't).[50] The central hypothesis of MIRACL was that early rapid and profound cholesterol lowering therapy with atorvastatin could reduce early recurrent ischemic events in patients with unstable angina or acute non-Q wave infarction. Often with acute intervention studies, the primary outcome is assessed at 30 days after the sentinel event. From Fig. 3.9 one can see that there was no difference in the primary outcome at 30 days. Fortunately the study specified a 16 week follow-up, and a significant difference was seen at that time point. Had the study been stopped at 30 days the ultimate benefit would not have been realized. Finally, an example from the often cited controversial ALLHAT study which demonstrated a greater incidence in new diabetes in the diuretic arm as assessed at the study end of 5 years.[51] The investigators pointed out that this increase in diabetes did not result in a statistically significant difference in adverse

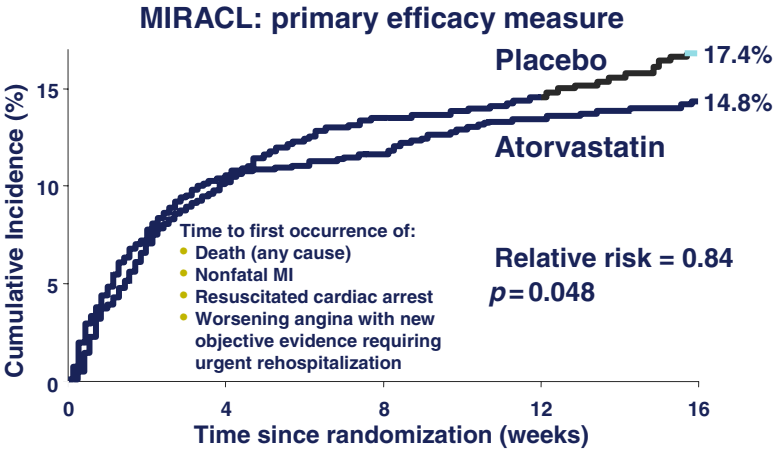



**Fig. 3.9** The results of MIRACL for the primary outcome. What would have been the conclusion for the intervention if the pre-specified study endpoint was 1 month?

outcomes when the diuretic arm was compared to the other treatment arms. Many experts have subsequently opined that the trial duration was too short to assess adverse outcomes from diabetes, and had the study gone on longer that it is likely that a significant difference in adverse complications from diabetes would have occurred.

## The Devil Lies in the Interpretation

It is interesting to consider and important to reemphasize, that intelligent people can look at the same data and render differing interpretations. MRFIT is exemplary of this principal, in that it demonstrates how mis-interpretation can have far-reaching effects. One of the conclusions from MRFIT was that reduction in cigarette smoking and cholesterol was effective, but '*possibly an unfavorable response to antihypertensive drug therapy in certain but not all hypertensive subjects'* led to mixed benefits.[22] This 'possibly unfavorable response' has since been at least questioned if not proven to be false.

Differences in interpretation was also seen in the alpha-tocopherol, beta carotene cancer study.[23] To explain the lack of benefit and potential worsening of cancer risk in the treated patients, the authors opined that perhaps the wrong dose was used, or that the intervention period was to short, since '*no known or described mechanisms and no evidence of serious toxic effects of this substance* (beta carotene) *in humans*' had been observed. This points out how ones personal bias can influence ones 'shaping' of the interpretation of a trials results. Finally, there are many examples of trials where an interpretation of the results is initially presented only to find that after publication differing interpretations are rendered. Just consider the recent controversy over the interpretation of the ALLHAT results.[51]

**Causal Inference**, and the role of **the Media** in reporting clinical research will be discussed in chapters 16 and 20.

## Conclusions

While randomized clinical trials remain a 'gold standard', there remains many aspects of trial design that must be considered before accepting the studies results, even when the study design is a RCT. Starzi et al in their article entitled 'Randomized Trialomania? The Multicentre Liver Transplant Trials of Tacrolimus' outline many of the roadblocks and pitfalls that can befall even the most conscientious clinical investigator.[52] Ioannidis presents an even more somber view of clinical trials, and has stated 'there is increasing concern that in modern research, false findings may be the majority or even the vast majority of published research claims. However, this should not be surprising. It can be proven that most claimed research findings are false.'[53] One final note of caution revolves around the use of reading or reporting only abstracts in decision making. As Toma et al noted, 'not all research presented at scientific meetings is subsequently published, and even when it is, there may be inconsistencies between these results and what is ultimately printed.[54] They compared RCT abstracts presented at the American College of Cardiology sessions between 1999 and 2002, and subsequent full length publications. Depending upon the type of presentation (e.g. late breaking trials vs. other trials) 69-79% were ultimately published; and, discrepancies between meeting abstracts and publication results were common even for the late breaking trials.[54]

## References

1. Grady D, Herrington D, Bittner V, et al. Cardiovascular disease outcomes during 6.8 years of hormone therapy: Heart and Estrogen/progestin Replacement Study follow-up (HERS II). Jama. Jul 3 2002;288(1):49-57.
2. Hulley S, Grady D, Bush T, et al. Randomized trial of estrogen plus progestin for secondary prevention of coronary heart disease in postmenopausal women. Heart and Estrogen/progestin Replacement Study (HERS) Research Group. Jama. Aug 19 1998;280(7):605-613.
3. Rossouw JE, Anderson GL, Prentice RL, et al. Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results From the Women's Health Initiative randomized controlled trial. Jama. Jul 17 2002;288(3):321-333.
4. Grady D, Rubin SM, Petitti DB, et al. Hormone therapy to prevent disease and prolong life in postmenopausal women. Ann Intern Med. Dec 15 1992;117(12):1016-1037.
5. Stampfer MJ, Colditz GA. Estrogen replacement therapy and coronary heart disease: a quantitative assessment of the epidemiologic evidence. Prev Med. Jan 1991;20(1):47-63.
6. Sullivan JM, Vander Zwaag R, Hughes JP, et al. Estrogen replacement and coronary artery disease. Effect on survival in postmenopausal women. Arch Intern Med. Dec 1990;150(12):2557-2562.
7. Glasser SP, Howard G. Clinical trial design issues: at least 10 things you should look for in clinical trials. J Clin Pharmacol. Oct 2006;46(10):1106-1115.

8. Grimes DA, Schulz KF. An overview of clinical research: the lay of the land. Lancet. Jan 5 2002;359(9300):57-61.
9. Loscalzo J. Clinical trials in cardiovascular medicine in an era of marginal benefit, bias, and hyperbole. Circulation. Nov 15 2005;112(20):3026-3029.
10. Bienenfeld L, Frishman W, Glasser SP. The placebo effect in cardiovascular disease. Am Heart J. Dec 1996;132(6):1207-1221.
11. Clark PI, Leaverton PE. Scientific and ethical issues in the use of placebo controls in clinical trials. Annu Rev Public Health. 1994;15:19-38.
12. Rothman KJ, Michels KB. The continuing unethical use of placebo controls. N Engl J Med. Aug 11 1994;331(6):394-398.
13. Montori VM, Devereaux PJ, Adhikari NK, et al. Randomized trials stopped early for benefit: a systematic review. Jama. Nov 2 2005;294(17):2203-2209.
14. Medical Research Council. Streptomycin treatment of pulmonary tuberculosis. BMJ. 1948;ii:769-782.
15. Reviews of statistical and economic books, Student's Collected Papers. J Royal Staitistical Society. 1943;106:278-279.
16. A Village of 100 A Step Ahead.
17. Beneficial effect of carotid endarterectomy in symptomatic patients with high-grade carotid stenosis. North American Symptomatic Carotid Endarterectomy Trial Collaborators. N Engl J Med. Aug 15 1991;325(7):445-453.
18. Endarterectomy for asymptomatic carotid artery stenosis. Executive Committee for the Asymptomatic Carotid Atherosclerosis Study. Jama. May 10 1995;273(18):1421-1428.
19. Lang JM. The use of a run-in to enhance compliance. Stat Med. Jan-Feb 1990;9(1-2):87-93; discussion 93-85.
20. Shem S. The House of God: Palgrace Macmillan; 1978:280.
21. Smith DH, Neutel JM, Lacourciere Y, Kempthorne-Rawson J. Prospective, randomized, open-label, blinded-endpoint (PROBE) designed trials yield the same results as double-blind, placebo-controlled trials with respect to ABPM measurements. J Hypertens. Jul 2003;21(7):1291-1298.
22. Multiple risk factor intervention trial. Risk factor changes and mortality results. Multiple Risk Factor Intervention Trial Research Group. Jama. Sep 24 1982;248(12):1465-1477.
23. The effect of vitamin E and beta carotene on the incidence of lung cancer and other cancers in male smokers. The Alpha-Tocopherol, Beta Carotene Cancer Prevention Study Group. N Engl J Med. Apr 14 1994;330(15):1029-1035.
24. Hollis S, Campbell F. What is meant by intention to treat analysis? Survey of published randomised controlled trials. Bmj. Sep 11 1999;319(7211):670-674.
25. Influence of adherence to treatment and response of cholesterol on mortality in the coronary drug project. N Engl J Med. Oct 30 1980;303(18):1038-1041.
26. Sulfinpyrazone in the prevention of sudden death after myocardial infarction. The Anturane Reinfarction Trial Research Group. N Engl J Med. Jan 31 1980;302(5):250-256.
27. Sackett DL, Gent M. Controversy in counting and attributing events in clinical trials. N Engl J Med. Dec 27 1979;301(26):1410-1412.
28. Howard G, Chambless LE, Kronmal RA. Assessing differences in clinical trials comparing surgical vs nonsurgical therapy: using common (statistical) sense. Jama. Nov 5 1997;278(17):1432-1436.
29. Assmann SF, Pocock SJ, Enos LE, Kasten LE. Subgroup analysis and other (mis)uses of baseline data in clinical trials. Lancet. Mar 25 2000;355(9209):1064-1069.
30. Sleight P. Debate: Subgroup analyses in clinical trials: fun to look at - but don't believe them! Curr Control Trials Cardiovasc med. 2000;1(1):25-27.
31. Amarenco P, Goldstein LB, Szarek M, et al. Effects of intense low-density lipoprotein cholesterol reduction in patients with stroke or transient ischemic attack: the Stroke Prevention by Aggressive Reduction in Cholesterol Levels (SPARCL) trial. Stroke. Dec 2007;38(12): 3198-3204.
32. Black HR, Elliott WJ, Grandits G, et al. Principal results of the Controlled Onset Verapamil Investigation of Cardiovascular End Points (CONVINCE) trial. Jama. Apr 23-30 2003;289(16):2073-2082.

33. Weir MR, Ferdinand KC, Flack JM, Jamerson KA, Daley W, Zelenkofske S. A noninferiority comparison of valsartan/hydrochlorothiazide combination versus amlodipine in black hypertensives. Hypertension. Sep 2005;46(3):508-513.
34. Kaul S, Diamond GA, Weintraub WS. Trials and tribulations of non-inferiority: the ximelagatran experience. J Am Coll Cardiol. Dec 6 2005;46(11):1986-1995.
35. Le Henanff A, Giraudeau B, Baron G, Ravaud P. Quality of reporting of noninferiority and equivalence randomized trials. Jama. Mar 8 2006;295(10):1147-1151.
36. Fleming TR, DeMets DL. Surrogate end points in clinical trials: are we being misled? Ann Intern Med. Oct 1 1996;125(7):605-613.
37. Prentice RL. Surrogate endpoints in clinical trials: definition and operational criteria. Stat Med. Apr 1989;8(4):431-440.
38. Anand IS, Florea VG, Fisher L. Surrogate end points in heart failure. J Am Coll Cardiol. May 1 2002;39(9):1414-1421.
39. Kelsen DP. Surrogate endpoints in assessment of new drugs in colorectal cancer. Lancet. Jul 29 2000;356(9227):353-354.
40. Buyse M, Thirion P, Carlson RW, Burzykowski T, Molenberghs G, Piedbois P. Relation between tumour response to first-line chemotherapy and survival in advanced colorectal cancer: a meta-analysis. Meta-Analysis Group in Cancer. Lancet. Jul 29 2000;356(9227):373-378.
41. Greene HL, Roden DM, Katz RJ, Woosley RL, Salerno DM, Henthorn RW. The Cardiac Arrhythmia Suppression Trial: first CAST ... then CAST-II. J Am Coll Cardiol. Apr 1992;19(5):894-898.
42. FDA Adviser Questions Surrogate Endpoints for Diabetes Drug Approvals. Medpage Today; 2007.
43. Boissel JP, Collet JP, Moleur P, Haugh M. Surrogate endpoints: a basis for a rational approach. Eur J Clin Pharmacol. 1992;43(3):235-244.
44. Montori VM, Busse JW, Permanyer-Miralda G, Ferreira I, Guyatt GH. How should clinicians interpret results reflecting the effect of an intervention on composite endpoints: should I dump this lump? ACP J Club. Nov-Dec 2005;143(3):A8.
45. Freemantle N, Calvert M, Wood J, Eastaugh J, Griffin C. Composite outcomes in randomized trials: greater precision but with greater uncertainty? Jama. May 21 2003;289(19):2554-2559.
46. Lauer MS, Topol EJ. Clinical trials--multiple treatments, multiple end points, and multiple lessons. Jama. May 21 2003;289(19):2575-2577.
47. Kip K, Hollabaugh K, Marroquin O, Williams D. The problem with composite endpoinmts in cardiovascular studies. J Am Coll Cardiol. 2008;51:701-707.
48. Connolly SJ, Gent M, Roberts RS, et al. Canadian implantable defibrillator study (CIDS) : a randomized trial of the implantable cardioverter defibrillator against amiodarone. Circulation. Mar 21 2000;101(11):1297-1302.
49. Bokhari F, Newman D, Greene M, Korley V, Mangat I, Dorian P. Long-term comparison of the implantable cardioverter defibrillator versus amiodarone: eleven-year follow-up of a subset of patients in the Canadian Implantable Defibrillator Study (CIDS). Circulation. Jul 13 2004;110(2):112-116.
50. Schwartz GG, Olsson AG, Ezekowitz MD, et al. Effects of atorvastatin on early recurrent ischemic events in acute coronary syndromes: the MIRACL study: a randomized controlled trial. Jama. Apr 4 2001;285(13):1711-1718.
51. Major outcomes in high-risk hypertensive patients randomized to angiotensin-converting enzyme inhibitor or calcium channel blocker vs diuretic: The Antihypertensive and Lipid-Lowering Treatment to Prevent Heart Attack Trial (ALLHAT). Jama. Dec 18 2002;288(23):2981-2997.
52. Starzl TE, Donner A, Eliasziw M, et al. Randomised trialomania? The multicentre liver transplant trials of tacrolimus. Lancet. Nov 18 1995;346(8986):1346-1350.
53. Ioannidis JPA. Why most published research findings are false. PLoS. 2005;2:0696-0701.
54. Toma M, McAlister FA, Bialy L, Adams D, Vandermeer B, Armstrong PW. Transition from meeting abstract to full-length journal article for randomized controlled trials. Jama. Mar 15 2006;295(11):1281-1287.

# Chapter 4
# Alternative Interventional Study Designs

**Stephen P. Glasser**

*A man who does not habitually wonder is but a pair of spectacles behind which there is no eye.*

Thomas Carlyle[1]

**Abstract** There are many variations to the classical randomized controlled trial. These variations are utilized when, for a variety of reasons, the classical randomized controlled trial would be impossible, inappropriate, or impractical. Some of the variations are described in this chapter and include: equivalence and non-inferiority trials; crossover trials; N of 1 trials, case-crossover trials, and externally controlled trials. Large simple trials, and prospective randomized, open-label, blinded endpoint trials are discussed in another chapter.

## Introduction

There are a number of variations of the 'classical' RCT design. For instance, many view the classical RCT as having an exposure group compared to a placebo control group, using a parallel design, and a 1:1 randomization scheme. However, in a given RCT, there may be several exposure groups (e.g. several different doses of the drug under study), and the comparator group may be an active control rather than a placebo control; and, some studies may have both. By an active control, it is meant that the control group receives an already approved intervention. For example, a new anti-hypertensive drug could be compared to placebo or could be compared to a drug already approved by the FDA and used in the community (frequently, in this case, the manufacturer of the investigational drug will compare their drug to the most frequently prescribed drug for the indication of interest). The decisions regarding the use of a comparator are based upon a number of considerations and discussed more fully under the topic entitled equivalence testing. Also, the randomization sequence may not be 1:1, particularly if (for several reasons, ethical issues may be one example) one wanted to reduce the number of subjects exposed to placebo. Also, rather than parallel groups there may be a titration schema built into the design. On occasion, the study design could incorporate a

S.P. Glasser (ed.), *Essentials of Clinical Research*,

placebo withdrawal period in which at the end of the double blind comparison, the intervention group is subsequently placed on placebo (this can be done single-blind or double-blind). In this latter case, retesting 1 or 2 weeks later occurs with comparison to the original placebo group. Other common variants to the classical RCT are discussed in more detail below.

## Traditional Versus Equivalence/Non-inferiority Testing

As discussed in Chapter 3, most clinical trials have been designed to assess if there is a difference in the efficacy to two (or more) alternative treatment approaches (with placebo ideally being the comparator treatment) (see Tables 3.6 and 4.1). Consider the fact that for evidence of efficacy there are two distinct approaches: to demonstrate a difference-showing superiority of test drug to control (placebo, active, lower dose) which then demonstrates the drug effect; or, to show equivalence or non-inferiority to an active control (i.e. the investigational drug is of equal efficacy or not worse than an active control). That is, one can attempt to demonstrate that there is similarity to a known effective therapy (active control) and attributing the efficacy of the active control drug to the investigational drug, thereby demonstrating a drug effect (i.e. equivalence). Since nothing is perfectly equivalent, equivalence means within a margin predetermined by the investigator (termed the equivalence margin). Non-inferiority trials on the other hand aim to demonstrate that the investigational drug is not worse than the control, but once again by a defined amount (i.e. not worse by a given amount – the non-inferiority margin), the margin (M or $\delta$) being that amount no larger than the effect the active control would be expected to have in the study. As will be discussed later, this margin is not easy to determine and requires clinical judgment; and, this represents one of the limitations of these kinds of trials.[2]

As discussed in Chapter 3, there are a number of reasons for the increased interest in equivalence and non-inferiority trials including the ethical issues associated with placebo controls. In general, placebo-controls are preferable to active controls, due to the placebo's ability to distinguish an effective treatment from a less effective treatment. The ethical issues surrounding the use of a placebo-control aside, there are other issues that have led to the increasing interest and use of equivalence and non-inferiority studies. For example, clinical trials are increasingly being required to show benefits on clinical endpoints rather than on surrogate endpoints

**Table 4.1** RCT hypothesis testing

| Question asked | Superior | Equivalence | Non-inferior |
|---|---|---|---|
| Null | A = B | A < B + margin | A not less than B |
| Alternative | A ≠ B (i.e. A < B or A > B | A ≥ B + margin | A = B |
| Rejection of null | A is different than B | A is equivalent to B | A is at least as effective as B |
| Failure to reject null | Did not show that A is different from B | Did not show that A is equivalent to B | Did not show that A is as effective as B |

at the same time that the incremental benefit of new treatments is getting smaller. This has led to the need for larger, longer, and more costly trials; and, this has resulted in the need to design trials less expensive. Additional issues are raised by the use of equivalence/non-inferiority trials, such as assay sensitivity, the aforementioned limitations of defining the margins, and the constancy assumption.

## *Assay Sensitivity*

Assay sensitivity is a property of a clinical trial defined as the ability of the trial to distinguish effective from ineffective treatments.[3] That is, assay sensitivity is the ability of a specific clinical trial to demonstrate a treatment difference if such a difference truly exists.[3] Assay sensitivity depends on the effect size one needs to detect. One, therefore, needs to know the effect of the control drug in order to determine the trials assay sensitivity. There is then an inherent, usually unstated, assumption in an equivalence/non-inferiority trial, namely that the active control was similarly effective in the particular study one is performing (i.e., that one's trial has assay sensitivity), compared to a prior study that utilized a placebo comparator. However, this aforementioned assumption is not necessarily true for all effective drugs, is not directly testable in the data collected (because there is no placebo group to serve as an internal standard); and thus, in essence, causes an active control equivalence study to have elements of a historically controlled study.[4]

A trial that demonstrates superiority has inherently demonstrated assay sensitivity; but, a trial that finds the treatments to be similar cannot distinguish (based upon the data alone) between a true finding, and a poorly executed trial that just failed to show a difference. Thus, an equivalence/non-inferiority trial must rely on the assumption of assay sensitivity, based upon quality control procedures and the reputation of the investigator. The International Conference on Harmonization (ICH) guidelines (see Chapter 6) list a number of factors that can reduce assay sensitivity, and include: poor compliance, poor diagnostic criteria, excessive measurement variability, and biased endpoint assessment.[5] Thus, assay sensitivity can be more directly ascertained in an active control trial only if there is an 'internal standard,' a control vs. placebo comparison as well as the control vs. test drug comparison (e.g. a three-arm study).

## *Advantages of the Equivalence/Non-inferiority Approach*

As discussed above, the application of equivalence testing permits a definitive statement that the new treatment is '*as good* or better' (if the null hypothesis is rejected), and depending upon the circumstances, this statement may meet the needs of the manufacturer, who may only want to make the statement that the new treatment is as good as the established treatment, with the implication that the new treatment is preferred because it may require less frequent dosing, or be associated with fewer side effects, etc. On the other hand, the advantage of superiority testing is that one can definitively state if one treatment is better (or worse) than the other, with the

downside that if there is not evidence of a difference, you cannot state that the treatments are the same (recall, that the null hypothesis is never 'accepted' – it is simply a case where it cannot be rejected, i.e. 'there is not sufficient evidence in these data to establish if a difference exists').

### *Disadvantages or Limitations of Equivalence/Non-inferiority Studies*

The disadvantages of equivalence/non-inferiority testing include: (1) that the choice of the margin chosen to define whether two treatments are equivalent or not inferior to one another; (2) requires clinical judgment and should have clinical relevance (variables that are difficult to measure); (3) the assumption that the control would have been superior to placebo (assumed assay sensitivity) had a placebo had been employed (constancy assumption – that is, one expects the same benefit in the equivalence/non-inferiority trial as occurred in a prior placebo controlled trial); and (4) having to determine the margin such that it is not greater than the smallest effect size (that of the active drug vs. placebo) in prior placebo controlled trials.[6] In addition there is some argument as to whether the analytic approach in equivalence/non-inferiority trials should be ITT or Per Protocol (Compliers Only).[7] While ITT is recognized as valid for superiority trials, the inclusion of data from patients not completing the study in equivalence/non-inferiority trials, could bias the results towards the treatments being the same, which could then result in an inferior treatment appearing to be non-inferior or equivalent. On the other hand, using the compliers only (per protocol) analysis may bias the results in either direction. Most experts in the field argue that the Per Protocol analysis is preferred for equivalence/non-inferiority trials but some argue for the ITT approach.[7] Also, blinding does not protect against bias as much in equivalence/non-inferiority trials as it does with superiority trials-since the investigator, knowing that the trial is assessing equality may subconsciously assign similar ratings to the treatment responses of all patients.

## The Null Hypothesis in Equivalence/Non-inferiority Trials

"*It is a beautiful thing, the destruction of words…Take 'good' for instance, if you have a word like 'good' what need is there for the word "bad"? 'Ungood' will do just as well*"[8]

Recall that with traditional hypothesis testing, the null hypothesis states that 'there is no difference between treatment groups (i.e. New = Established, or placebo). Rejecting the null, then allows one to definitively state if one treatment is better than another (i.e. New > or < Established). The disadvantage is if at the conclusion of an RCT there is not evidence of a difference, one cannot state that the treatments are the same, or as good as one to the other.

Equivalence/non-inferiority testing in essence 'flips' the traditional null and alternative hypotheses. Using this approach, the null hypothesis is that the new treatment is worse than the established treatment (i.e. New < Old); that is, rather than assuming that there is no difference, the null hypothesis in equivalence/non-inferiority trials is that a difference exists and the new treatment is inferior. Just as in traditional testing, the two actions available resulting from statistical testing are (1) reject the null hypothesis, or (2) failure to reject the null hypothesis. However, with equivalence testing, rejecting the null hypothesis is making the statement that the new treatment is not worse than established treatment, implying the alternative, that is, that the new treatment is as good as (or better than the established i.e. New ≥ Established). Hence, this approach allows a definitive conclusion that the new treatment is at least as good, if not better, or is not inferior to the established.

As mentioned before, a caveat is the definition of 'as good as,' which is defined as being in the 'neighborhood' or having a difference that is so small as to be considered clinically unimportant (generally, event rates within ±2% – this is known as the equivalence or non-inferiority margin usually indicted by the symbol δ). The need for this 'neighborhood' that is considered 'as good as' exposes the first shortcoming of equivalence/non-inferiority testing – having to make a statement that "I reject the null hypothesis that the new treatment is worse than the established, and accept the alternative hypothesis that it is as good or better – and by that I mean that it is within at least 2% of the established" (the wording in italics are rarely included in the conclusions of a manuscript). A second caveat of equivalence/non-inferiority testing is that no definitive statement can be made that there is evidence that the new treatment is worse. Just as in traditional testing, one never accepts the null hypothesis – one only fails to reject it. Hence if the null is not rejected, all one can really say is that there is no evidence in these data that the new treatment is as good as or better than the old treatment.

In summary, one might ask, which is the 'correct' approach, traditional, equivalence, or non-inferiority testing? There is simply no general answer to this question; rather, the answer depends on the major goal of the study. But, once an approach is taken, the decision cannot be changed in post-hoc analysis. That is, the format of the hypotheses has to be tailored to the major aims of the study and must then be followed.

## *Crossover Design*

In crossover designs, both treatments (investigational and control) are administered sequentially to all subjects, and randomization occurs in terms of which treatment each patient receives first. In this manner each patient serves as their own control. The two treatments can be an experimental drug vs. placebo or an experimental drug compared to an active control. The value of this approach beyond being able to use each subject as their own control, centers on the ability (in general) to use smaller sample sizes. For example, a study that might require 100 patients in a par-

allel group design might require fewer patients in a crossover design. But like any decision made in clinical research there is always a 'price to pay.' For example, the washout time between the two treatments is arbitrary, and one has to assume that they have eliminated the likelihood of carryover effects from the first treatment period (plasma levels of the drug in question are usually used to determine the duration of the crossover period, but in some cases the tissue level of the drug-not measured clinically – is more important). Additionally, there is some disagreement as to which baseline period measurement (the first baseline period or the second baseline period – they are almost always not the same) should be used to compare the second period effects.

## N of 1 Trials

During a clinical encounter, the benefits and harms of a particular treatment are paramount; and, it is important to determine if a specific treatment is benefiting the patient or if a side effect is the result of that treatment. This is particularly a problem if adequate trials have not been performed regarding that treatment. Inherent to any study is the consideration of why a patient might improve as a result of an intervention. Of course, what is generally hoped for is that the improvement is the result of the intervention. However, improvement can also be a result of the disease's natural history, placebo effect, or regression to the mean (see Chapter 7). Clinically, a response to a specific treatment is assessed by a trial of therapy, but this is usually performed without rigorous methodological standards so the results may be in question; and, this has led to the N of 1 trial (sometimes referred to as an RCT crossover study in a single patient at a time). The requirements of this study design are: the patient receives active, investigational therapy during one period, and alternative therapy during another period. As is true of crossover designs, the order of treatment from one patient to another is randomly varied, and other attributes-blinding/masking, ethical issues, etc. – are adhered to just as they are in the classical RCT.

## Factorial Designs

Many times it is possible in one trial to evaluate two or even three treatment regimens in one study. In the Physicians Health Study, for example, the effect of aspirin and beta carotene were assessed.[9] Aspirin was being evaluated for its ameliorating effect on myocardial infarction, and beta carotene on cancer. Subjects were randomized to one of four groups; placebo and placebo, aspirin and placebo, beta carotene and placebo, and aspirin plus beta carotene. In this manner, each drug could be compared to placebo, and any interaction of the two drugs in combination could also be evaluated. This type of design certainly can add to the efficiency of a trial,

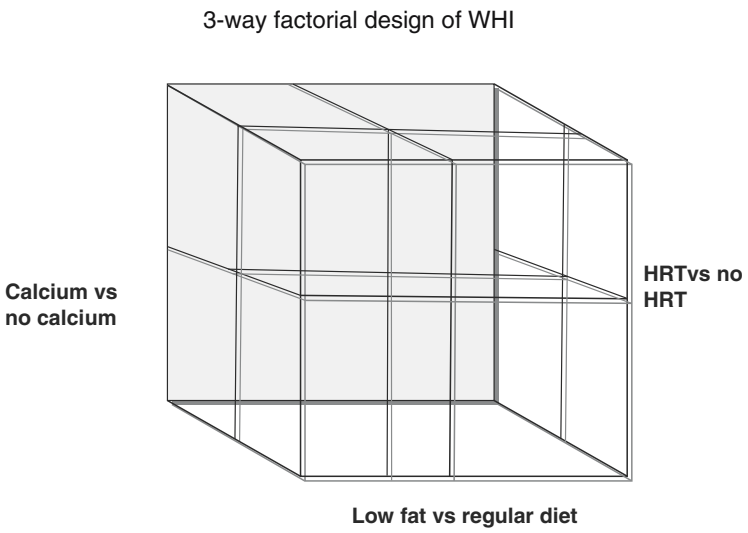



**Fig. 4.1** Three-way factorial design of WHI

but this is counterbalanced by increased complexity in performing and interpreting the trial results. In addition, the overall trial sample size is increased (four randomized groups instead of the usual two), but the overall sample size is likely to be less than the total of two separate studies, one addressing the effect of aspirin and the other of beta carotene. In addition two separate studies would lose the ability to evaluate treatment interactions, if that is a concern. Irrespective, costs (if it is necessary to answer both questions) should be less with a factorial design compared to two separate studies, since recruitment, overhead etc. should be less. The Woman's Health Initiative is an example of a three-way factorial design.[10] In this study, hormone replacement therapy, calcium/vitamin D supplementation, and low fat diets are being evaluated (see Fig. 4.1). Overall, factorial designs can be seductive but can be problematic, and it is best used for unrelated research questions, both as it applies to the intervention as well as the outcomes.

## Case Crossover Design

Case cross over designs are a variant of a RCT designed with components of a crossover, and a case-control design. The case cross over design was first introduced by Maclure in 1991.[11] It is usually applied to study transient effects of brief exposures on the occurrence of a 'rare' acute onset disease. The presumption is that if there are precipitating events, these events should be more frequent during the period immediately preceding the event, than at a similar period which is more distant from the event. For example, if physical and/or mental stress trigged sudden

cardiac death (SCD), one should find that SCD occurred more frequently during or shortly after these stressors. In a sense, it is a way of assessing whether the patient was doing anything unusual just before the outcome of interest. As mentioned above, it is related to a prospective crossover design in that each subject passes through both the exposure (in the case-crossover design this is called the hazard period) and 'placebo' (the control period). The case cross over design is also related to a case-control study in that it identifies cases and then looks back for the exposure (but in contrast to typical case-control studies, in the case-crossover design the patient serves as their own control). Of course, one needs to take into account the times when the exposure occurs but is not followed by an event (this is called the exposure-effect period). The hazard period is defined empirically (one of this designs limitations, since this length of time may be critical yet somewhat arbitrary) as the time period before the event (say an hour or 30 minutes) and is the same time given to the exposure-effect period. A classic example of this study design was reported by Hallqvist et al., where the triggering of an MI by physical activity was assessed.[12] To study possible triggering of first events of acute myocardial infarction by heavy physical exertion, Halqvist et al. conducted a case-crossover analysis. Interviews were carried out with 699 myocardial infarction patients after onset of the disease. The relative risk from vigorous exertion was 6.1 (95% confidence interval: 4.2, 9.0), while the rate difference was 1.5 per million person-hours.[12]

In review, the strengths of this study design include using subjects as their own control (self matching decreases between-person confounding, although if certain characteristics change over time there can be individual confounding), and improved efficiency (since one is analyzing relatively rare events). In the example of the Halqvist study, although MI is common, MI just after physical exertion is not.[12] Weaknesses of the study design, besides the empirically determined time for the hazard period, include: recall bias, and that the design can only be applied when the time lag between exposure and outcome is brief and the exposure is not associated with a significant carryover effect.

## Externally Controlled Trials (Before-After Trials)

Using historical controls as a comparator to the intervention is problematic, since the natural history of the disease may have changed over time, and certainly sample populations may have changed (e.g. greater incidence of obesity, more health awareness, new therapies, etc. now vs. the past). However, when an RCT with a concomitant control cannot be used (this can occur for a variety of reasons-see example below) there is a way to use a historical control that is not quite as problematic. Olson and Fontanarosa cite a study by Cobb et al to address survival during out of hospital ventricular fibrillation.[13] The study design included a pre-intervention period (the historical control) during which emergency medical technicians (EMT) administered defibrillation as soon as possible after arriving on scene of a patient in cardiac arrest. This was followed by an intervention period where the

EMT performed CPR for 90 seconds before defibrillation. In this way many of the problems of typical historical controls can be overcome in that in the externally controlled design, one can use the same sites and populations in the 'control' and intervention groups as would be true of a typical RCT, it is just that the control is not concomitant.

## Large Simple Trials (LSTs) and Prospective, Randomized, Open-Label, Blinded Endpoint Designs (PROBE)

In summary, in this chapter, various clinical research study designs were discussed, and the differing 'levels of scientific evidence' that are associated with each were addressed. A comparison of study designs is complex, with the metric being that the study design providing the highest level of scientific evidence is the one that yields the greatest likelihood of implying causation. The basic tenet of science is that it is almost impossible to absolutely prove something, but it is much easier to disprove it. Causal effect focuses on outcomes among exposed individuals; but, what would have happened had they not been exposed? Causality is further discussed in the chapter on Associations, Cause, and Correlations (Chapter 16).

## References

1. Cited in Breslin JEcb. *Quote Me*. Ontario, CA: Hounslow Press; 1990.
2. Siegel JP. Equivalence and noninferiority trials. *Am Heart J*. Apr 2000; 139(4):S166–170.
3. Assay Sensitivity. *Wikipedia*.
4. Snapinn SM. Noninferiority trials. *Curr Control Trials Cardiovasc Med*. 2000; 1(1):19–21.
5. The International Conference on harmonization (ICH) Guidelines.
6. D'Agostino RB Sr., Massaro JM, Sullivan LM. Non-inferiority trials: design concepts and issues – the encounters of academic consultants in statistics. *Stat Med*. Jan 30, 2003; 22(2):169–186.
7. Wiens BL, Zhao W. The role of intention to treat in analysis of noninferiority studies. *Clin Trials*. 2007; 4(3):286–291.
8. Diamond GA, Kaul S. An orwellian discourse on the meaning and measurement of noninferiority. *Am J Cardiol*. Jan 15, 2007; 99(2):284–287.
9. Hennekens CH, Eberlein K. A randomized trial of aspirin and beta-carotene among U.S. physicians. *Prev Med*. Mar 1985; 14(2):165–168.
10. Rossouw JE, Anderson GL, Prentice RL, et al. Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results from the women's health initiative randomized controlled trial. *JAMA*. July 17, 2002; 288(3):321–333.
11. Maclure M. The case-crossover design: a method for studying transient effects on the risk of acute events. *Am J Epidemiol*. Jan 15, 1991; 133(2):144–153.
12. Hallqvist J, Moller J, Ahlbom A, Diderichsen F, Reuterwall C, de Faire U. Does heavy physical exertion trigger myocardial infarction? A case-crossover analysis nested in a population-based case-referent study. *Am J Epidemiol*. Mar 1, 2000; 151(5):459–467.
13. Olson CM, Fontanarosa PB. Advancing cardiac resuscitation: lessons from externally controlled trials. *JAMA*. Apr 7, 1999; 281(13):1220–1222.

# Chapter 5
# Postmarketing Research*

**Stephen P. Glasser, Elizabeth Delzell, and Maribel Salas**

**Abstract** In the past, postmarketing research, postmarketing surveillance and pharmacovigilance were synonymous with phase IV studies because the main activities of the regulatory agency (e.g. FDA) were focused on the monitoring of adverse drug events and inspections of drug manufacturing facilities and products. (1) However, the fact that not all FDA mandated (classical phase IV trials) research consists of randomized controlled trials (RCTs), and not all postmarketing activities are limited to safety issues (pharmacovigilance), these terms require clarification. This chapter attempts to clarify the confusing terminology; and, to discuss many of the postmarketing research designs-both their place in clinical research as well as their limitations.

## Introduction

In the past, postmarketing research, postmarketing surveillance and pharmacovigilance were synonymous with phase IV studies because the main activities of the regulatory agency (e.g. FDA) were focused on the monitoring of adverse drug events and inspections of drug manufacturing facilities and products.[1] However, the fact that not all FDA mandated (classical phase IV trials) research consists of randomized controlled trials (RCTs), and not all postmarketing activities are limited to safety issues (pharmacovigilance), these terms require clarification. Information from a variety of sources is used to establish the efficacy and short-term safety (< 3 years) of medications used to treat a wide range of conditions. Premarketing studies (Table 5.1) consist of phase I–III trials, and are represented by pharmacokinetic and pharmacodynamic studies, dose ranging studies, and for phase III trials the gold standard randomized, placebo-controlled (or active controlled), double blind, trial (RCT). Approximately only 20% of the drugs that enter phase I are approved for marketing.[1] RCTs remain the 'gold standard' for assessing the efficacy and to a lesser extent, the safety of new therapies[2,3]; however, they do have significant limitations that promote caution in generalizing their results to routine clinical practice.

* Over 50% of this chapter is taken from "Importance and challenges of studying marketed drugs: what is a phase IV study? Common clinical research designs, registries, and self-reporting systems".[8] With permission of the publisher.

S.P. Glasser (ed.), *Essentials of Clinical Research*,

**Table 5.1** Premarketing study designs for FDA approval

I. Phase I–III studies
  a. Pharmacokinetic and pharmacodynamic studies
  b. Dose-ranging studies
  c. RCTs (efficacy studies)
    1. With or without crossover designs
    2. Drug withdrawal designs
    3. Placebo or active controls

**Table 5.2** Estimated necessary study size to find adverse events

| Frequency of adverse events (%) | Number of patients | Trial type |
|---|---|---|
| 1 | 1,000 | Clinical trial |
| 0.1 | 10,000 | Large clinical trial |
| 0.01 | 100,000 | Postmarket survey |
| 0.001 | 1,000,000 | Long-term survey |

For example, because of the strict inclusion and exclusion criteria mandated in most controlled studies a limited number of patients who are relatively homogeneous are enrolled. Elderly patients, women, and those deemed not competent to provide informed consent are often excluded from such trials.[4–7] RCTs may also suffer from selection or volunteer bias. For example, clinical studies that include extended stays in a clinic may attract unemployed patients, and studies that involve a free physical examination may attract those concerned that they are ill. Studies that offer new treatments for a given disease may inadvertently select patients who are dissatisfied with their current therapy.[7]

RCTs have other limitations as well. For example, the stringent restrictions regarding concomitant medications and fixed treatment strategies bear only modest resemblance to the ways in which patients are treated in actual practice.[2,9] This difference creates a situation dissimilar from routine clinical practice in which many or even most patients are taking multiple prescription and over-the-counter medications or supplements to manage both acute and chronic conditions.[10,11] RCTs also generally include intensive medical follow-up in terms of number of medical visits, number and/or type of tests and monitoring events, that is usually not possible in routine clinical care.[12] Also, unintended adverse events (UAEs) are unlikely to be revealed during phase III trials since the usual sample sizes of such studies and even the entire NDA may range from hundreds to only a few thousand patients. For example, discovering an UAE with a frequency of 0.1% would require a sample size of more than 10,000 participants (Table 5.2). Castle[13] further elaborated on this issue by asking the question ‘how large a population of treated patients should be followed up to have a good chance of picking up one, two, or three cases of an adverse reaction?’ He notes that if one defines ‘good chance’ as a 95% probability, one has to still factor in the expected incidence of the adverse event. If one assumes no background incidence of adverse event, and the expected incidence is 1 in 10,000, then by his assumptions, it would require 65,000 patients to pick up an excess of three adverse events.

Phase III trials also are not useful for detecting UAEs that occur only after long-term therapy because of insufficient length of follow-up time of the majority of

phase III trials, nor do they provide information on long-term effectiveness and safety. All of the restrictions characteristic of controlled clinical studies may result in overestimation of the efficacy and underestimation of the potential for UAEs of the medication being evaluated.[9,12,14] As a result of these limitations, additional complementary approaches to evaluation of medication efficacy, effectiveness and safety are taking on increasing importance.

Postmarketing research (Table 5.3) is a generic term used to describe all activities after the drug approval by the regulatory agency, such as the Food and Drug Administration (FDA). Postmarketing studies concentrate much more (but not exclusively) on safety and effectiveness and they can contribute to the drugs implementation through labeling changes, length of the administrative process, pricing negotiations and marketing. The most commonly used approaches for monitoring drug safety are based on spontaneous reporting systems, automated linkage data, patient registries, case reports, and data obtained directly from a study. Since there are major limitations from relying on case reports on voluntary reporting, postmarketing research has become an integral part of the drug evaluation process for assessing adverse events.[15–20] However, falling under the rubric of postmarking research is a wide variety of study designs and approaches, each with its own strengths and limitations. Postmarketing studies (Fig. 5.1; Table 5.3) are not only represented by a much broader array of study designs, they have clearly differentiated goals compared to premarketing studies. Examples of study designs that might fall under the rubric of postmarketing research are phase IV clinical trials, practice-based

**Table 5.3** Postmarketing study designs

I. FDA 'Mandated or Negotiated' Studies (phase IV)
  (a) Any study design may be requested including studies of
    (i) Drug-drug interactions
    (ii) Formulation advancement
    (iii) Special safety
    (iv) Special populations (e.g. elderly, pediatrics, etc.)
  (b) 'Phase V' trials
II. Non FDA 'Mandated or Negotiated' Studies
  (a) RCTs
    (i) Superiority vs. equivalence testing
    (ii) Large simple trials
    (iii) PROBE designs
    (iv) 'Phase V' trials
  (b) Surveillance studies
    (i) Pharmacovigilance studies
    (ii) Effectiveness studies
    (iii) Drug utilization studies
    (iv) Observational epidemiology studies
III. Health Services Research (HSR)
IV. Health Outcomes Research (HOR)
V. Implementation Research

Note: we have not included a discussion of HSR or HOR in this review. Implementation Research will be discussed in Chapter 13

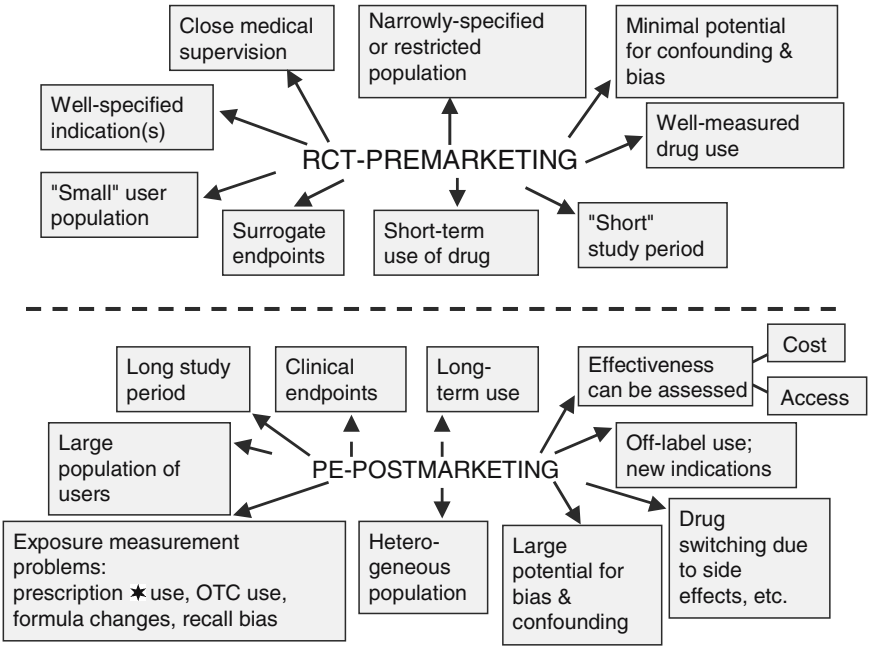



**Fig. 5.1** Contrasts between pre- and post-marketing studies

clinical experience studies, large simple trials (LSTs), equivalence trials, post-marketing surveillance studies such as effectiveness studies, pharmacovigilance studies, and pharmacoeconomic studies.

There are several initiating mechanisms for postmarketing studies: (1) those required by a regulatory agency as a condition of the drug's approval (these are referred to as postmarketing commitments or PMCs; (2) those that are initiated by the pharmaceutical company to support various aspects of the development of that drug; (3) investigator initiated trials that may be as scientifically rigorous as phase III RCTs, but occur after drug approval (a recent example is some of the Vioxx studies that ultimately questioned the drugs safety); and (4) investigator initiated observational studies. The more scientifically rigorous postmarketing studies (particularly if they are RCTs) are sometime referred to as 'phase V' trials. This review will discuss each of the common types of postmarketing research studies and examples will be provided in order to highlight some of the strengths and limitations of each.

## FDA 'Mandated or Negotiated' Studies (Phase IV Studies)

Phase IV studies are most often concerned with safety issues and usually have prospectively defined end points aimed at answering these questions. Any type of study (these include standard RCTs, observational studies, drug-drug interaction

studies, special population studies, etc. – see Table 5.3) may be requested by the FDA upon NDA (New Drug Application) approval, and these are frequently called Phase IV Post Marketing Commitment Studies (PMCs). Phase IV PMCs are studies required of, or agreed to (i.e. 'negotiated'), by the sponsor at the time of NDA approval and this is particularly true of those drugs that have had accelerated approval. Phase IV clinical trials usually include large and more heterogeneous population than phase III trials with emphasis on the replication of usual clinical care conditions.[21] For some special populations, phase IV commitment trials represent a unique opportunity to determine safety and efficacy of a drug.[22] This is particularly important for pediatric population because only a small fraction of all drugs approved in the United States have been studied in pediatric patients, and more than 70% of new molecular entities were without pediatric labeling. Adequate designed phase IV clinical trials will impact drug utilization and prescriber's decisions particularly in children. For example, Lesko and Mitchell designed a practitioner-based, double-blind, randomized trial in 27,065 children younger than 2 years old to compare the risk of serious adverse clinical events of ibuprofen versus acetaminophen suspension. They found small risk of serious adverse events and no difference by medication.[23] Phase IV commitments trials have also been used in exploratory special population studies, such as neonatal abstinence syndrome,[24] and pregnant opiate-dependency[25,26] In those studies, the main research question is focused on the efficacy and/or safety of a drug in small number of patients. For example, in the pregnant-opiate dependent study, Jones successfully transferred four drug-dependent pregnant inpatients from methadone to morphine and then buprenorphine.[27]

An analysis of phase IV studies during 1987–1993 showed that each of the phase IV drugs had, on average, a commitment to conduct four studies.[24] The regulations regarding phase IV studies began in 1997 as part of the FDA Modernization Act. As a result of that act, the FDA was required to report annually on the status of postmarketing study commitments. In 1999 (a rule which became officially effective in 2001), the FDA published rules and formatting guidelines for the phase IV reports. Although these studies are a 'requirement' of NDA approval and are called 'commitment' studies, significant problems exist. In March 2006, the Federal Register reported on the status of postmarketing study commitments. Of 1,231 commitments, 787 were still pending (65%), 231 were ongoing, and only 172 (14%) were completed. The problem associated with these studies has been extensively discussed. For example, a recommendation by Public Citizen (a public advocacy group) followed the release of this FDA report, and noted that the FDA needs the ability to impose financial penalties as an incentive for drug companies to submit required annual postmarket study reports on time. Peter Lurie, deputy director of Public Citizen's Health Research Group, told FDA news; 'The only thing the agency can do is take the drug off the market, which is a decision that often would not serve the public health very well,' he said.[28] In addition, the only mechanism that was available to remove a drug from the market was through a difficult legal channel. The FDA did not have the authority itself to withdraw a drug from the market, or suspend sales of a drug. In fact, the FDA could not even compel

completion of a post-marketing study agreed upon at the time of approval, limit advertising of the drug, compel manufacturer to send out 'Dear Doctor' letters, or revise the product label of a drug without the approval of the company involved. Lurie noted that 'the great majority of postmarketing studies address safety issues, at least in part, so patients and physicians are denied critical safety information when these studies are not completed in a timely fashion.' Lurie also criticized the FDA's report on the status of postmarketing commitments, noting there is no way of knowing what the deadlines are for each stage of the commitment and if they are being met or not, and for inadequate tracking system for those who are initiating and those ongoing trials. In the past, the FDA set the schedule for firms to complete a battery of studies on products that require a phase IV study. The agency then evaluated each study to see if the drug company had fulfilled the requirements of the study commitment. If the company failed to submit data on time, the commitment was considered delayed. The reports were to contain information on the status of each FDA-required study specifically for clinical safety, clinical efficacy, clinical pharmacology, and non-clinical toxicology. The pharmaceutical firm then continued to submit the report until the FDA determined that the commitment had been fulfilled or that the agency no longer needed the reports.

In 2007, the FDA Amendments Act of 2007 was signed into law. Among other things, the Law addressed the need for ongoing evaluations of drug safety after drug approval, a way of addressing safety signals and performing high quality studies addressing those signals, new authority to require post marketing studies, civil penalties for non-compliance, the registration of all phase II–IV trials, and the designation of some of the user's fees (10%) to be earmarked for safety issues.

Some examples of phase IV studies follow.

## Practice Based Clinical Experience Studies

Physician Experience Studies (PES) may be mandated by the FDA or initiated by the pharmaceutical company that has marketed a particular drug. The name is descriptive of the intent of the study and it is most often associated with the phase IV study. PES is generally not a RCT, and therefore has been most often criticized for its lack of scientific rigor. It does, however, in addition to providing physicians with experience in using a newly marketed drug, expose a large number of patients to the drug, potentially providing 'real world' information about the drugs adverse event profile.

An example of a recently reported PES is that of graded release diltiazem. The Antihypertensive Safety and Efficacy and Physician and Patient Satisfaction in Clinical Practice: Results from a Phase IV Practice-based Clinical Experience Trial with Diltiazem LA (DLA). The study enrolled a total of 139,965 patients with hypertension, and involved 15,155 physicians who were to perform a baseline evaluation and two follow-up visits.[26] Usual care treatment any other drug therapy was allowed as long as they were candidates for the addition of DLA. The potential to record efficacy and safety data for this large number of 'real world' patients was

great. However, as a characteristic of these kinds of studies, only 50,836 (26%) had data recorded for all three visits, and data on ADEs were missing for many as well. On the other hand, ADEs for 100,000 patients were collected, and none of the ADEs attributed to DLA were reported in more than 1% of patients, supporting the general safety profile of DLA.

## Non FDA Studies

Non FDA mandated postmarketing studies may utilize the wide array of research designs available and should not be confused with phase IV or PES studies. Examples of postmarketing studies include (1) RCTs with superiority testing, equivalence testing, or non-inferiority testing; large simple trials, 'phase V' trials; and (2) surveillance studies such as effectiveness studies, drug utilization trials, epidemiologic observational studies that usually concentrate on a safety profile of a drug, and classical RCTs. Not included in this review are health services research and health outcomes research which can also be studies of marketed drugs. Following is a discussion of some of the more common postmarketing research study designs. Postmarketing research falls under the umbrella of pharmacoepidemiologic studies (see Chapter 12).

## Equivalence and non-inferiority trials are discussed in chapters 3 and 4 Large Simple Trials

Not infrequently, an already marketed drug needs to be evaluated for a different condition than existed for its approval, or at a different dose, different release system, etc. In the aforementioned instance, the FDA might mandate a phase IV RCT that has all the characteristics of a classical phase III design. Some have suggested that this be termed a phase V study to distinguish it from the wide variety of other phase IV trials with all their attendant limitations and negative perceptions.

One type of postmarketing research is the Large Simple Trial (LST). The concept of large simple clinical trials has become more popular. The idea is that it is increasingly necessary to just demonstrate modest benefits of an intervention, particularly in common conditions. The use of short-term studies, implemented in large populations is then attractive. In these types of trials, the presumption is that the benefits are similar across participant types, so that the entry criteria can be broad, and the data entry and management can be simplified, and the cost thereby reduced. This model further depends on a relatively easily administered intervention and an easily ascertained outcome; but if these criteria are met, the size of the study also allows for a large enough sample size to assess less common ADEs. An example of the organization for this type of trial is the Clinical Trial of Reviparin and Metabolic Modulation of Acute Myocardial Infarction (CREATE), as discussed

by Yusuf et al.[29] In this trial over 20,000 subjects from 21 countries were enrolled in order to compare two therapies-glucose-insulin-potassium infusion, and low molecular weight heparin.

## Prospective, Randomized, Open-Label, Blinded Endpoint (PROBE) Design

A variation of the LST that also addresses a more 'real-world' principal is the prospective randomized open-label blinded endpoint design (PROBE design). By using open-label therapy, the drug intervention and its comparator can be clinically titrated as would occur in a doctor's office as compared to the fixed dosing of most RCTs. Of course, blinding is lost with the PROBE design, but only as to the therapy. Blinding is maintained as to the outcome. To test whether the use of open-label vs. double-blind therapy affected outcomes differentially, a meta-analysis of PROBE trials and double-blind trials in hypertension was reported by Smith et al.[30] They found that changes in mean ambulatory blood pressure from double-blind controlled studies and PROBE trials were statistically equivalent.

## Surveillance Studies

Pharmacovigilance deals with the detection, assessment, understanding and prevention of adverse effects or other drug-related problems. Traditionally, pharmacovigilance studies have been considered as part of the postmarketing phase of drug development because clinical trials of the premarketing phase are not powered to detect all adverse events particularly uncommon adverse effects. It is known that in the occurrence of adverse drug reactions other factors are involved such as the individual variation in pharmacogenetic profiles, drug metabolic pathways, the immune system, and drug-drug interactions. Additionally, the dose range established in clinical trials is not always representative of that used in the postmarketing phase. Cross, et al. analyzed the new molecular entities approved by FDA between 1980 and 1999 and they found that dosage changes occurred in 21% of the approved entities, and of these, 79% were related to safety. The median time to change following approval ranged from 1 to 15 years and the likelihood of a change in dosage was three times higher in new molecular entities approved in the nineties compared to those approved in the eighties,[31] and this would suggest that a wider variety of dosages and diverse populations need to be included in the premarketing phase and/or additional studies should be requested and enforced in the postmarketing phase. Further amplifying this point is a recent FDA news report[32] in which it was noted that there had been 45 Class I recalls (very serious potential to cause harm, injury, or death) in the last fiscal year (in many of the past years there had been only one or two such recalls) and also 193 Class II recalls (potential to cause harm).

Recently, a clinical trial in 8,076 patients with rheumatoid arthritis that examined the association of rofecoxib (Vioxx) vs. naproxen on the incidence of gastrointestinal events reported higher percentage of incident myocardial infarction in the arm of rofecoxib compared to naproxen during a median follow-up of 9 months,[33,34] which questioned the drug safety of COX 2 inhibitors. Then, the cardiac toxicity was corroborated in a metanalysis[35] database studies,[34] and in the APPROVe trial (Adenomatous Polyps Prevention on Vioxx),[36] a study in which cardiovascular events were found to be associated with rofecoxib in a colorectal adenoma chemoprevention trial.[34] The APPROVe trial is an example of phase IV trial that was organized for another potential indication of rofecoxib, the reduction of the risk of recurrent adenomatous polyps among patients with a history of colorectal adenomas. In that multicenter, randomized, placebo-controlled, double-blind study, 2,600 patients with history of colorectal adenoma was enrolled but after 3,059 patient-years of follow-up there was an increased risk of cardiovascular events. All of the above evidence resulted in the final decision of the manufacturer to withdraw rofecoxib from the market.[37]

The type of scandals that are associated with drug safety and the pressure of the society have contributed to the development of initiatives for performing more pharmacovigilance studies. Some countries, for example, are now requiring manufacturers to monitor the adverse drug events of approved medications. In France for example, manufacturers must present a pre-reimbursement evaluation and a post-marketing impact study.[38] In fact, France has a policy for the overall assessment of the public health impact of new drugs.[38]

In the United States, the recent withdrawals from the market (particularly for drugs that were approved through the expedited process by the FDA) indicate a need to start pharmacovigilance programs at the earliest stages of drug development, encouraging the identification of safety signals, risk assessment, and communication of those risks. The FDA has started developing algorithms to facilitate detection of adverse-event signals using the 'MedWatch', a spontaneous reporting adverse event system, to institute risk-management measures.

The 'MedWatch' is a voluntary system where providers, patients or manufacturers can report serious, undesirable experiences associated with the use of a medical product in a patient. An event is considered serious if it is associated with patient's death or increases the risk of death; the patient requires hospitalization, the product causes disability, a congenital anomaly occurs, or the adverse event requires medical or surgical intervention to prevent permanent impairment or damage.[39] The main obstacle of MedWatch is the high rate of underreporting adverse drug reactions which is then translated into delays in detecting adverse drug reactions of specific drugs.[40,41] Adverse events that are associated with vaccines or with veterinary products are not required to be reported to the Medwatch. The FDA revises those reports and determines if more research is needed to establish a cause-effect relationship between the drug and the adverse event. Then, the FDA defines the actions that manufacturers, providers and patients should take.

Another consequence from the recent drug withdrawals is the release of more safety information form the FDA to the public and press, as well as the creation of a new board to help monitoring drugs.[42]

**Table 5.4** Efficacy vs effectiveness

| | Efficacy | Effectiveness |
|---|---|---|
| Objective | Optimal | Usual |
| Motivation | FDA approval | Formulary |
| Intervention | Fixed regimen | Flexible |
| Comparator | Placebo | Usual care |
| Design | RCT | Open label |
| Subjects | Selected, compliant | Anyone |
| Outcomes | Condition | Comprehensive |
| Other | Short term, MOA | Long term |

As mentioned before, one of the limitations of phase III RCTs is their limited generalizability. Although the RCT may be the best way to evaluate efficacy under optimal conditions, it may not accurately reflect the drugs effectiveness under usual case ('real world') conditions. Clearly, clinical practice would follow evidence-based medicine which is derived from the RCT and meta-analyses of RCTs. But often the outcomes of clinical practice are not equal to that of the RCTs (due to differences in patients, the quality of the other treatments they receive, drug-drug and drug-disease interactions they may experience-these being much more common in the heterogeneity of clinical practice patients compared to the highly selected clinical trial patients). It is in this aforementioned setting that Effectiveness Trials are increasingly important. As is true of any decision made in research, there is always trade-offs (compromises) one has to make. While effectiveness trials may improve generalizability, it does so at the expense of internal validity. Table 5.4 contrasts important considerations between efficacy and effectiveness studies. An example of some of these issues was reported by Taylor et al.[43] The British Association for Cardiac Rehabilitation performs an annual questionnaire of the 325 cardiac rehabilitation programs in the UK. Taylor et al. compared the patient characteristics and program details of this survey with RCTs included in the 2004 Cochrane review. They found 'considerable differences' between the RCTs of cardiac rehabilitation and the actual practice in the UK (Table 5.5), differences suggesting that the real world practice of cardiac rehabilitation is unlikely to be as effective as clinical trials would suggest.

## Drug Utilization and Pharmacoeconomic Studies

One of the main reasons to conduct postmarketing studies is to demonstrate the economic efficiency of prescribing a new drug. In this instance, the manufacturer is interested in showing the relationship of risks, benefits and costs involved in the use of a new drug in order to show the value for the products cost. That value is essential for decision makers and prescriber's, who will select medications for formularies or prescribe the most appropriate medication for patients.

**Table 5.5** Comparison of Clinical Trials and Actual Practice of British Cardiac Rehabilitation Programs (Br J Cardiol © 2007 Sherboune Gibbs, Ltd.)

| | www.medscape.com | | |
|---|---|---|---|
| Medscape® | Cochrane report | British Association Of Cardiac Rehabilitation survey | Coronary Prevention Group survey |
| **Population characteristics** | | | |
| Mean age (SD) | 54.3 years (3.9) | 64.2 years (11.6) | Unknown |
| Women (SD) | 10.4% (14.1) | 26.4% | Unknown |
| Myocardial Infarction | 86% | 53% | Unknown |
| Coronary artery bypass graft | 6% | 24% | |
| Percutaneous transluminal coronary angloplasty | 5% | 13% | |
| **Intervention characteristics** | | | |
| Exercise-only programmes (%) | 17/44 (39%) | 0/242 (0%) | 0/28 (0%) |
| Overall duration (SD) | 18 weeks (21) | 7.5 weeks (3.2) | 7 weeks (2.1) |
| Mean exercise duration/session, minutes | 58 | Unknown | 60 |
| Mean frequency exercise sessions/week | 2.80 | 1.66 | 1.67 |
| Mean exercise intensity, %$VO_2$ or HR max | 75 | Unknown | Unknown |
| Mean number of sessions | 50 | 12.4 | 12 |
| Hospital based (%) | 40/44 (91%) | 166/302 (66%) | 28/28 (100%) |

**Key:** $VO_2$ = estimated peak oxygen consumption per minute; HR = heart rate

Most of the pharmacoeconomic studies have been carried out in the postmarketing phase using modeling techniques. Simulation models are mathematical abstractions of reality, based on both assumptions and judgements.[44] Those models are built using decision analysis, state transition modeling, discrete event simulation and survival modeling techniques.[45] The aforementioned models could allow for the adjustment of various parameters in outcomes and costs, and could explore the effect of changes in healthcare systems and policies if they clearly present and validate the assumptions made. Unfortunately, many economic models have issues related to model building; model assumptions and lack of data which limits their acceptability by decision makers and consumers.

One of the issues for simulated models is that they usually get information from different sources. For example, a cost-effectiveness model of antidiabetic medications obtained information from the literature and expert panels to determine algorithms of treatment; success, failures and adverse events were obtained from product labeling, literature and the drugs NDA; resource utilization data (i.e. physician office visits, laboratory tests, eye exams, etc.) were acquired from the American Diabetes Association guidelines, and costs were obtained from the literature.[46] This mixture of heterogeneous information raises questions related to the validity of the model. As a potential solution, some manufacturers have started including pharmacoeconomic evaluations alongside clinical trials. This ‘solution’ might appear logical

but the approach has limitations, such as the difficulty in merging clinical and economic outcomes in one study, limitations regarding the length of the trial as these may differ for the clinical vs. the economic measures, differing sample size considerations and finally differences in efficacy vs. effectiveness.

Frequently, trials are organized to show the efficacy of new medications but most phase II (and for that matter phase III) trials use surrogate measures as their primary end points and the long-term efficacy of the drug is unknown. For example, glycosylated hemoglobin ($HbA_{1c}$) or fasting plasma glucose is frequently used as an indicator of drug efficacy for phase II or phase III trials. However, when those efficacy data are used for simulation models, there is a lack of long-term efficacy information which then requires a series of controversial assumptions. To overcome that latter issue, economists are focusing on short term models adducing that health maintenance organizations (HMOs) are more interested in those outcomes while society is interested in both short and long-term outcomes. For example, a decision-tree model was developed to assess the direct medical costs and effectiveness of achieving glycosylated hemoglobin ($HBA_{1c}$) values with antidiabetic medications during the first 3 years of treatment. The authors justified the short-term period arguing that it was more relevant for decision makers to make guideline and formulary decisions.[46] Although it may look easy to switch short-term for long-term outcomes this switch may be problematic, because short term outcomes may not reflect long term outcomes. Another factor to consider is the length of a trial because if there is a considerable lag-time between premarketing trials and post-marketing trials, practice patterns may have changed affecting HMO decisions.

The size of the trial is also a very important factor to take into account in pharmacoeconomics because trials are powered for clinical outcomes and not for economic outcomes. If economic outcomes are used to power a trial, then a larger sample size will be required because economic outcomes have higher variation than clinical outcomes.[47]

In addition, the use of surrogate outcomes may not be economically relevant, a factor that needs to be considered by health economists and trialists during a trials planning phase. A question could then arise: could costs be used as endpoints? The short-answer is no, because costs data are not sensitive surrogates endpoints since cost and clinical outcomes may be disparate.

Finally, the efficiency of a new product requires that the manufacturer demonstrate the effectiveness of the product that is how the product behaves in the real world and not under ‘experimental’ conditions (efficacy). For example, the manufacturers may want to show that the new product is more cost-effective than current therapies or at least as good as new alternatives, but they need real-life data which are almost always absent when that product is launched. This is an important issue because premarketing trials are usually carried out in selective sites that are not representative of the practice community at large. Why are ‘real’ data so important? It is known that once a product is in the market, there is a wide variation in how the product is used by providers (e.g. indications, target population – different age, gender, socioeconomic status, patients with co-morbidities or multiple medications; adherence to medical guidelines, and variation among providers), or used by

patients (e.g. patient adherence to medications, variation in the disease knowledge, access to care, and type of care). Additionally, the new product might prompt changes in the resource utilization for a particular disease. For example, when repaglinide was introduced into the market, it was recommended that in patients with type 2 diabetes postprandial and fasting glucose, as well as $HbA_{1c}$ be monitored,[48,49] this type of monitoring would require testing that is additional to the usual management of patients with diabetes.

Because of the aforementioned issues, economic data alongside (or 'merged with') clinical trials are important because data obtained in premarketing trials could shed light on the goal of anticipating results in postmarketing trials, they could contribute to developing cost weights for future studies, and they could help to identify the resources that have the highest impact of the new drug.

## Discussion

The term 'phase IV study' has become misunderstood and has taken on negative connotations that have led some experts to question the validity of such trials. This latter point is emphasized by Pocock – 'such a trial has virtually no scientific merit and is used as a vehicle to get the drug started in routine medical practice.' He was undoubtedly referring to phase IV physician experience studies at the time. But even the phase IV PES has some merit, even given that adverse event reporting is voluntary, and underreporting of events is believed to be common (this is contrast to phase III trials where UAEs are arguably over reported). It is true that many phase IV studies have limitations in their research design, that the follow-up of patients enrolled in phase IV trials may be less vigorous than in controlled clinical trials (which can decrease the quantity and quality of information about the safety and efficacy of the medication being evaluated)[50,51] but, due to the highly varied designs of phase IV studies the utility of the information they provide will vary substantially from one study to another.

Due to the limitations of the current system for identifying adverse events, Strom has suggested a paradigm shift from the current traditional model of drug development and approval. He supports this paradigm shift based upon the fact that '…51% of drugs have label changes because of safety issues discovered after marketing, 20% of drugs get a new black box warning after marketing, and 3–4% of drugs are ultimately withdrawn for safety reasons.' The FDA website lists 12 drugs withdrawn from the market between 1997 and 2001 as shown in Table 5.5.

Strom's suggested paradigm for studying drug safety has a shortened phase III program followed by conditional approval during which time, required postmarketing studies would need to be performed (and the FDA would need to be given the power to regulate this phase in the same manner that they now have with phase I–III studies). He further recommends that once the conditional approval phase has ascertained safety in an additional 30,000 or more patients, the current system of optional and/or unregulated studies could be performed (Fig. 5.2).

**Table 5.5** Drugs withdrawn from the market between 1997 and 2001

| Drug name | Use | Adverse risk | Year approved |
|---|---|---|---|
| Cerivastatin | LDL reduction | Rhabdomyolysis | 1997 |
| Rapacuronium bromide | Anesthesia | Bronchospasm | 1999 |
| Alosetron | Irritable bowel | Ischemic colitis | 2000 |
| Cisapride | Heartburn | Arrhythmia | 1993 |
| PPA[1] | Decongestant | Stroke | * |
| Troglitazone | Type 2 diabetes | Liver toxicity | 1997 |
| Astemizole | Antihistamine | Arrhythmia | 1988 |
| Grepafloxacin | Antibiotic | Arrhythmia | 1997 |
| Mibefradil | High BP & angina | Arrhythmia | 1997 |
| Bromfenac | Pain relief | Liver toxicity | 1997 |
| Terfenadin | Antihistamine | Arrhythmia | 1985 |
| Fenfluramine | Appetite suppressant | Valve disease | 1973 |
| Dexfenfluramine | Appetite suppressant | Valve disease | 1996 |

*PPA (phenylpropanolamine) was in use prior to 1962, when an amendment to food and drug laws required a review of the effectiveness of this and other drugs while they remained on the market. It was deferred from final approval because of safety concerns about a possible association between phenylpropanolamine use and an increased risk of stroke. Based on previous case reports of stroke and data from a recent safety study, the FDA is proposing to remove phenylpropanolamine from the market.[1]

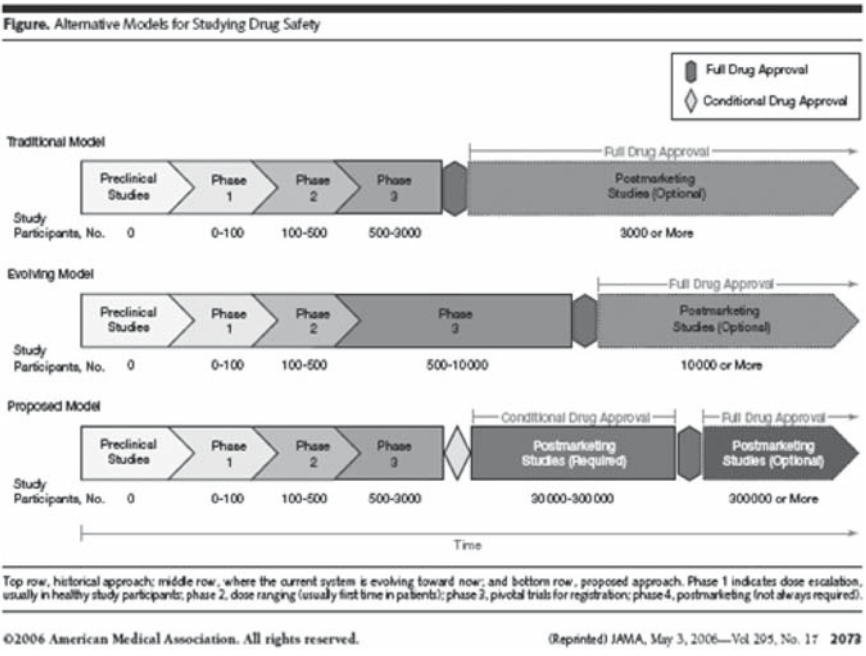



**Fig. 5.2** The current vs. some proposed paradigms for drug development

The conditional approval concept has been supported by the Institute of Medicine, and it goes further. The Institute of Medicine proposes to include a symbol for new drugs, new combinations of active substances, and new systems of delivery of existing drugs in the product label. This symbol would last 2 years and it would indicate the conditional approval of a drug until enough information of postmarketing surveillance is available, and during this period, the manufacturer would limit the use of direct-to-consumer advertising.[52] The question is how much impact that label would have on prescriber's since some studies have shown that prescriber's often fail to follow black box warnings labels[53]. The Institute of Medicine also recommends that FDA should reevaluate cumulative data on safety and efficacy no later than 5 years after approval. However, these changes are expected to have low impact if they are not accompanied by changes in the law commitments.

It is also important not to lump the phase IV study with other postmarketing research, research that may be every bit as scientifically rigorous as that associated with RCTs. Postmarketing studies are essential to establish patterns of physician prescribing and patient drug utilization and they are usually carried out using observational designs. Investigators frequently relate postmarketing surveillance studies with pharmacovigilance studies, and this might be a signal of what is happening in practice. In the last 25 years, 10% of the new drugs marketed in the United States have been withdrawn or were the subject of major warnings about serious or life-threatening side effects during the postmarketing phase. This situation has called for concrete actions such as closer monitoring of new drugs, the development of better notification systems for adverse events and presentation of transparent and high quality data.

Clinical pharmacologists and pharmacoepidemiologists are trying to promote the collection of blood samples at the population level for pharmacokinetic analysis. A study in psychiatric inpatients treated with alprazolam collected two blood samples at different time intervals to assess the pharmacokinetic variability of heterogeneous patient population.[54] This information could contribute to establishing dosages and frequency of drug administration in patients with co-morbidities, those treated with multiple medications and special populations. Clearly, the rubric of the phase IV study has taken on an expanded and meaningful role in drug development, use, and safety.

## Appendix

The following definitions were used in this manuscript

Definitions of phase IV trials:

- Post-marketing studies to delineate additional information including the drug's risks, benefits, and optimal use.
- clinicaltrials.mayo.edu/glossary.cfm.

- Postmarketing studies, carried out after licensure of the drug. Generally, a phase IV trial is a randomized, controlled trial that is designed to evaluate the long-term safety and efficacy of a drug for a given indication. Phase IV trials are important in evaluating AIDS drugs because many drugs for HIV infection have been given accelerated approval with small amounts of clinical data about the drugs' effectiveness.
- www.amfar.org/cgi-bin/iowa/bridge.html.
- In medicine, a clinical trial (synonyms: clinical studies, research protocols, medical research) is a research study.
- en.wikipedia.org/wiki/Phase_IV_trials.

1. Adverse drug event or adverse drug experience: 'an untoward outcome that occurs during or following clinical use of a drug, whether preventable or not' (does not mention causality)
2. Adverse experience: 'any adverse event associated with the use of a drug or biological product in humans, whether or not considered product related' (causality not assumed)
3. Adverse drug reaction: 'an adverse drug event that is judged to be caused by the drug' (specifically refers to causality)
4. 'Studies of adverse effects examine case reports of adverse drug reactions, attempting to judge subjectively whether the adverse events were indeed caused by the antecedent drug exposure' (specifically focuses on causality)
5. 'Studies of adverse events explore any medical events experienced by patients and use epidemiologic methods to investigate whether any given event occurs more often in those who receive a drug than in those who do not receive the drug' (a bit equivocal about causality: positive association v. causal association)

'Pharmacovigilance is a type of continual monitoring for unwanted effects and other safety-related aspects of drugs that are already on the market. In practice, pharmacovigilance refers almost exclusively to the spontaneous reporting systems which allow health care professionals and others to report adverse drug reactions to a central agency. The central agency can then combine reports from many sources to produce a more informative safety profile for the drug product than could be done based on one or a few reports from one or a few health care professionals.'

## References

1. Hartzema A. *Pharmacoepidemiology*. Vol 41. 3rd ed. Cincinnati, OH: Harvey Whitney Books Company; 1998.
2. Gough S. Post-marketing surveillance: a UK/European perspective. *Curr Med Res Opin*. Apr 2005; 21(4):565–570.
3. Olsson J, Terris D, Elg M, Lundberg J, Lindblad S. The one-person randomized controlled trial. *Qual Manag Health Care*. Oct–Dec 2005; 14(4):206–216.
4. Bugeja G, Kumar A, Banerjee AK. Exclusion of elderly people from clinical research: a descriptive study of published reports. *BMJ*. Oct 25, 1997; 315(7115):1059.

5. Corrigan OP. A risky business: the detection of adverse drug reactions in clinical trials and post-marketing exercises. *Soc Sci Med*. Aug 2002; 55(3):497–507.
6. Gurwitz JH, Col NF, Avorn J. The exclusion of the elderly and women from clinical trials in acute myocardial infarction. *JAMA*. Sept 16, 1992; 268(11):1417–1422.
7. Simon SD. Is the randomized clinical trial the gold standard of research? *J Androl*. Nov–Dec 2001; 22(6):938–943.
8. Glasser SP, Salas M, Delzell E. Importance and challenges of studying marketed drugs: what is a phase IV study? Common clinical research designs, registries, and self-reporting systems. *J Clin Pharmacol*. Sept 2007; 47(9):1074–1086.
9. Farahani P, Levine M, Gaebel K, Thabane L. Clinical data gap between phase III clinical trials (pre-marketing) and phase IV (post-marketing) studies: evaluation of etanercept in rheumatoid arthritis. *Can J Clin Pharmacol*. Fall 2005; 12(3):e254–263.
10. Gex-Fabry M, Balant-Gorgia AE, Balant LP. Therapeutic drug monitoring databases for postmarketing surveillance of drug-drug interactions. *Drug Saf*. 2001; 24(13):947–959.
11. Kaufman DW, Kelly JP, Rosenberg L, Anderson TE, Mitchell AA. Recent patterns of medication use in the ambulatory adult population of the United States: the Slone survey. *Jama*. Jan 16, 2002; 287(3):337–344.
12. Vijan S, Kent DM, Hayward RA. Are randomized controlled trials sufficient evidence to guide clinical practice in type II (non-insulin-dependent) diabetes mellitus? *Diabetologia*. Jan 2000; 43(1):125–130.
13. Castle WM, Lewis JA. Postmarketing surveillance of adverse drug reactions. *Br Med J (Clin Res Ed)*. May 12, 1984; 288(6428):1458–1459.
14. Hayward RA, Kent DM, Vijan S, Hofer TP. Reporting clinical trial results to inform providers, payers, and consumers. *Health Aff (Millwood)*. Nov–Dec 2005; 24(6):1571–1581.
15. Edwards C, Blowers DA, Pover GM. Fosinopril national survey: a post-marketing surveillance study of fosinopril (Staril) in general practice in the UK. *Int J Clin Pract*. Sept 1997; 51(6):394–398.
16. Fallowfield JM, Blenkinsopp J, Raza A, Fowkes AG, Higgins TJ, Bridgman KM. Post-marketing surveillance of lisinopril in general practice in the UK. *Br J Clin Pract*. Nov–Dec 1993; 47(6):296–304.
17. Marsh BT, Atkins MJ, Talbot DJ, Fairey IT. A post-marketing acceptability study in 11,685 patients of the efficacy of timolol/bendrofluazide in the management of hypertension in general practice. *J Int Med Res*. Mar–Apr 1987; 15(2):106–114.
18. Riley J, Wilton LV, Shakir SA. A post-marketing observational study to assess the safety of mibefradil in the community in England. *Int J Clin Pharmacol Ther*. June 2002; 40(6):241–248.
19. Schmidt J, Kraul H. Clinical experience with spirapril in human hypertension. *J Cardiovasc Pharmacol*. Aug 1999; 34 Suppl 1:S25–30.
20. Ueng KC, Chen ZC, Yeh PS, et al. Nifedipine OROS in Chinese patients with hypertension–results of a post-marketing surveillance study in Taiwan. *Blood Press Suppl*. July 2005; 1:32–38.
21. Tognoni G, Alli C, Avanzini F, et al. Randomised clinical trials in general practice: lessons from a failure. *BMJ*. Oct 19, 1991; 303(6808):969–971.
22. Ben-Menachem E. Data from regulatory studies: what do they tell? What don't they tell? *Acta Neurol Scand Suppl*. 2005; 181:21–25.
23. Lesko SM, Mitchell AA. The safety of acetaminophen and ibuprofen among children younger than two years old. *Pediatrics*. Oct 1999; 104(4):e39.
24. Jackson L, Ting A, McKay S, Galea P, Skeoch C. A randomised controlled trial of morphine versus phenobarbitone for neonatal abstinence syndrome. *Arch Dis Child Fetal Neonatal Ed*. July 2004; 89(4):F300–304.
25. Fischer G, Ortner R, Rohrmeister K, et al. Methadone versus buprenorphine in pregnant addicts: a double-blind, double-dummy comparison study. *Addiction*. Feb 2006; 101(2):275–281.

26. Vocci F, Ling W. Medications development: successes and challenges. *Pharmacol Ther*. Oct 2005; 108(1):94–108.
27. Jones HE, Suess P, Jasinski DR, Johnson RE. Transferring methadone-stabilized pregnant patients to buprenorphine using an immediate release morphine transition: an open-label exploratory study. *Am J Addict*. Jan–Feb 2006; 15(1):61–70.
28. Lurie P. FDA Report Highlights Poor Enforcement of Post-Marketing Follow-up. http://www.citizen.org/pressroom/release.cfm?ID=2147. Accessed October 12, 2006.
29. Yusuf S, Mehta SR, Xie C, et al. Effects of reviparin, a low-molecular-weight heparin, on mortality, reinfarction, and strokes in patients with acute myocardial infarction presenting with ST-segment elevation. *JAMA*. Jan 26, 2005; 293(4):427–435.
30. Smith DH, Neutel JM, Lacourciere Y, Kempthorne-Rawson J. Prospective, randomized, open-label, blinded-endpoint (PROBE) designed trials yield the same results as double-blind, placebo-controlled trials with respect to ABPM measurements. *J Hypertens*. July 2003; 21(7):1291–1298.
31. Cross J, Lee H, Westelinck A, Nelson J, Grudzinskas C, Peck C. Postmarketing drug dosage changes of 499 FDA-approved new molecular entities, 1980–1999. *Pharmacoepidemiol Drug Saf*. Sept 2002; 11(6):439–446.
32. *FDA news Drug Daily Bulletin*. Oct 2006; 3(207).
33. Bombardier C, Laine L, Reicin A, et al. Comparison of upper gastrointestinal toxicity of rofecoxib and naproxen in patients with rheumatoid arthritis. VIGOR Study Group. *N Engl J Med*. Nov 23, 2000; 343(21):1520–1528, 1522 p following 1528.
34. Mukherjee D, Nissen SE, Topol EJ. Risk of cardiovascular events associated with selective COX-2 inhibitors. *JAMA*. Aug 22–29, 2001; 286(8):954–959.
35. Juni P, Nartey L, Reichenbach S, Sterchi R, Dieppe PA, Egger M. Risk of cardiovascular events and rofecoxib: cumulative meta-analysis. *Lancet*. Dec 4–10, 2004; 364(9450):2021–2029.
36. Bresalier RS, Sandler RS, Quan H, et al. Cardiovascular events associated with rofecoxib in a colorectal adenoma chemoprevention trial. *N Engl J Med*. Mar 17, 2005; 352(11): 1092–1102.
37. Merck & Co I. http://www.vioxx.com/vioxx/documents/english/vioxx_press_release.pdf. Accessed October 4.
38. Abenhaim L. Lessons from the withdrawal of rofecoxib: France has policy for overall assessment of public health impact of new drugs. *BMJ*. Dec 4, 2004; 329(7478):1342.
39. FDA. MedWatch: Voluntary Reporting by Health Professionals. http://www.fda.gov/medwatch/report/hcp.htm. Accessed October 12, 2006.
40. Grootheest van A, Graafe e L, Jong van den Berg de L. Consumer reporting: a new step in pharmacovigilance? An overview. *Drug Safety*. 2003; 26:211–217.
41. Improving ADR reporting. *Lancet*. Nov 9, 2002; 360(9344):1435.
42. Zwillich T. How Vioxx is changing US drug regulation. *Lancet*. Nov 19, 2005; 366(9499):1763–1764.
43. Taylor R, Bethell H, Brodie D. Clinical trial versus the real world: the example of cardiac rehabilitation. *Br J Cardiol*. 2007; 14:175–178.
44. Tappenden P, Chilcott J, Ward S, Eggington S, Hind D, Hummel S. Methodological issues in the economic analysis of cancer treatments. *Eur J Cancer*. Nov 2006; 42(17):2867–2875.
45. Chilcott J, Brennan A, Booth A, Karnon J, Tappenden P. The role of modelling in prioritising and planning clinical trials. *Health Technol Assess*. 2003; 7(23):iii, 1–125.
46. Ramsdell JW, Braunstein SN, Stephens JM, Bell CF, Botteman MF, Devine ST. Economic model of first-line drug strategies to achieve recommended glycaemic control in newly diagnosed type 2 diabetes mellitus. *Pharmacoeconomics*. 2003; 21(11):819–837.
47. Briggs A, Gray A. The distribution of health care costs and their statistical analysis for economic evaluation. *J Health Serv Res Policy*. Oct 1998; 3(4):233–245.
48. Leiter LA, Ceriello A, Davidson JA, et al. Postprandial glucose regulation: New data andnew implications. *Clin Ther*. 2005; 27 Suppl 2:S42–56.

49. Plosker GL, Figgitt DP. Repaglinide: a pharmacoeconomic review of its use in type 2 diabetes mellitus. *Pharmacoeconomics*. 2004; 22(6):389–411.
50. Heeley E, Riley J, Layton D, Wilton LV, Shakir SA. Prescription-event monitoring and reporting of adverse drug reactions. *Lancet*. Dec 1, 2001; 358(9296):1872–1873.
51. Lassila R, Rothschild C, De Moerloose P, Richards M, Perez R, Gajek H. Recommendations for postmarketing surveillance studies in haemophilia and other bleeding disorders. *Haemophilia*. July 2005; 11(4):353–359.
52. Institute of Medicine of the National Academies. The Future of Drug Safety. http://www.nap.edu/books/0303103045/html/1.html. Accessed April 3, 2007.
53. Public Health Newswire. Drug's Black Box Warning Violations in Outpatient Settings Putting Patients at Risk. http://www.medicalnewstoday.com/medicalnews.php?newsid=37735. Accessed April 3, 2007.
54. DeVane CL, Grasela TH, Jr., Antal EJ, Miller RL. Evaluation of population pharmacokinetics in therapeutic trials. IV. Application to postmarketing surveillance. *Clin Pharmacol Ther*. May 1993; 53(5):521–528.

# Chapter 6
# The United States Federal Drug Administration (FDA) and Clinical Research

**Stephen P. Glasser, Carol M. Ashton, and Nelda P. Wray**

*Some bargains are Faustian, and some horses are Trojan. Dance carefully with the porcupine, and know in advance the price of intimacy*[1]

**Abstract** The USFDA is an agency of the US Department of Health and Human Services and is the nation's oldest consumer protection agency whose function it is to review drugs before marketing, monitor marketed drugs, monitor drug manufacturing and advertising, protect drug quality, and to conduct applied research. It is charged with overseeing of not only human drugs and biologics, but also veterinary drugs, foods, medical devices, and radiopharmaceuticals, and as such serves as a watch dog over industry. This chapter discusses the historical development of the FDA, and what the FDA is today. The phases of research development leading to the marketing of a new drug, the role of the FDA in surgical interventions, and the FDA's role in advertising and adverse event reporting are discussed.

## Historical Considerations

The history leading up to the formation of the FDA is interesting. In the early days of our country, epidemics were common (diphtheria, typhoid, yellow fever, small pox etc.), there were few if any specific treatments, the few patent medicines were largely unregulated, some were dangerous, and few were effective. In fact, drugs readily available in the 1800s would likely astound the average citizen today. For example Winslow's Soothing Syrup and Koop's Babyfriend contained liberal amounts of morphine, a marketed cough syrup contained heroin, and so on. Beginning as the Division of Chemistry and then (after July 1901) the Bureau of Chemistry, the modern era of the FDA dates to 1906 with the passage of the Federal Food and Drugs Act; this added regulatory functions to the agency's scientific mission and was the result of recurrent food safety scares (Fig. 6.1). The Division of Chemistry investigation into the adulteration of agricultural commodities was actually initiated as early as 1867. When Harvey Washington Wiley arrived as chief

S.P. Glasser (ed.), *Essentials of Clinical Research*,



**Fig. 6.1** Depictions of historical developments in drug safety

chemist in 1883, the government's handling of the adulteration and misbranding of food and drugs took a decidedly different course, which eventually helped spur public indignation at the problem. Wiley expanded the division's research in this area, exemplified by Foods and Food Adulterants, a ten-part study published from 1887 to 1902. He demonstrated his concern about chemical preservatives as adulterants in the highly publicized "poison squad" experiments, in which able-bodied volunteers consumed varying amounts of questionable food additives to determine their impact on health. And Wiley unified a variety of groups behind a federal law to prohibit the adulteration and misbranding of food and drugs, including state chemists and food and drug inspectors, the General Federation of Women's Clubs, and national associations of physicians and pharmacists.[2]

Languishing in Congress for five years, the bill that would replace the 1906 Act was ultimately enhanced and passed in the wake of a therapeutic disaster in 1937. In September and October of 1937, people across the country started dying after drinking a new cough medicine known as Elixer Sulfanilamide. Produced by the S.E. Massengill Company in response to consumer demand for a liquid cough medicine, it was released to the public after testing for flavor, appearance and fragrance – but not toxicity. At the time, federal regulations did not require companies to certify that their drugs were safe, and the solution used to liquefy the sulfanilamide was diethylene glycol, a deadly poison that is found in anti-freeze. From the first death to the FDA's no-holds-barred response, John Swann, in a DVD entitled the ELIXER OF DEATH, tells the remarkable story of the incident that led to passage of the 1938 Food, Drug, and Cosmetic Act, which increased the FDA's authority to regulate drugs. Survivor's, recall their harrowing ordeals, and FDA historians reveal how agents located 234 of the 240 gallons produced – often one bottle at a time![3]

The public outcry not only reshaped the drug provisions of the new law to prevent such an event from happening again, it propelled the bill itself through Congress. FDR signed the Food, Drug, and Cosmetic Act on 25 June 1938. The new law brought cosmetics and medical devices under control, and it required that drugs be labeled with adequate directions for safe use. Moreover, it mandated premarket approval of all new drugs, such that a manufacturer would have to prove to

FDA that a drug were safe before it could be sold. It irrefutably prohibited false therapeutic claims for drugs, although a separate law granted the Federal Trade Commission jurisdiction over drug advertising. The act also corrected abuses in food packaging and quality, and it mandated legally enforceable food standards. Tolerances for certain poisonous substances were addressed. The law formally authorized factory inspections, and it added injunctions to the enforcement tools at the agency's disposal.

The Bureau of Chemistry's name changed to the Food, Drug, and Insecticide Administration in July 1927, when the non-regulatory research functions of the bureau were transferred elsewhere in the department. In July 1930 the name was shortened to the present version. FDA remained under the Department of Agriculture until June 1940, when the agency was moved to the new Federal Security Agency. In April 1953 the agency again was transferred, to the Department of Health, Education, and Welfare (HEW). Fifteen years later FDA became part of the Public Health Service within HEW, and in May 1980 the education function was removed from HEW to create the Department of Health and Human Services, FDA's current home. To understand the development of this agency is to understand the laws it regulates, how the FDA has administered these laws, how the courts have interpreted the legislation, and how major events have driven all three.[4]

During the time period from 1906 and 1938, there were the beginnings of pharmaceutical research and drug discovery. For example, penicillin was discovered in 1928 and insulin was also uncovered during this period. In 1951 the Durham-Humphrey Amendment defined the Over the Counter drugs. The 1940–1950s was also the golden age for pharmaceutical companies with over 90% of all drugs used in 1964 were unknown before 1938. In 1960 the Kefauver Hearings were evaluating drug costs, prices and profits but the 1962 thalidomide tragedy resulted in the 1962 Kefauver-Harris Drug Amendments. The original impetus for the effectiveness requirement was Congress's growing concern about the misleading and unsupported claims made by pharmaceutical companies about their drug products coupled with high drug prices.[5] In this 1962 Act, Congress amended the Federal Food, Drug, and Cosmetics Act to add the requirement that to obtain marketing approval, manufacturers demonstrate the effectiveness (with substantial evidence) of their products through the conduct of adequate and well-controlled studies (prior to this amendment there were only safety requirements). This amendment also established informed consent procedures, the reporting process for adverse drug events and placed drug advertising under FDA jurisdiction. Another important milestone for the FDA came in 1968 when the Drug Efficacy Study Implementation (DESI) was enacted. This resulted in over 4,000 drugs marketed between 1938 and 1962 to undergo evaluation for efficacy and safety based upon the existent literature (pre 1938 drugs were "grandfathered"). Other significant actions followed, including the Medical Device Amendment of 1978 which put medical devices under the same kinds of Good Medical Practice (GMP) and Good Clinical Practice (GCP) guidelines that applied to drug development. GCP is an international ethical and scientific standard for designing, conducting, recording, and reporting trials that involve

the participation of human subjects. The GCP principles have their origin in the Declarations of Helsinki.

## The FDA Now

The U.S. Food and Drug Administration is a scientific, regulatory, and public health agency that oversees items accounting for 25 cents of every dollar spent by consumers. Its jurisdiction encompasses most food products (other than meat and poultry), human and animal drugs, therapeutic agents of biological origin, medical devices, radiation-emitting products for consumer, medical, and occupational use, cosmetics, and animal feed. As prior mentioned the agency grew from a single chemist in the U.S. Department of Agriculture in 1862 to a staff of approximately 9,100 employees and a budget of $1.294 billion in 2001, comprising chemists, pharmacologists, physicians, microbiologists, veterinarians, pharmacists, lawyers, and many others. About one-third of the agency's employees are stationed outside of the Washington, D.C. area, staffing over 150 field offices and laboratories, including five regional offices and 20 district offices. Agency scientists evaluate applications for new human drugs and biologics, complex medical devices, food and color additives, infant formulas, and animal drugs. Also, the FDA monitors the manufacture, import, transport, storage, and sale of about $1 trillion worth of products annually at a cost to taxpayers of about $3 per person. Investigators and inspectors visit more than 16,000 facilities a year, and arrange with state governments to help increase the number of facilities checked.

An era of rapid change for the FDA also occurred with an increase in drug development and approval beginning in the early 1970s. During the period of 1970–2002, reports on the adverse events of over 6,000 marketed drugs numbered in the millions, with 75 drugs removed from the market and another 11 that had severe restrictions placed on their use. From 1975 to 1999, 584 new chemical entities were approved, and over 10% of these either were withdrawn or received a "black-box' warning. This rapid increase in marketed drugs placed a tremendous burden on the post-marketing safety systems which the FDA had in place to protect public safety. More recently, a number of dug 'embarrassments' occurred that has again reshaped the FDA. These embarrassments included concealed studies (studies that the manufacturer did not publish), fraudulent data (as exemplified in the development of telithromycin), rofecoxib (the withdrawal of Vioxx still has pending litigation), and rosiglitazone withdrawal. After rofecoxib was withdrawn from the market, the Center for Drug Evaluation and Research (CDER) asked the Institute of Medicine (IOM) to assess the US drug-safety system. Their report was released in 2006.

As to telithromycin, French pharmaceutical company Hoechst Marion Roussel (later Sanofi-Aventis) started phase II/III trials of telithromycin (HMR-3647) in 1998. Telithromycin was approved by the European Commission in July 2001

and subsequently came on sale in October 2001. In USA, telithromycin gained FDA approval April 1, 2004. FDA staffers publicly complained that safety problems were ignored, and congressional hearings were held to examine those complaints. Some of the data in clinical trials submitted to the FDA turned out to be fabricated, and one doctor went to prison. The House Committee on Energy and Commerce held hearings. Study 3014 was a key clinical trial of more than 24,000 patients that Sanofi-Aventis submitted to the FDA seeking approval for Ketek. An indictment said that one doctor fabricated data she sent to the company. Documents, including internal Sanofi-Aventis emails, show that Aventis was worried about this doctor early in study 3014 but didn't tell the FDA until the agency's own inspectors discovered the problem independently.[6]

Due to the rapid increase in new drug development during the 1970s and on, the Prescription Drug User Fee Act (PDUFA), was enacted in 1992 and was revised in 1997 and 2002. PDUFA is a program under which the pharmaceutical/biotechnology industry pays certain "user fees" to the Food and Drug Administration (FDA). In exchange for these fees, the FDA agreed, via correspondence with Congress, to a set of performance standards intended to reduce the approval time for New Drug Applications (NDA) and Biological License Applications (BLA). PDUFA assess three types of user fees: (1) fees on applications (NDA/BLA); (2) annual fees on establishments; and (3) renewal fees on products. The law includes a set of "triggers" designed to ensure that appropriations for application review are not supplanted by user fees. These triggers require that Congressional appropriations for such review reach certain levels before user fees may be assessed, and that the FDA devotes a certain amount of appropriated funds annually to drug review activities. However, little provision was made for post marketing drug surveillance. PDUFA resulted in a reduction in review times from 33 to 14 months. Also, prior to PDUFA, the testing ground for new drugs occurred predominantly in Europe. In 1980, only 2% of new drugs were first being first used in the USA; by 1988 60% were first used in the USA. The glut of new approved and arguably understudied drugs on the US market, placed a stress on the already inadequate post marketing surveillance systems, and ultimately lead to the commission of an Institute of Medicine review. This IOM review lead to the FDA Amendments Act of 2007.[7] This 156 page document expands the authority of the FDA particularly as it relates to marketed drugs (see Chapter 5). Briefly, this new act grants the FDA the power to require postmarketing studies, to order changes in a drug's label, and to restrict distribution of a drug. The Act also provides new resources (225 million dollars over five years aimed at improving drug safety).

## International Conference on Harmonization (ICH)

More recently, an international effort was initiated that was designed to bring together the regulatory authorities of Europe, Japan and the United States and experts from the pharmaceutical industry in the three regions to discuss scientific and technical aspects of product registration. Their stated purpose is to

make recommendations on ways to achieve greater harmonization in the interpretation and application of technical guidelines and requirements for product registration in order to reduce or obviate the need to duplicate the testing carried out during the research and development of new medicines. The objective of such harmonization is a more economical use of human, animal and material resources, and the elimination of unnecessary delay in the global development and availability of new medicines whilst maintaining safeguards on quality, safety and efficacy, and regulatory obligations to protect public health. The Mission Statement of the ICH (as taken from their website)[8] is "to maintain a forum for a constructive dialogue between regulatory authorities and the pharmaceutical industry on the real and perceived differences in the technical requirements for product registration in the EU, USA and Japan in order to ensure a more timely introduction of new medicinal products, and their availability to patients; to contribute to the protection of public health from an international perspective; To monitor and update harmonised technical requirements leading to a greater mutual acceptance of research and development data; To avoid divergent future requirements through harmonisation of selected topics needed as a result of therapeutic advances and the development of new technologies for the production of medicinal products; To avoid divergent future requirements through harmonisation of selected topics needed as a result of therapeutic advances and the development of new technologies for the production of medicinal products; To facilitate the adoption of new or improved technical research and development approaches which update or replace current practices, where these permit a more economical use of human, animal and material resources, without compromising safety; and, to facilitate the dissemination and communication of information on harmonised guidelines and their use such as to encourage the implementation and integration of common standards."

## USA Drug Development Phases

Since the FDAs 1962 amendment mentioned above, the issue of what constitutes sufficient evidence of effectiveness has been debated by the FDA, Industry, and academia. Before getting to that point, there is a fairly regimented program for drug development which will be discussed in the following paragraphs.

### *Preclinical Evaluation*

First, when a new drug is identified as possibly active, it undergoes chemical and physical characterization and screening for biological activity by testing in appropriate animal models. This includes toxicity studies followed by preclinical pharmacology

where dosage, mode of action, chronic toxicology, safety, efficacy, and teratogenicity are evaluated. If the drug seemingly has merit, it advances to clinical investigation where it undergoes three phases of evaluation (phase 1, 2 and 3 trials). However, before clinical testing can take place, an Investigational New Drug (IND) Application must be submitted and approved by the FDA. Since across state transfer of drugs is necessary, and there is a federal law against such transport, an IND allows for an exemption in the law so that a drug can be shipped via interstate commerce. This is a rapid process (the FDA must respond within 30 days of the application). Parenthetically, the FDA uses a broad definition for "new drug". It is not just a new chemical moiety; rather a new drug is any drug or drug use that is not included in current labeling of that drug. If the drug has no prior approval the definition is fairly obvious. However, an approved drug now being studied with a new release system (e.g. a transdermal patch, or a new salt side chain), a new indication, or a new combination (even if the two separate drugs are approved) is considered "new". So, for example, when aspirin was to be studied in the Coronary Drug Project, an IND had to be submitted for this "new" drug.[9]

## *Phase 1–3 Studies*

Following IND approval a phase 1 study can be launched and these are sometimes referred to as "first-in-man" studies. In general phase 1 trials have relatively small sample sizes and are usually performed in normal human volunteers. The goal is to evaluate pharmacokinetics (PK) and to determine if there are any differences compared to the preclinical studies. Early, phase 1 studies are acute PK evaluations; later the studies may include chronic PK and dose escalation in order to determine the maximum tolerated dose. First in man studies have received renewed interest as a result of the TGN-1412 study,[10] which in its first human clinical trials, caused catastrophic systemic failure in the subjects, despite being administered at a supposed sub-clinical dose. The adverse event resulted in the hospitalization of six volunteers. At least four of the six these suffered multiple organ dysfunction, and one trial volunteer is said to be showing signs of developing cancer. Tentative opinions from an as yet uncompleted inquiry suggest that the problems arose due to an "unforeseen biological action in humans", rather than any breach of trial protocols; and, the case, therefore, has had important ramifications for future trials of potentially powerful clinical agents. In part, as a result of this aforementioned trial, the European Medicines Agency (EMEA the European equivalent of the USFDA) is in the process of developing guidelines for first in man studies.[11] This initial draft guidance has been the subject of wide comment in the clinical trials community, and as a result of a wide variety of opinions has a challenge to finalize the guidelines.[12]

Phase 2 trials are slightly larger and also examine PKs, but now in patients with the disease of interest. In addition, these are referred to as "proof of concept" studies. They are also dose-ranging and safety studies. Recently, there has been a suggestion that phase 2 trials be sub-classified into 2A and 2B. Phase 2B studies can

be thought of as smaller early RCTs, while phase 2A studies are an extension of phase 1 studies, but in patients rather than subjects. These classifications are not firm, however, and there are many exceptions. Phase 2 studies are also feasibility studies, in which efficacy, response rates, and response durations are determined. This is also a phase in which ineffective treatment can be rejected prior to the more expensive phase 3 trials (there are an increasing number of phase 3 efficacy trials which fail to find benefit). In order to save money and time, phase 2 futility studies are becoming more common. In this variant of phase 2 trials, futility studies can be designed as a clever way of dealing with the trade-off between investment risk and clinical promise. That is, one way to reduce the studies sample size is to focus on futility-that is designing a study to identify which agents are least likely to demonstrate benefits rather than the more typical goal of identifying the most promising agents. The null hypothesis in a futility study is that the treatment has promise and will therefore produce results exceeding a meaningful threshold. If that threshold is not met, the null is rejected and further study is considered futile. Remember, the same provisos hold regarding the null discussed in Chapters 3 and 18. That is, agents passing an efficacy criterion are winners, but agents meeting the futility criterion are merely non-losers. For example, Palesch et al. evaluated six phase 2 futility study designs of therapeutic stroke trials. They identified three trials as futile in phase 2, and none of the three subsequently showed benefit in subsequent phase 3 trials. In the remaining three phase 2 trials which did not show futility, 1 showed efficacy in phase 3.[13]

More specifically, the way phase 2 futility studies are designed is first to estimate the proportion of favorable outcomes in untreated controls (this is usually done from historical case-series or control groups from previous trials) and this becomes the proportion of favorable outcomes for the single arm phase 2 futility study. The minimally worthwhile improvement of the drug under study is then estimated as one does in determining the sample size in phase 3 studies. If the null hypothesis is rejected that there is a minimally worthwhile improvement, we conclude that the benefit of treatment is less than what we would want, and it is therefore futile to proceed to a phase 3 trial. Additionally, in phase 2 futility trials, one would want to minimize the risk of drawing false negative conclusions (that is that the drug shows no efficacy when it in fact does-one would not want to miss studying a potentially effective agent). The sample size is then "hedged" towards this aforementioned goal, with less concern about a false positive conclusion (that is that the drug is effective when in fact it is not).[13]

Phase 3 trials are classical efficacy studies generally using RCT designs discussed in Chapter 3; and, phase 4 studies are discussed in Chapter 5. However, it is the phase 3 study that was the topic of discussion as a result of the 1962 Kefauver-Harris amendment. The main issue of contention about phase 3 studies surrounded the words "substantial evidence" of effectiveness that the FDA required for drug approval. In the above mentioned FDA Act, substantial evidence was defined as "evidence consisting of adequate and well-controlled investigations, including clinical investigations, by experts qualified by scientific training and experience to evaluate the effectiveness of the drug involved, on the basis of

which it could be fairly and responsible concluded by such experts that the drug will have the effect it purports or is represented to have under the conditions of use prescribed, recommended, or suggested in the labeling or proposed labeling thereof." The argument that ensued from this definition centered on what the specific quality of evidence was in order to establish efficacy. It was the FDA's position that Congress intended to require at least two adequate and well-controlled studies, each convincing on its own, to establish efficacy. There has been some subsequent flexibility by the FDA in regard to the above as it applies to a specific drug in development. In some cases, for example, the FDA has relied on information from adequate and well-controlled studies published in the literature. In other cases where it would be difficult to perform a second study due to ethical concerns, the result of a single study could be accepted (as long as it was of excellent design, provided highly reliable and statistically strong – $p < 0.001$ – evidence of important clinical benefit-such as survival).

The requirement of more than 1 adequate and well-controlled investigation reflects the need for independent substantiation of experimental results and refers back to the question posed in Chapter 3 that asked why studies can presumably be of similar design and yet lead to different results. Indeed, the FDA realized that any clinical trial may be subject to unanticipated, undetected, systematic biases that may be operative irrespective of the best intentions of sponsors and investigators. They also note that the inherent variability in biological systems may produce a positive trial by chance alone. In addition, results may be dependent on specific issues related to the site or the investigator (e.g. concomitant treatments, diets etc.) that may impact the generalizability of the results. Finally (and fortunately rarely), favorable efficacy might be the product of scientific fraud. Independent substantiation of experimental results then addresses these problems by providing consistency across more than one study, thus greatly reducing the possibility that a biased, chance, site-specific, or fraudulent result will lead to an erroneous conclusion that a drug is effective.

The concept of independent substantiation of trial results, has often been referred to as replication, but replication may imply precise repetition of the same experiment. Actually, studies that are of different design, in different populations, with different endpoints or dosage forms may provide evidence of efficacy, and this may be even more convincing than repetition of the same study. It should be noted, that it is usually not necessary to rely on a single study to support the efficacy of a drug under development. This is because, in most situations there is a need to explore the appropriate dose range, to study patients with differing complexities and severities of disease, to compare the drug to other therapies, to perform safety studies, so that before marketing, most drugs will have been evaluated in more than one study.

Another trend seen by the FDA is the increase in new drug applications from foreign studies. In 2000, 27% of NDA's contained pivotal data from foreign studies.[14] There is no current restriction on non-US studies being used to support an NDA so long as they are well designed and conducted and the study sites are available for inspection.

## FDA and Surgical Interventions

**Carol M. Ashton MD MPH and Nelda P. Wray MD MPH**

Whereas prescription drugs are regulated by the FDA, and for drug approval there is the requirement that there be pre-release demonstration of efficacy and safety in randomized trials, there are no FDA regulations governing surgical interventions. Rather, new surgical interventions are developed based on anatomic and clinicopathological correlations in humans and studies in animals, and then used in humans, with the initial experience reported as case reports or a series of cases. Subsequent large scale dissemination of the procedure occurs as additional surgical groups begin using it. It is only subsequently, when doubts set in about a given procedure, that its efficacy is evaluated in a randomized controlled trial. These RCTs generally demonstrate that the procedure is less beneficial or more harmful than originally thought, no better than a nonoperative course of action, beneficial for only certain subgroups, or no better than a placebo (sham procedure). A classic example of the above principles is the story of lung volume reduction surgery (LVRS) for emphysema.[15] The first report of the use of LVRS in humans was published in 1957[16] but the procedure did not become widely used until it was modified in the mid 1990s by Joel Cooper.[17] Dr. Cooper reported his experience with 20 cases in 1994 (abstract) and 1995 (paper). By 1996, 1,200 LVRS were performed in Medicare beneficiaries, at an estimated cost of $30,000–70,000 each, not counting physician charges. But here is where the LVRS story diverges from the typical scenario. Scrutiny of LVRS by a consensus of experts as well as Medicare officials led to concerns about the procedure's effectiveness and safety. In a landmark decision,[18] Medicare officials decided that coverage for LVRS would only be provided in the context of a clinical trial. This decision was challenged by Dr. Cooper and others championing the procedure as unethical because of the "obvious benefit of the procedure." In record time, the NIH, Health Care Financing Administration (now the Centers for Medicare and Medicaid Services) and the Agency for Healthcare Research and Quality launched a randomized trial of LVRS vs. medical therapy for severe emphysema, the National Emphysema Treatment Trial, enrolling the first patient in 1997. The initial results, reported in 2003,[19] indicated that, in 1,219 patients followed for an average of 29 months, in certain subgroups of patients, LVRS resulted in higher mortality rates than medical therapy. Based on the trial results, Medicare officials limited coverage to patient subgroups that appeared to benefit or at least not be harmed by LVRS. But the trial seems to have quenched demand for LVRS. By 2006, as reported in the New York Times, "Medicare says it will pay, but patients say 'no thanks,'" only 458 Medicare claims for LVRS were filed between January 2004 and September 2005.[20]

Two other examples of this "evolutionary pattern" in the development of surgical interventions are provided by carotid artery endarterectomy for stroke prevention; and, arthroscopic treatment for relief of knee pain due to osteoarthritis. The first case report of carotid artery endarterectomy in a human appeared in 1956.[21] By 1971, 15,000 carotid endarterectomies were performed in USA. By 1985, this had increased to 107,000.[22] Criteria were then developed for the appropriate use of this

procedure; when they were retrospectively applied to the carotid endarterectomies performed on Medicare beneficiaries in 1981, only 35% of patients were found to have undergone the procedure for "appropriate" reasons, and in another 32% the reasons were equivocal.[22] Definitive randomized trials of carotid endarterectomy were not conducted and reported until the mid 1990s.[23–25] The volume of carotid artery endarterectomies in the US increased from 68,000 in 1990 to 134,000 in 2002, but the trials changed clinical practice: based upon the appropriateness criteria, by 1999 only 8.6% could be deemed "inappropriate".[26] On the other hand, 75% of all carotid artery endarterectomies are now performed in asymptomatic patients, in whom the risk:benefit ratio of the procedure is much narrower. In 2004, the FDA approved for use the first carotid artery stent, and now carotid artery stenting is being compared with carotid artery endarterectomy in RCTs. This fact illustrates the fact the FDA's role *vis a vis* surgical procedures is limited to regulating the various devices that may be used in the course of performing them.

A final example of the evolution of new surgical approaches is that of arthroscopic lavage with or without debridement for knee pain due to osteoarthritis. Fiberoptic arthroscopic debridement for this condition began to be used in the mid-1970s. By 1996, more than 650,000 of these procedures were performed in US.[27] A definitive randomized trial of the efficacy of this procedure was not begun until 1995. That trial was a single site study in which 180 people were randomized in the operating room to arthroscopic lavage, arthroscopic lavage plus debridement, or a sham procedure (skin incisions with no entry into the joint) and followed for two years. The study showed that arthroscopic lavage with our without debridement was no better than the sham procedure in relieving pain and restoring function.[27] That same year, the Veterans Health Administration issued a directive that it would no longer cover arthroscopic surgery for the relief of pain due to osteoarthritis, and the Centers for Medicare and Medicaid shortly followed suit. Between 2000 and 2005, the volume of these procedures in VHA declined by 26%.

Clearly, there are challenges in designing an RCT to evaluate the efficacy of an invasive therapeutic procedure. Potential randomized designs that could be used to evaluate the efficacy of a procedure include comparing the operative procedure to a non-operative course of therapy, the operative procedure against a sham or placebo procedure, and the operative procedure against an alternate operative procedure. Evaluating an operative intervention against a non-operative comparator is by far the most commonly used design, but blinding as to group assignment is impossible, and expectancy bias on the part of patients and outcome assessors can affect estimates of treatment effect, especially if the surgical procedure is intended to alter subjective endpoints such as symptoms or function rather than more objective endpoints, e.g., death rates. In addition, because of participants' and doctors' treatment preferences, crossovers may be a serious problem. For example, in a recent RCT of diskectomy vs. nonoperative therapy for lumbar disk herniation, only 60% of people randomized to surgery actually had the surgery, while 45% of those randomized to the nonoperative arm crossed over and had the surgery.[28] The use of a sham procedure as a comparator in an RCT is limited, among other things, by the risks associated with sham anesthesia and a sham procedure. These are dictated by the nature

of the active invasive procedure that is under evaluation. For many procedures, it would be impossible to design a sham that would maintain blinding yet still be safe for the patient. Ethical controversies about sham-procedure controlled surgical trials continue to be debated.[29,30] Few placebo-controlled trials of surgical procedures have been conducted; beside the knee arthroscopy trial already mentioned, the Parkinson's disease "burr hole" study is another recent example.[31] Finally, comparing an invasive intervention to an invasive procedure that is part of the accepted standard of care is that such a comparison is only of value if we are certain about the efficacy of the comparator and if one can assume that that efficacy is the same in the experiment to be performed as it has been in the past. Blinding as to treatment group assignment is possible with the latter design, as it is with sham procedure controls. As Baruch Brody has said regarding the issue of blinding in invasive intervention trials, one needs a "…balancing of the scientific gains from blinding against the burdens imposed on the subjects and deciding when the burdens are too great".[32] Table 6.1 summarizes the limitations of each of the above approaches.

Invasive therapeutic procedures pose other challenges in the design of randomized trials to evaluate their efficacy, including:

- The need to refine the surgical technique in humans: implications for the timing of RCTs
- Learning curves of individual surgeons
- Unequal technical skill in the individual surgeon for various procedures
- Patient – and doctor! preferences for operative vs. nonoperative intervention
- Clinical uncertainty and equipoise: who defines these?
- Modest effect sizes expected from most therapeutic interventions and implications for sample size and number of participating surgical centers
- Difficulty of evaluating effects of an intervention aimed at alleviating subjective parameters such as pain and discomfort
- Placebo effect associated with invasive therapeutic procedures and
- Control of expectancy bias in outcome assessments (blinding of patient, surgeon, outcome assessors)

**Table 6.1** Choices of comparator in controlled trials of invasive therapeutic procedures

| | Comparator | | |
|---|---|---|---|
| | Nonoperative therapy | Alternative invasive procedure | Sham procedure |
| Random allocation possible (controls selection bias) | Yes | Yes | Yes |
| Blinding of patients possible (controls expectancy bias) | No | Yes | Yes |
| Blinding of outcome assessors possible | No | Sometimes | Yes |
| Minimization of crossovers (preserves best attributes of random allocation) | No | Yes | Yes |

In summary, the current standard of practice is that invasive therapeutic procedures are devised and become widely used in the public without first having been put to scientifically valid demonstrations in humans (i.e., randomized controlled trials); and, that their benefits exceed their harms and costs and those of alternative courses of therapy. Additionally, "promising but unproven" procedures are performed for decades before being tested in well planned and well conducted RCTs, and many in common use have never been tested under such circumstances. Compared with pre-release standards for prescription drugs, those for invasive procedures seem antiquated at best. As Weinberg stated, "we need a way to assure the American people that the needed evaluations of clinical theory are done in a timely way, before plausible but wrong ideas get institutionalized into the everyday practice of medicine".[33]

## Adverse Event Reporting

The aforementioned paragraphs address the industries role in drug development, and its lack of a role in surgical procedure development. From the FDA standpoint, one of the more important interests is in monitoring the trials as they proceed and to ensure patient safety during the process. Thus, for each trial, a mechanism must be in place for a timely review of adverse events. In fact, one FDA report cited the failure to report adverse events as required as one of the top ten problems surrounding clinical trials. The FDA definition of an adverse event is "any unfavorable and unintended sign, symptom, or disease temporally associated with the use of a medical treatment or procedure regardless of whether it is considered related to the treatment or procedure."

Adverse drug events (ADEs) are classified by the FDA as serious when death, life threatening occurrences, hospitalization, persistent or permanent disability, or the need for medical or surgical intervention occurs during (and up to 30 days after) a clinical trial. An example of this is the report by Suntharalingam et al. which occurred during a phase 1 trial. They describe the events that occurred when six healthy volunteers received a dose of TGN1412 (a monoclonal antibody that affects T-cells). In all six subjects, a life threatening cytokine-release syndrome developed.

There are a number of questions that address adverse event reporting as follows:

### *Are clinical trials powered in such a way as to address differences in ADE's vs. placebo or active control?*

The answer to this is generally no. Phase 1–3 trials are powered based on presumed efficacy beyond that of the control treatment, not based upon any ADE frequency. Also, the entire drug development portfolio submitted to the FDA for drug approval

may consist of fewer than 5,000 patients exposed and certainly fewer than 10,000. Most of those patients are represented by phase 3 trials, and by the time a phase 3 trial is launched common ADE's will have already been ascertained. Given this, ADE's that occur even at a rate of 1 in 10,000 will not be revealed.

## *Does the manner in which ADE'S are ascertained matter?*

This is a frequently argued point in which there is insufficient information to come to a meaningful conclusion. Of course, most studies report ADE frequency, but the absolute frequency depends upon whether ADE's are ascertained verbally either by general questions (e.g. "have you had any new symptoms since the last visit" or specifically, e.g. "have you had any headaches since the last visit?"); or ascertained by checklists either filled out by the patient or elicited by the study coordinator and/or the PI. One can immediately see the strengths and weaknesses of each approach. One of the attempts to evaluate these differences comes from the Acute Myocardial Infarction Study (AMIS) as shown in Table 6.2. Not surprisingly, compared to controls, the frequency of GI bleeding elicited by specific questions was greater than those that were volunteered observations, but the relative difference between the active and control treatments was nearly the same.

## *Does the use of surrogate endpoints affect the determination of ADE frequency?*

Recall that a surrogate endpoint is an outcome used in lieu of the real outcome of interest, and the main reason surrogate endpoints are used is so the clinical trial will be of shorter duration and/or have a smaller sample size. It is thus obvious that this would decrease ones ability to uncover infrequent ADE's. Surrogate endpoints are more fully discussed in Chapter 3.

**Table 6.2** Percentage reporting selected ADEs in AMIS

| *Volunteered* | Hematemesis | Tarry stools | Bloody stools |
|---|---|---|---|
| ASA | 0.27 | 1.34 | 1.29 |
| Placebo | 0.09 | 0.67 | 0.45 |
| *Elicited* | | | |
| ASA | 0.62 | 2.81 | 4.86 |
| Placebo | 0.27 | 1.74 | 2.99 |

The percentage is different for volunteered vs elicited; but, placebo: ASA differences were the same

### *Does the use of intention-to-treat analysis affect the determination of ADE frequency?*

As with the use of surrogate endpoints, ITT analysis can reduce ones ability to determine the true ADE frequency. This is because, if a patient drops out from a trial before completion, and does not receive the drug for the entire trial duration, they will not have been fully exposed to the drug under study for the full time period. Even if they are dropped for an ADE (which of course would be counted), they might have had an additional ADE, had they been able to continue. Since ITT is the primary analysis of a RCT (already a relatively short trial for the reasons mentioned in Chapter 3) most RCTs underestimate the true ADE frequency.

## The FDA and Advertising

The FDA has a clear mission of protecting the public health by assuring the safety, efficacy, and security of human drugs…… The FDA is also responsible for advancing the public health by helping to speed innovations that make medicines more effective, safer, and more affordable.[34] If we consider that the FDA is also responsible to help the public get accurate, science-based information that is needed for medicines to improve their health, then it is understandable that a key role of the FDA is as a regulator and supervisor of manufacturer promotional activities.

The Division of Drug Marketing and Communications (DDMAC) in the Center for Drug Evaluation and Research, at the US Food and Drug Administration (FDA), is responsible for reviewing sponsor promotional materials, including prescription drug advertising, promotional labeling, and materials prepared for prescribers.[35] The main objective of the Division is to ensure that information about prescription drugs disseminated by sponsors to health care providers and consumers is not false or misleading, that there is fair balance of benefit/risk information,[36] and that it is accurately communicated.[37]

Since 1962, the FDA was granted the responsibility to regulate prescription drug advertising and labeling.[38,39] The regulations include reviewing written, printed, or graphic material accompanying a regulated product (“promotional labeling”) and materials published in journals and newspapers, broadcast, and telephone communications systems.[38,40] However, the FDA does not have the authority to require sponsors to submit promotional materials for approval prior to their use.[41] According to the Food, Drug and Cosmetics Act, manufacturers in their advertisements should include a brief summary which truthfully communicates the product’s indication, major side effects and contraindications, major warnings, significant precautions, drug interactions, and they should present an adequate balance of risks and benefits. For broadcast ads, two options are available to communicate drug information: a brief summary or a toll-free telephone number or website.[42]

Because manufacturers are not required to submit copies of advertisements at the time of initial dissemination nor copies of advertising at the time of initial pub-

lication,[43] the FDA sees promotional materials only after they have been released or broadcasted.[44] However, many manufacturers do submit their materials before airing to avoid future problems. Once an advertisement is disseminated, if it contains violative messages, the FDA can require corrective actions by means of untitled letters, warning letters, injunctions and consent decrees, referrals for criminal investigation, or prosecution and seizures.[44]

Untitled letters or notices of violation are issued for less serious violations and they usually require the sponsor to discontinue use of false or misleading advertising materials. Warning letters are usually issued when there are more serious violations (e.g. repetitive misconduct or there is a potential for serious health risks to the public).[37] Warning letters contain a statement that failure to respond may result in another regulatory action and that the FDA can initiate court proceedings for a seizure, injunction, or criminal prosecution.[39] Therefore, when manufacturers receive a warning letter, they are supposed to correct the problem immediately and disseminate the correct message using mailings and journals. However, a previous study showed that the FDA enforcement actions against false and misleading drug ads declined in 2002 and that there were delays in enforcement actions.[45–47]

In November 2005, The Pharmaceutical Research and Manufacturers of America (PhRMA) issued some principles on the advertising of prescription drugs but the effect of those guidelines on warning letters is unknown. As a result of the above, Salas et al. described the number, type, and content of warning letters for prescribed medications and to assess if PhRMA guidelines had an effect on the number and content of warning letters issued. They found that 25% of the overall warning letters issued by the FDA were related directly with drugs and that 10% were focused on drug-related promotional activities. They also found that half of the warning letters were issued because of superiority claims which encourage prescriber's not only to use drugs but also to try the use of drugs for non approved indications (i.e. off-label uses). In addition, they found an increase in warning letters issued in 1998 compared to previous years, which may be an effect of changes in the 1997 law. According to this law, the Food and Drug Administration Modernization Act of 1997 reauthorizes the Prescription Drug User Fee Act of 1992, regulating advertising of unapproved uses of approved drugs,[48] and it released a draft guidance for direct to consumer advertising, which might have influenced an increase in the production of promotional materials.

In summary, the USFDA has a long history of regulating new drug development, and in trying to insure the safety of drugs both before and after they reach the marketplace. The regulatory authority granted to the FDA is a dynamic process and the constant changes require continual updating of ones knowledge.

## References

1. Lewis S, Baird P, Evans RG, et al. Dancing with the porcupine: rules for governing the university-industry relationship. *CMAJ*. Sept 18, 2001; 165(6):783–785.
2. *The Historical Guide to American Government*. New York: Oxford Press; 1998.
3. http://store.aetv.com/html/product/index.jhtml?id = 73174.

4. Swann R. History of the FDA. www.fda.gov/oc/history. Accessed May 9, 2007.
5. Guidance for Industry. www.fda.gov/cber/guidelines.
6. Thelithromycin. *Wikipedia*.
7. FDA Amendment Act of 2007; 2007.
8. The Mission Statement of the ICH. http://www.ich.org/
9. Coronary Drug Project. www.fda.gov
10. Suntharalingam G, Perry MR, Ward S, et al. Cytokine storm in a phase 1 trial of the anti-CD28 monoclonal antibody TGN1412. *N Engl J Med*. Sept 7, 2006; 355(10):1018–1028.
11. European Medicines Agency (EMEA). http.www.emea.europa.
12. O'Donnell P. Not yet the last word on first-in-man. *Appl Clin Trials*. 2007; 16:34–38.
13. Palesch YY, Tilley BC, Sackett DL, Johnston KC, Woolson R. Applying a phase II futility study design to therapeutic stroke trials. *Stroke*. Nov 2005; 36(11):2410–2414.
14. Henderson L. The long arm of the FDA. *Appl Clin Trials*. 2007.
15. Ramsey SD, Sullivan SD. Evidence, economics, and emphysema: medicare's long journey with lung volume reduction surgery. *Health Aff (Millwood)*. Jan–Feb 2005; 24(1):55–66.
16. Brantigan OC, Mueller E. Surgical treatment of pulmonary emphysema. *Am Surg*. Sept 1957; 23(9):789–804.
17. Cooper JD, Trulock EP, Triantafillou AN, et al. Bilateral pneumectomy (volume reduction) for chronic obstructive pulmonary disease. *J Thorac Cardiovasc Surg*. Jan 1995; 109(1): 106–116; discussion 116–109.
18. Tunis SR, Pearson SD. Coverage options for promising technologies: medicare's 'coverage with evidence development'. *Health Aff (Millwood)*. Sept–Oct 2006; 25(5):1218–1230.
19. Fishman A, Martinez F, Naunheim K, et al. A randomized trial comparing lung-volume-reduction surgery with medical therapy for severe emphysema. *N Engl J Med*. May 22, 2003; 348(21):2059–2073.
20. Kolata G. Medicare says it will pay, but patients say 'no thanks'. *New York Times*. March 3, 2006, 2006; C:1.
21. Al-Naaman YD, Carton CA, Cooley DA. Surgical treatment of arteriosclerotic occlusion of common carotid artery. *J Neurosurg*. Sept 1956; 13(5):500–506.
22. Winslow CM, Solomon DH, Chassin MR, Kosecoff J, Merrick NJ, Brook RH. The appropriateness of carotid endarterectomy. *N Engl J Med*. Mar 24, 1988; 318(12):721–727.
23. Beneficial effect of carotid endarterectomy in symptomatic patients with high-grade carotid stenosis. North American Symptomatic Carotid Endarterectomy Trial Collaborators. *N Engl J Med*. Aug 15, 1991; 325(7):445–453.
24. Endarterectomy for asymptomatic carotid artery stenosis. Executive Committee for the Asymptomatic Carotid Atherosclerosis Study. *JAMA*. May 10, 1995; 273(18):1421–1428.
25. Barnett HJ, Taylor DW, Eliasziw M, et al. Benefit of carotid endarterectomy in patients with symptomatic moderate or severe stenosis. North American Symptomatic Carotid Endarterectomy Trial Collaborators. *N Engl J Med*. Nov 12, 1998; 339(20):1415–1425.
26. Halm EA, Tuhrim S, Wang JJ, Rojas M, Hannan EL, Chassin MR. Has evidence changed practice?: appropriateness of carotid endarterectomy after the clinical trials. *Neurology*. Jan 16, 2007; 68(3):187–194.
27. Moseley JB, O'Malley K, Petersen NJ, et al. A controlled trial of arthroscopic surgery for osteoarthritis of the knee. *N Engl J Med*. July 11, 2002; 347(2):81–88.
28. Weinstein JN, Tosteson TD, Lurie JD, et al. Surgical vs nonoperative treatment for lumbar disk herniation: the Spine Patient Outcomes Research Trial (SPORT): a randomized trial. *JAMA*. Nov 22, 2006; 296(20):2441–2450.
29. Horng S, Miller FG. Ethical framework for the use of sham procedures in clinical trials. *Crit Care Med*. Mar 2003; 31(3 Suppl):S126–130.
30. Macklin R. The ethical problems with sham surgery in clinical research. *N Engl J Med*. Sept 23, 1999; 341(13):992–996.
31. Freed CR, Greene PE, Breeze RE, et al. Transplantation of embryonic dopamine neurons for severe Parkinson's disease. *N Engl J Med*. Mar 8, 2001; 344(10):710–719.

32. Brody BA. *The Ethics of Biomedical Research: An International Perspective*. New York: Oxford University Press; 1998.
33. Wennberg JE. An apple a day? *N Engl J Med*. Sept 22, 1994, 1994; 331(12):815–816.
34. FDA Website. 6/10/07; http://www.fda.gov/opacom/morechoices/mission.html
35. Division of Drug Marketing, Advertising, and Communications, Food and Drug Administration. 5/31/07; http://www.fda.gov/cder/ddmac
36. 21 CFR Part 310 section 502(a) of the Food and Drug Administration Modernization Act of 1997. In: Department of Health and Human Services FaDA, ed. Vol 21 U.S.C. 352(a).
37. Baylor-Henry M, Drezin N. Regulation of prescription drug promotion: direct-to consumer advertising. *Clin Ther*. 1998; 20(C):C86–C95.
38. Section 502(n) of the Food Drug and Cosmetics Act, and Title 21 Code of Federal Regulations. Vol 202.1(1)(1).
39. Kessler DA, Pines WL. The federal regulation of prescription drug advertising and promotion. *JAMA*. Nov 14, 1990; 264(18):2409–2415.
40. 21 CFR Part 310 section 502(a) of the Food and Drug Administration Modernization Act of 1997 (Modernization Act).. In: Department of Health and Human Services FaDA, ed. Vol 21 U.S.C. 352 (n).
41. Section 502(n) of the Food Drug and Cosmetics Act, and Title 21 Code of Federal Regulations. Vol 202.1(j)(1).
42. Section 502 (n) of the Food Drug and Cosmetics Act, and Title 21 Code of Federal Regulations. Vol 202.1.
43. 21 CFR Part 310 section 502(a) of the Food and Drug Administration Modernization Act of 1997 (Modernization Act). Vol 21 314.81(b)(3).
44. Woodcock J. *Statement by Janet Woodcock, MSD Director, Center of Drug Evaluation and Research. US Drug Administration*. Rockville, MD: Department of Health and Human Services; 2003.
45. Gahart MT, Duhamel LM, Dievler A, Price R. Examining the FDA's oversight of direct to consumer advertising. *Health Affairs*. 2003; W3:120–123.
46. Waxman HA. Ensuring that consumers receive appropriate information from drug ads: what is the FDA's role? *Health Affairs*. 2004; W4:256–258.
47. Waxman RHA. Letter from Rep. Henry A. Waxman to the Honorable Tommy G. Thompson. http://oversight.house.gov/story.asp?ID = 441. Accessed June 1, 2007.
48. Food and Drug Administration. http://www.fda.gov/opacom/backgrounders/miles.html. Accessed October 6, 2007.

# Chapter 7
# The Placebo and Nocebo Effect

**Stephen P. Glasser and William Frishman**

*If a placebo were submitted to the FDA for approval, they would no doubt be impressed with its efficacy, but would probably not approve it due to its frequent side effects.*

*Anon*

**Abstract** There are four general reasons for clinical improvement in a patient's condition: (1) natural history of the disease; (2) specific effects of the treatment; (3) regression to the mean; and (4) nonspecific effects of the treatment that are attributable to factors other than the specific active components. The latter effect is included under the heading 'placebo effect'. In this chapter the placebo effect will be discussed, with some emphasis on regression to the mean. Placebos ('I will please') and their lesser known counterpart's nocebo's (I will harm') are sham treatments. The difference is in the response to the inert therapy. A beneficial response to an inert substance is a placebo response; a side effect to an inert substance is a nocebo response.

Placebo has been cited in PubMed over 100,000 times indicating that placebo has set the standard for how clinical research and particularly clinical trials are conducted. On the other hand, some have argued that placebo effects are overstated and can be explained by other variables (e.g. changes in the natural history of the disease, regression to the mean, methodological issues, conditioned answers, etc.). Because of the importance, controversy, and to date inadequate study of the placebo effect, this chapter presents more detail than many of the other chapters. In addition, the discussion of placebos requires an understanding of the ethics of clinical trials, intention to treat analysis, surrogate endpoints and many of the other areas that have been discussed. As such this chapter can also be used to review those concepts.

Placebos ('I will please') and their lesser known counterpart's nocebo's (I will harm') are sham treatments. The difference is in the response to the inert therapy. A beneficial response to an inert substance is a placebo response; a side effect to an inert substance is a nocebo response.

There are four general reasons for clinical improvement in a patient's condition: (1) natural history of the disease; (2) specific effects of the treatment; (3) regression

S.P. Glasser (ed.), *Essentials of Clinical Research*,

to the mean; and (4) nonspecific effects of the treatment that are attributable to factors other than the specific active components. The latter effect is included under the heading 'placebo effect'.[1] Each time a physician recommends a diagnostic or therapeutic intervention for a patient, built into this clinical decision is the possibility of a placebo effect, that is, a clinical effect unrelated to the intervention itself.[2] Simple diagnostic procedures such as phlebotomy or more invasive procedures such as cardiac catheterization have been shown to have important associated placebo effects.[3,4] Chalmers[5] has stated that a simple review of the many abandoned therapies reveals that many patients would have benefited by being assigned to a placebo control group. In fact, what might represent the first known clinical trial, and one in which the absence of a placebo control group led to erroneous conclusions, is a summary attributed to Galen in 250 BC, who stated that 'some patients that have taken this herbivore have recovered, while some have died; thus, it is obvious that this medicament fails only in incurable diseases.'[6]

Placebo effects are commonly observed in patients with cardiac disease who also receive drug and surgical therapies as treatments. Rana et al. noted the 'tremendous power of the placebo effect' in patients with end-stage coronary disease in clinical trials of angiogenesis and laser myocardial revascularization.[7] They also commented on the fact that the observed improvements were not limited to 'soft' symptomatic endpoints but were also observed with 'hard' endpoints such as exercise time, and in magnetic resonance imaging.[7] Rana et al. also studied the longevity of the placebo effect from published clinical trials. They found that the beneficial effects of placebo (on angina class, angina frequency, and exercise time) persisted over the long term (up to 2 years).

## Definition

Stedman's Medical Dectionary[8] defines the word 'placebo,' which originates from Latin verb meaning 'I shall please,' to have two meanings. First, a placebo may be an inert substance prescribed for its suggestive value. Second, it may be an inert substance identical in appearance with the compound being tested in experimental research, and the use of which may or may not be known by the physician or the patient; it is given to distinguish between the action of the compound and the suggestive effect of the compound under study.[9]

Currently, there is some disagreement as to the exact definition of a placebo.[8,9] Many articles on the subject include a broader definition, as given by Shapiro in 1961.[10]

*Any therapeutic procedure (or that component of any therapeutic procedure) which is given deliberately to have an effect or unknowingly has an effect on a patient, symptom, syndrome, or disease, but which is objectively without specific activity for the condition being treated. The therapeutic procedure may be given with or without conscious knowledge that the procedure is a placebo, may be an*

*active (noninert) or nonactive (inert) procedure, and includes, therefore, all medical procedures no matter how specific—oral and parenteral medication, topical preparations, inhalants, and mechanical, surgical and psychotherapeutic procedures. The placebo must be differentiated from the placebo effect, which may or may not occur and which may be favorable or unfavorable. The placebo effect is defined as the changes produced by placebos. The placebo is also used to describe an adequate control in research'.*

A further refinement of the definition was proposed by Byerly[11] in 1976 as 'any change in a patient's symptoms that are the result of the therapeutic intent and not the specific physiochemical nature of a medical procedure.

## Placebo Effect in Clinical Trials

The use of placebo controls in medical research was advocated in 1753 by Lind[12] in an evaluation of the effects of lime juice on scurvy. After World War II, research protocols designed to assess the efficacy and safety of new pharmacologic therapies began to include the recognition of the placebo effect. Placebos and their role in controlled clinical trials were recognized in 1946, when the Cornell Conference on Therapy devoted a session to placebos and double-blind methodology. At that time, placebos were associated with increased heart rate, altered respiration patterns, dilated pupils, and increased blood pressure.[9] In 1951, Hill[13] concluded that for a change for better or worse in a patient to be attributable to a specific treatment, this result must be repeatable a significant number of times in similar patients. Otherwise, the result was due simply to the natural history of the disease or the passage of time. He also proposed the inclusion of a control group that received identical treatment except for the exclusion of an 'active ingredient.' Thus the 'active ingredient' was separated from the situation within which it was used. This control group, also known as a placebo group, would help in the investigations of new and promising pharmacologic therapies.[13]

Beecher[14] was among the first investigators to promote the inclusion of placebo controls in clinical trials. He emphasized that neither the subject nor the physician should know what treatment the subject was receiving and referred to this strategy as the 'double unknown technique.' Today, this technique is called the 'double-blind trial' and ensures that the expectations and beliefs of the patient and physician are excluded from evaluation of new therapies. In 1955, Beecher reviewed 15 studies that included 1,082 patients and found that an average of 35% of these patients significantly benefited from placebo therapy (another third had a lesser benefit).[14] He also concluded that placebos can relieve pain from conditions with physiologic or psychological etiologies. He described diverse objective changes with placebo therapy. Some medical conditions improved; they included severe postoperative wound pain, cough, drug-induced mood changes, pain from angina pectoris, headache, seasickness, anxiety, tension, and the common cold.

## Characteristics of the Placebo Effect

There appears to be an inverse relation between the number of placebo doses that needs to be administered and treatment outcomes. In a study of patients with postoperative wound pain, 53% of the subjects responded to one placebo dose, 40% to two or three doses, and 15% to four doses.[14] In analyzing the demographics of those who responded to placebo and those who did not, Lasagna et al.[15] could find no differences in gender ratios or intelligence quotients between the two groups. They did find significant differences in attitudes, habits, educational backgrounds, and personality structure between consistent responders and nonresponders.[15] In attempting to understand the reproducibility of the placebo effect, they observed that there was no relation between an initial placebo response and subsequent responses with repeated placebo doses of saline.[14] Beecher[14] concluded that placebos are most effective when stress, such as anxiety and pain, is greatest. Placebo responses are associated with dose response characteristics, frequency of dosing, pill color (e.g. blue vs. pink pills are more sedating, yellow vs. green more stimulating) and, "branded placebo" is more effective than generic placebo. The magnitude of effect is difficult to quantitate due to its diverse nature but it is estimated that a placebo effect accounts for 30–40% of an interventions benefit.

Placebos can produce both desirable and adverse reactions. Some now use the term placebo for the beneficial effects and nocebo for the adverse effects. Beecher et al.[14] described >35 adverse reactions from placebos; the most common are listed in Table 7.1. These reactions were recorded without the patient's or physician's knowledge that a placebo had been administered. In one study in which lactose tablets were given as a placebo, major adverse reactions occurred in three patients[16] The first patient had overwhelming weakness, palpitation, and nausea after taking the placebo and the test drug. In the second patient, a diffuse rash developed and then disappeared after placebo administration was discontinued. The third patient had epigastric pain followed by watery diarrhea, urticaria, and angioneurotic edema of the lips after receiving the placebo and the test drug.[16]

**Table 7.1** Most common adverse reactions from placebo therapy (nocebo effect)

| Reaction | Incidence (%) |
|---|---|
| Dry mouth | 9 |
| Nausea | 10 |
| Sensation of heaviness | 18 |
| Headache | 25 |
| Difficulty concentrating | 15 |
| Drowsiness | 50 |
| Warm glow | 8 |
| Relaxation | 9 |
| Fatigue | 18 |
| Sleep disturbance | 10 |

Indeed, because of the substantial evidence of placebo 'efficacy' and placebo 'side effects,' some investigators have wittingly suggested that if placebo were submitted to the United States Food and Drug Administration (FDA) for approval, that the agency, though impressed with the efficacy data, would probably recommend disapproval on the basis of the high incidence of side effects. Some authors have questioned whether placebos are truly inert. Davis[17] pointed out that part of the problem with the placebo paradox is our failure to separate the use of an inert medication (if there is such as substance) from the phenomenon referred to as the placebo effect. It might help us if we could rename the placebo effect the 'obscure therapeutic effect.'

For instance, in trials of lactase deficiency therapy, could the amount of lactose in placebo tablets actually cause true side effects? The small amount of lactose makes this possibility seem unlikely. Perhaps it is more likely that allergies to some of the so-called inert ingredients in placebos cause reactions in predisposed persons, although this explanation probably could not explain more than a small percentage of placebo side effects.

The most recent validation of the placebo effect occurred in 1962 when the United States enacted the Harris-Kefauver amendments to the Food, Drug, and Cosmetic Act. These amendments required proof of efficacy and documentation of relative safety, in terms of the risk-benefit ratio for the disease to be treated, before an experimental agent could be approved for general use.[18] In 1970, the FDA published rules for 'adequate and well-controlled clinical evaluations.' The federal regulations identified five types of controls (placebo, dose-comparison, active, historical, and no treatment) and identified use of the placebo control as an indispensable tool to achieve the standard.[19] However, the FDA does not mandate placebo controls, and in fact has stated that placebo groups are 'desirable, but need not be interpreted as a strict requirement…The speed with which blind comparisons with placebo and/or positive controls can be fruitfully undertaken varies with the nature of the compound.'[19] In the publication regarding 'Draft Guidelines for the Clinical Evaluation of Anti-anginal Drugs,' the FDA further states that 'it should be recognized that there are other methods of adequately controlling studies. In some studies, and in some diseases, the use of an active control drug rather than a placebo is desirable, primarily for ethical reasons.'[19]

## *Regression Towards the Mean (or Towards Mediocrity)*

An important statistical concept and one that many mimic a placebo response or a clinical response is regression towards the mean or regression towards mediocrity (RTM). RTM identifies a phenomenon that a variable that is extreme on its first measurement will tend to be closer to the center of the distribution on a later measurement. The term originated with Sir Francis Galton who studied the relationship between the height of parents and their adult offspring. He observed that children of tall parents were (on average) shorter than their parents; while, children of short

parents were taller than their parents. Galton called this regression towards mediocrity.[20] Another example of RTM from Ederer, who observed that during the first week of the 1968 baseball season the top 10 and bottom 10 batters averaged 0.414 and 0.83 respectively. The following week they hit 0.246 and 0.206 while the average for the league remained stable.[21]

At least three types of studies are potentially affected by RTM: a survey in which subjects are selected for subsequent follow-up based upon an initial extreme value, studies with no control groups, and even controlled trials. An example is taken from the Lipid Research Clinics Prevalence Study, a sample population who had elevated total cholesterol was asked to return for reevaluation. It would be expected that the second measurement would on average be lower, and this would not be so had a randomly selected sample been chosen for reevaluation.[22] The reason that a randomly selected sample would be less likely to demonstrate RTM is because the random sample would have representative values across the spectrum of cholesterol measurements at the start, whereas the selected sample all initially had elevated values. In studies that lack a control group, it is difficult to estimate RTM since the best way to evaluate for RTM is to have a placebo control. But, even in controlled clinical trials, RTM can be problematic. For example, in many trials subjects are identified in two stages; at first screen, subjects with extreme values are asked to return (and invariably have lower values) for entrance into the study. The choice of baseline from which to measure the treatment effect then becomes an issue.

There are ways to limit the RTM effect. For example one can use the control group to estimate RTM. Also, taking at least two pretreatment measures and using the first to classify the subject and the second for baseline comparison, or using the average of two or more measures, will be helpful. An example of the RTM principal comes from the National Diet-Heart Study.[23] It had been repeatedly observed that a low cholesterol diet given to subjects with high cholesterol values results in greater cholesterol lowering that when the same diet is given to someone with lower cholesterol values. In the National Diet-Heart Study subjects with a baseline cholesterol > 242 mg/dL had a 15% reduction while those whose baseline cholesterol was 210–241 mg/dL had a 12% reduction.[23] There are two possible explanations of this observation: one, that the diet hypothesis holds i.e. that subjects with high cholesterol are more responsive to cholesterol lowering treatment than those with lower cholesterol values; and two, that independent of dietary intervention subjects with high cholesterol will (on average) decrease more than those with lower values due to RTM. In fact, it is likely that both could occur simultaneously.

RTM then, is a phenomenon that can make a natural variation in repeated data look like a real change. In biologic systems, most variables increase and decrease around a mean (as, for instance, might be visualized as a sine wave). Thus, it is likely that any value measured at a specific point in time will, by chance, either be above or below the mean, and that a second measurement will be at a different point around the mean and therefore different from the first measurement (Fig. 7.1). The presumption is that this variability about the mean will be the same in the placebo group as in the active treatment group (assuming adequate sample size and randomization), so that differences between the two groups relative to regression to the

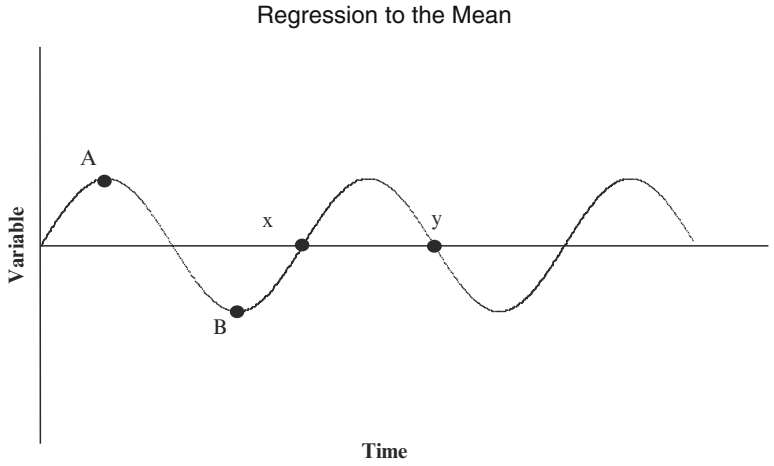



**Fig. 7.1** If one measures a variable at its peak value (A in the example) the next measurement is likely to be lower (B, x, or y in this example). Conversely, if one were to measure a variable at its lowest point (B), the next measurement is likely to be higher

mean will cancel out. When there is no placebo group, the distinction regarding whether RTM has occurred is even more problematic. In an intervention study, RTM cannot be observed because it is mixed into the genuine intervention effect. This is particularly true of intervention studies where the population selected for study generally is in the high risk groups – that is with values that are high at baseline. Yudkin and Stratton evaluated this by analyzing a group with high baseline cholesterol, and observing a 9% fall without any intervention.[24] These authors go on to point out several ways of estimating the impact of RTM, and three suggested approaches to minimizing the RTM problem. These approaches include the use of an RCT design, since the RTM effect will be part of the total effect of the response in both the intervention and control groups. However, the response in both groups will be inflated by the RTM so the true impact of the intervention is not known and is likely somewhat less that that observed. A second approach to minimizing RTM is to obtain several measurements and average them to determine baseline. The third approach is to use the first measurement as the basis for selection of the subject into the study, and a second measurement which will be used as the baseline from which to assess the effect of the intervention.

The ideal comparator for a study would actually be no therapy vs. the investigational agent, however, the loss of blinding makes this problematic. There has been little study of the no therapy control, however, Asmar et al. did attempt to evaluate this as part of a larger interventional trial.[25] They used a randomized cross-over approach with a 1 month run-in followed by a 1 month placebo vs. no treatment period. BP and ABPM were measured. The results could be then analyzed in terms of the no treatment effect (no parameters changed in the two periods) and the RTM effect shown in Fig. 7.2.

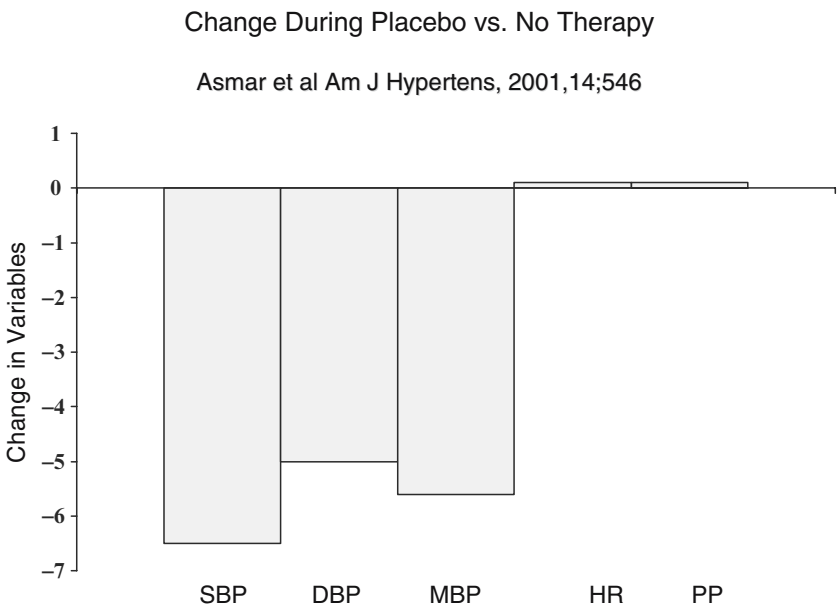



**Fig. 7.2** Change during placebo vs. no therapy[25]

## Mechanism of the Placebo Effect

Much has been discussed about the mechanism of the placebo response in patients. However, the mechanism at the cellular level and the role of biochemical mediators continues to escape detection.

Beecher[14] described two phases of suffering: first, the initial pain sensation or other symptom, and second the person's reaction to this sensation or experience by the central nervous system. The first, or somatic, phase is associated with the source of the pain or symptom; the second, or cortical, phase is superimposed on the pain or symptom. An example of the influence of the effect of the mind on the body is the 'Anzio Effect.' During World War II, injured soldiers at Anzio, Italy, complained less of pain after surgery, than typical patients after surgery. This difference was recognized because less than one third of the injured soldiers required morphine, compared with four fifths of patients undergoing similar recovery from the same surgery in non combatants. For the soldiers, the knowledge that they had survived, combined with the anticipation of returning home, probably reduced their pain. Typical surgical patients are required to comply with hospital procedures, probably producing anxiety or fear that acts to increase pain.[26]

The physiologic mechanism begins when fear or anxiety activates the hypothalamus-hypophysis-adrenal axis, resulting in release of catecholamines. These catecholamines act on the body, which then sends feedback to the cerebral cortex via

neural connections. The thalamus in the diencephalons, which processes sensory input before relaying it to the cerebral cortex, then sends recurrent axons to the thalamus, presumably to allow modulation of the input received from the thalamus.[26,27]

One theory to explain the placebo effect is classical conditioning, the pairing of an unconditioned stimulus with a conditioned stimulus until eventually the conditioned stimulus alone elicits the same response as the unconditioned stimulus. This effect of the environment on behavior was tested in a study by Voudouris et al.[28] They studied responses to pain stimulation with and without a placebo cream. A visual analogue scale determined pain perception. To evaluate the effect of verbal expectancy, the patients were informed that the placebo cream had powerful analgesic properties (expectancy) or that the cream was neutral (no expectancy). To determine the role of conditioning, the level of pain stimulus was reduced after application of the cream (conditioning) or was maintained at the same level of pain (no conditioning). The patients were divided into four groups: a group receiving expectancy and conditioning, a group receiving only expectancy, a group receiving only conditioning, and a group receiving neither. Both conditioning and verbal expectancy were important mediators on the placebo response, but conditioning was more powerful.[28]

A second explanation for the placebo effect is response by neurohormones, including motor or autonomic nervous systems, hormone systems, and immune systems. Endogenous neuroendocrine polypeptides, including β-endorphins, enkephalins, and antiopioids, are activated by many factors. These factors include placebos, vigorous exercise, and other stressors. Modulation of the opioid system may occur by an antiopiod system of neurotransmitters. γ-Aminobutyric acid, and peptide neurotransmitter, is associated with the secretion of β-endorphin and β-lipotropin.[26]

The endorphin group of neurotransmitters is created from the proopiomelanocortitrophin peptide and is linked through β-lipotropin with the regulation of the hypothalamus-hypophysis-adrenal axis. There is no understanding of the exact link between the opioid-antiopioid and β-lipotropin systems of neuroendocrine peptides. The brain peptides and their actions on presynaptic and postsynaptic receptors on neurons also are not understood. Experiments in animals provide most of the information about control of the genetic expression of the peptides.[26]

In a double-blind study by Levine et al.,[29] patients received placebo and then intravenous naloxone after tooth extraction. Naloxone, a partial opioid antagonist that competes with β-endorphins for the same receptor in the brain, blocked the placebo effect previously experienced by the patients. Levine et al.[29] concluded that placebo activates β-endorphins in the brain and that naloxone increases the pain by inhibiting the placebo effect.

A double-blind study by Hersh et al.[30] found ibuprofen to be more efficacious than placebo or codeine. Naltrexone, a long-acting oral form of naloxone, given before oral surgery reduced the analgesic response to placebo and to codeine received after surgery. In an additional noteworthy finding, pretreatment with naltrexone prolonged the duration of ibuprofen's action rather than diminishing the

peak analgesic response. This prolongation of ibuprofen's action was hypothesized to result from increased central stimulation of endogenous opiates by ibuprofen or from competition by naltrexone for liver enzymes involved in the inactivation and elimination of ibuprofen.

A third model of the placebo response is the ability of mental imagery to produce specific and measurable physiologic effects. This model explains the relation between psychological and physiologic components of the placebo effect. There is a conversion in the brain of psychological placebo-related imagery into a physiologic placebo response. A patient may modify his or her imagery content in response to bodily reactions during treatment, in response to the behaviors and attitudes of doctors or nurses, or in response to information about the treatment from other sources (such as other patients, books, and journals).[31]

An example of this model is described in another study.[32] Two matched groups of patients preparing to undergo abdominal surgery received different types of care. In one group, the anesthesiologist told the patients about the operation but not about the postoperative pain. The other group was told about the postoperative pain and assured that medication was available. The study found that the patients informed about the postoperative pain needed only half the analgesic and left the hospital 2 days earlier. The authors concluded that this result showed 'a placebo effect without the placebo.'[32]

## Placebo Effect in Various Diseases

### *Placebo Effect in Ischemic Heart Disease and Chronic, Stable, Exertional Angina Pectoris*

The rate of improvement in the frequency of symptoms in patients with chronic, stable, exertional angina pectoris with placebo therapy has been assessed to be 30–80%.[33] A summary of subjective and objective placebo effects in cardiovascular disease is provided in Tables 7.2 and 7.3. Because of the magnitude of the placebo effect, most studies of new antianginal therapies were performed with placebo control. However, the safety of this practice came under scrutiny in the late 1980s because of concern that patients with coronary artery disease would have periods of no drug treatment. As a result, Glasser et al.[34] explored the safety of exposing patients with chronic, stable, exertional angina to placebos during short-term drug trials with an average double-blind period of 10 weeks. The study samples were taken from new drug applications submitted to the FDA. The results of these drug trials were submitted, whether favorable or not, and all adverse events were reported. Qualifying studies used symptom-limited exercise tolerance testing as an end point. No antianginal medication, except sublingual nitroglycerin, was taken after a placebo-free or drug-free washout period. The placebo-controlled samples consisted of 12 studies, 6 studies using β-adrenergic blocking agents and 6 studies using calcium

antagonists.[34] Of 3,161 patients who entered the studies, 197 withdrew because of adverse cardiovascular events. Adverse events with β-Blocker therapy was not significantly different when compared with placebo therapy, while calcium antagonist therapy had a significantly higher rate of cardiovascular events compared with placebo therapy, This analysis by Glasser et al.[34] found evidence that supported the safety of a placebo group in short-term drug trials for chronic, stable, exertional angina. An analysis of the safety of a placebo control in trials of anti-hypertensive drugs has also been published.[35] Although a slightly increased risk of reversible symptoms was identified, there was no evidence of irreversible harm as a result of participation in these trials. The same caveats apply as discussed in the angina trials-that is, these were short term trials of carefully monitored and selected patients.

**Table 7.2** Symptomatic placebo effects in cardiovascular disease

| | Placebo effect (%) |
|---|---|
| Improvement in chronic, stable Angina pectoris | 30–80 |
| Improvement in heart failure | 25–35 |

**Table 7.3** Objective placebo effects in cardiovascular disease

| | Placebo effect |
|---|---|
| Heart failure[38] | |
| Exercise tolerance testing | |
| 1 or 2 baseline measurements | 90–120 seconds |
| 3–10 baseline measurements | 10–30 seconds |
| Increase in ejection fraction of 5% | 20–30% of patients |
| Hypertension[54] measured by noninvasive automatic ambulatory 24-hour monitoring | 0% |
| arrhythmia study 1[64a] | |
| A reduction in mean hourly frequency of ventricular tachycardia | <65% |
| A reduction in mean hourly frequency of couplets | <75% |
| A reduction in mean hourly frequency of all ventricular ectopic beats without regard for complexity study 2[65b] | <83% |
| Baseline VPCs > 100/hour | <3 times baseline |
| Baseline VPCs < 100/hour | <10 times baseline |
| Silent ischemic disease26 reduction in frequency of ischemic events | 44% |
| Reduction in ST-segment integral | 50% |
| Reduction in duration of ST-segment depression | 50% |
| Reduction of total peak ST-segment depression | 7% |
| Other[68,70,73] | |
| Compliance with treatment at rate of ≥75% | <3 times baseline |

VPC, ventricular premature complexes

[a] Based on comparison of one control 24 hour monitoring period to one 24-hour treatment period. Variability is so great that it may be inadvisable to pool individual patient data to detect trends in ectopic frequency in evaluating new potential antiarrhythmic agents in groups of patients.

[b] When differentiating proarrhythmia in patients with mixed cardiac disease and chronic ventricular arrhythmias from spontaneous variability, with false-positive rate of only 1%

The safety of using placebo in longer-term drug trials for chronic, stable, exertional angina has not been established. A placebo-controlled trial by a European group in 1986 enrolled 35 patients and made observations during a 6-month period of placebo or short-acting nitroglycerin administration.[36] This study of the long-term effects of placebo treatment in patients with moderately severe, stable angina pectoris found a shift toward the highest dosage during the titration period. Seven patients continued to receive the lowest dosage, but the average ending dosage was 65% more than the initial dosage. Compliance, when determined by pill count, for 27 patients was >80%. During the first 2.5 months of the trial, noncompliance with the regimen or physical inability to continue to study was ascertained. No patients died or had myocardial infarction.[36]

There is a paucity of information regarding any gender differences in placebo response. Women represented 43% of the population in the aforementioned European study[36] and were more likely to have angina despite normal coronary arteries. Because the placebo effect may be more pronounced in patients with normal coronary arteries, data from men were analyzed separately to compare them with the overall results. However, the data from men were very similar to the overall results. In fact, the functional status of men showed more improvement attributable to placebo (61%) than overall (48%) at 8 weeks. The results of this study showed no adverse effects of long-term placebo therapy: 65% of patients reported subjective, clinical improvement and 27% of patients reported objective, clinical improvement in exercise performance.[36] Of note, improvement in exercise performance can occur when patients undergo repeated testing.[37]

There is a problem inherent in all modern trials of antianginal therapy: because anginal patterns vary and, with modern treatments, are infrequent, a surrogate measure of antianginal effect has been adopted by the FDA and consists of treadmill walking time to the point of moderate angina. Also, just as there is a placebo effect on angina frequency, a patient's treadmill walking time frequently (50–75%) improves with placebo therapy. Other potential mechanisms also partially explain the improvement in exercise walking time in antianginal studies and are unrelated to a treatment effect: they are the 'learning phenomenon,' and the 'training effect.' Because of the learning phenomenon, patients frequently show an improvement in walking time between the first and second treadmill test in the absence of any treatment. The presumption is that the first test is associated with anxiety and unfamiliarity, which is reduced during the second test. Of greater importance is the training effect, with which the frequency of treadmill testing may result in a true improvement in exercise performance irrespective of treatment.

The effect of placebo on exercise tolerance in patients with angina was demonstrated in the Transdermal Nitroglycerin Cooperative Study,[38] which analyzed various doses of transcutaneous-patch nitroglycerin administered for 24-hour periods, in comparison with placebo patch treatment. This study was particularly important because it was the first large study to address the issue of nitrate tolerance with transcutaneous patch drug delivery in outpatient ambulatory patients. The result of the study was the demonstration of tolerance in all treated groups; the treated groups performed no better than the placebo group at the study's end. However,

there was an equally striking improvement of 80–90 seconds in the placebo and active treatment groups in the primary efficacy end point, walking time on a treadmill. This improvement in the placebo group could have masked any active treatment effect, but it also demonstrated the importance of a placebo control, because without this type of control, significant improvement could have been attributed by deduction to active therapy.

It was once thought that internal mammary artery ligation improved angina pectoris until studies showed a similar benefit in patients in whom a sham operation, consisting of skin incision with no ligation, was performed. Beecher[39] tried to analyze the effect of doctors' personalities on clinical outcomes of internal artery ligation, by comparing the results of the same placebo procedure performed by one of two groups, the 'enthusiasts' or the 'skeptics.' His analysis indicated that the enthusiasts achieved nearly four times more 'complete relief' for patients than did the skeptics, even though the procedure has no known specific effects.[39] Five patients undergoing the sham operation emphatically described marked improvement.[40,41] In objective terms, a patient undergoing the sham operation had an increase in work tolerance from 4 to 10 minutes with no inversion of T waves on the electrocardiogram and no pain. The internal mammary artery ligation procedure was used in the United States for 2 years before it was discontinued, when the procedure was disproved by three small, well-planned, double-blind studies.[42]

Carver and Samuels[43] also addressed the issue of sham therapy in the treatment of coronary artery disease. They pointed out that although the pathophysiologic features of coronary artery disease are well known, the awareness of many of the expressions of myocardial ischemia are subjective, rendering the placebo effective more important. This factor has resulted in several treatments that are based on testimonials rather than scientific evidence and that have been touted as 'breakthroughs.' Among therapies cited by these authors are chelation therapy, various vitamin therapies, and mineral supplements. Chelation therapy is an instructive example of a widely used technique for which little scientific data are available. It has been estimated that 500,000 patients per year in the United States are treated by this technique. Before 1995, the data to support claims regarding the effectiveness of chelation therapy were obtained from uncontrolled open-label studies. In 1994, van Rij et al.[44] performed a double-blind, randomized, placebo-controlled study in patients with intermittent claudication and demonstrated no difference in outcomes between chelation and placebo treatments. The evaluated variables included objective and subjective measures, and improvement in many of the measures was shown with both therapies. Again, without the use of a placebo control, the results could have been interpreted as improvement as a result of chelation treatment.

## *Placebo Effect in Heart Failure*

Until recently, the importance of the placebo effect in patients with congestive heart failure (CHF) had not been recognized. In the 1970s and early 1980s, administration

of vasodilator therapy was given to patients in clinical trials without placebo control. Investigators believed that the cause of heart failure was predictable, so placebo-controlled trials were unnecessary. Another view of the unfavorable course of heart failure concluded that withholding a promising new agent was unethical. The ethical issues involved when placebo therapy is considered are addressed later in this article.

With the inclusion of placebo controls in clinical trails, a 25–35% improvement of patients' symptoms was documented. This placebo response occurred in patients with mild to severe symptoms and did not depend on the size of the study. The assessment of left ventricular (LV) function can be determined by several methods, including noninvasive echocardiography, radionuclide ventriculography, or invasive pulmonary artery balloon-floatation catheterization. These methods measure the patient's response to therapy or the natural progression of the patient's heart failure.[45]

Noninvasive measurements of LV ejection fraction vary, especially when the ventricular function is poor and the interval between tests is 3–6 months. Packer[45] found that when a 5% increase in ejection fraction was used to determine a beneficial response to a new drug, 20–30% of patients showed improvement while receiving placebo therapy. Overall, changes in noninvasive measures of LV function have not been shown to correlate closely with observed changes in the clinical status of patients with CHF. Most vasodilator and inotropic drugs can produce clinical benefit without a change in LV ejection fraction. Conversely, LV ejection fraction may increase significantly in patients who have heart failure and worsening clinical status.[45]

When invasive catheterization is used to evaluate the efficacy of a new drug, interpretation must be done carefully because spontaneous fluctuations in hemodynamic variables occur in the absence of drug therapy. To avoid the attribution of spontaneous variability to drug therapy, postdrug effects should be assessed at fixed times and threshold values should eliminate changes produced by spontaneous variability. Another factor that can mimic a beneficial drug response, by favorably affecting hemodynamic measurements, is measurement performed immediately after catheterization of the right side of the heart or after ingestion of a meal. After intravascular instrumentation, systemic vasoconstriction occurs and resolves after 12–24 hours. When predrug measurements are done during the postcatheterization period, any subsequent measurements will show beneficial effects because the original measurements were taken in the vasoconstricted state. Comparative data must be acquired after the postcatheterization vasoconstricted state has resolved.[45]

In the past, one of the most common tests to evaluate drug efficacy for heart failure was the exercise tolerance test. An increased duration of exercise tolerance represents a benefit of therapy. However, this increased duration is also recorded during placebo therapy and possibly results from the familiarity of the patient with the test, as in the learning phenomenon described earlier in this article for antianginal therapy; and, the increased willingness of the physician to encourage the patient to exercise to exhaustion. Placebo response to repeated exercise tolerance testing can result in an increase in duration of 90–120 seconds, when only one or two

baseline measurements are done. This response can be reduced to 10–30 seconds, when 3–10 baseline measurements are performed. Another interesting finding was that the magnitude of the placebo response was directly proportional to the number of investigators in the study! Attempts to eliminate the placebo response, including the use of gas exchange measurements during exercise tolerance testing, have failed.[45]

Because all methods used to measure the efficacy of a treatment for heart failure include placebo effects, studies must include controls for placebos to prove the efficacy of a new drug therapy. Statistical analysis of placebo-controlled studies must compare results between groups for statistical significance. 'Between groups' refers to comparison of the change in one group, such as one receiving a new drug therapy, with the change in another group, such one receiving as a placebo.[45]

In 1992, Archer and Leier[46] reported that placebo therapy for 8 weeks in 15 patients with CHF resulted in a mean improvement in exercise duration of 81 seconds, to 30% above baseline. This result was statistically significant compared with the 12-second improvement in the nine patients in the nonplacebo control group. There were no statistically significant differences between the placebo and nonplacebo groups at baseline or at week 8 of treatment by between-group statistical analysis. Echocardiography showed no significant improvement in either group and no significant differences between the two groups at baseline or during the treatment period. To prove the existence of and to quantitate the therapeutic power of placebo treatment in CHF, all studies were performed by the same principal investigator with identical study methods and conditions, and all patients were familiarized similarly with the treadmill testing procedure before baseline measurements. Also, the study used a well-matched, nonplacebo control group and this illustrated the spontaneous variability of CHF.[46]

## *Placebo Effect in Hypertension*

Some studies of the placebo response in patients with hypertension have shown a lowering of blood pressure,[47–52] but others have not.[53–57] In a Medical Research Council study, when active treatment was compared with placebo therapy (given to patients with mild hypertension for several months) similar results were produced in the two groups, an initial decrease in blood pressure followed by stabilization.[47] Of historical note is a study by Goldring et al.[58] published in 1956. These authors fabricated a sham therapeutic 'electron gun' designed to be as 'dramatic as possible, but without any known physiologic action other than a psychogenic one.' Initial exposure to 'the gun' lasted 1–3 minutes and was increased to 5 minutes three times daily. The investigators noticed substantially decreased blood pressure during therapy compared with pretherapy. In six of nine hospitalized patients there was a systolic/diastolic blood pressure reduction of 39/28 mmHg.

An important factor to consider is the method used to measure blood pressure. With the use of standard sphygmomanometry, blood pressure initially decreases. In

other studies of BP, 24-hour intraarterial pressure measurements and circadian curves did not show a decrease in blood pressure or heart rate during placebo therapy; however, Intraarterial blood pressure measurements at home were lower than measurements at the hospital. The circadian curves from intraarterial ambulatory blood pressure monitoring were reproducible on separate days, several weeks apart.[59]

Similar to 24-hour invasive intraarterial monitoring, 24-hour noninvasive automatic ambulatory blood pressure also is apparently devoid of a placebo effect. In one study, on initial application of the blood pressure device, a small reduction in ambulatory blood pressure values in the first 8 hours occurred with placebo therapy. This effect, however, did not change the mean 24-hour value. The home monitoring values were lower than the office measurements. Heart rate also was measured, with no variance in either setting. The office measurement of blood pressure was lower after 4 weeks of placebo therapy, but the 24-hour blood pressure measurement was not.[60] This study confirmed the absence of a placebo effect in 24-hour noninvasive ambulatory blood pressure monitoring, as suggested by several specific studies on large numbers of patients.[61,62] The 24-hour monitoring was measured by the noninvasive automatic Spacelabs 5300 device (Spacelabs, Redmond, Washington).[63] Another important factor in 24-hour noninvasive monitoring is that the intervals of measurement were <60 minutes.[64]

In a study on the influence of observer's expectation on the placebo effect in blood pressure measurements, 100 patients were observed for a 2-week single-blind period and for a 2-week double-blind period.[65] During this time, the patients' blood pressures were measured by two methods: a 30-minute recording with an automatic oscillometric device and a standard sphygomomanometric measurement performed by a physician. All patients were seen in the same examining room and seen by the same physician and their blood pressure monitored by the same automatic oscillometric device. The results during the single-blind period showed a slight but statistically significant decrease in diastolic blood pressure detected by the automatic oscillometric device and no decrease measured by the physician. During the double-blind period, there was no additional decline in diastolic blood pressure measured by the oscillometric device, but the physician measured significant decreases in systolic and diastolic blood pressures. Overall, the blood pressures measured by the automatic oscillometric device, in the absence of the physician, were lower than those measured by the physician. However, there was significant correlation between the two methods.

Although there was a placebo effect in the measurement of blood pressure in the Systolic Hypertension in the Elderly Program,[66,67] it was not as significant as the reduction in blood pressure produced by active therapy in patients ≥60 years of age who had isolated systolic hypertension.

As was true with angina studies, questions have been raised about the safety of placebo control studies in hypertension. As a result, two recent publications have addressed this issue.[35,68] Al-Khatib et al. performed a systematic review of the safety of placebo controls in short-term trials.[68] In their meta-analysis, they combined the data for death, stroke, MI, and CHF from 25 randomized trials. Each

study was relatively small (n = 20–734) but the combined sample size was 6,409. They found a difference between the two treatment groups and at the worst there were no more than 6/10,000 difference between placebo and active therapy. Lipicky et al. reviewed all original case report forms for deaths and dropouts were reviewed from al anti-hypertensive drug trials submitted to the FDA (as an NDA) between 1973 and 2001.[35] The population at risk was 86,137 randomized patients; 64,438 randomized to experimental drug, and 21,699 to placebo. Of the 9,636 dropouts more were from the placebo group (RR 1.33 for placebo), the majority of the dropouts were, as expected, due to treatment failures, and the patients were simply returned to their original therapies with no sequelae. When serious adverse events were compared (death, irreversible harm, etc.) there were no differences between placebo and experimental drug.

## *Placebo Effect in Arrhythmia*

Spontaneous variability in the natural history of disease or in its signs or symptoms is another reason that placebo controls are necessary. In a study of ventricular arrhythmias, Michelson and Morganroth[69] found marked spontaneous variability of complex ventricular arrhythmias such as ventricular tachycardia and couplets. These investigators observed 20 patients for 4-day periods of continuous electrographic monitoring. They recommended that when evaluating therapeutic agents, a comparison of one 24-hour control period to four 24-hour test periods must show a 41% reduction in the mean hourly frequency of ventricular tachycardia and a 50% reduction in the mean hourly frequency of couplets to demonstrate statistically significant therapeutic efficacy. They also suggested that individual patient data not be pooled to detect trends because individual variability was so great.

In a study by Morganroth et al.[70] an algorithm to differentiate spontaneous variability from proarrhythmia in patients with benign or potentially lethal ventricular arrhythmias was provided. Two or more Holter tracings were examined from each of 495 patients during placebo therapy. The algorithm defined proarrhythmia as a greater than threefold increase in the frequency of ventricular premature complexes (VPCs) when the baseline frequency of ventricular premature complexes VPCs/hour and a >10-fold increase when the frequency was <100 VPCs/hour. The false-positive rate was 1% when this algorithm was used.

The Cardiac Arrhythmia Suppression Trial[71,72] (CAST) evaluated the effect of antiarrhythmic therapy in patients with asymptomatic or mildly symptomatic ventricular arrhythmia. Response to drug therapy was determined by a ≥80% reduction in ventricular premature depolarizations or a ≥90% reduction in runs of unsustained ventricular tachycardia as measured by 24-hour Holter monitoring 4–10 days after initiation of pharmacologic treatment, a response previously considered to be an important surrogate measure of antiarrhythmic drug efficacy. One thousand four hundred fifty-five patients were assigned to drug regimens, and ambulatory electrocardiographic (Holter) recording screened for arrhythmias. The CAST Data and

Safety Monitoring Board recommended that encainide and flecainide therapy be discontinued because of the increased number of deaths from arrhythmia, cardiac arrest, or any cause compared with placebo treatment. The CAST investigators[71] conclusion emphasized the need for more placebo-controlled clinical trials of antiarrhythmic drugs with a mortality end point.

## *Relation of Treatment Adherence to Survival in Patients with or Without History of Myocardial Infarction*

An important consideration in determining study results is adherence to therapy and the presumption that any differences in adherence rates would be equal in the active versus the placebo treatment groups. The Coronary Drug Project Research Group[73] planned to evaluate the efficacy and safety of several lipid-influencing drugs in the long-term treatment of coronary heart disease. This randomized, double-blind, placebo-controlled, multicenter clinical trial found no significant difference in the 5-year mortality of 1,103 men treated with the fibric acid derivative clofibrate compared with 2,789 men given placebo. However, subjects showing good adherence (patients taking ≥80% of the protocol drug) had lower mortality than did subjects with low adherence in both the clofibrate group and the placebo group.[73]

A similar association between adherence and mortality was found in patients after myocardial infarction in the Beta-Blocker Heart Attack Trial[74] data. This phenomenon was extended to women after myocardial infarction. On analysis of the trial data for 505 women randomly assigned to β-blocker therapy or placebo therapy, there was a 2.5-fold to twofold increase in mortality within the first 2 years in patients taking <75% of their prescribed medication. Adherence among men and women was similar, at about 90%. However, the cause of the increased survival resulting from good adherence is not known. There is speculation that good adherence reflects a favorable psychological profile – a personal ability to make lifestyle adjustments that limit disease progression. Alternatively, adherence may be associated with other advantageous health practices or social circumstances not measured. Another possible explanation is that improved health status may facilitate good adherence.[75]

The Lipid Research Clinics Coronary Primary Prevention Trial[76] did not find a correlation between compliance and mortality. These investigators randomly assigned 3,806 asymptomatic hypercholesterolemic men to receive cholestyramine or placebo. The main effects of the drug compared with placebo on cholesterol level and death or nonfatal myocardial infarction were analyzed over a 7-year period. In the group receiving active drug, a relation between compliance and outcome existed, mediated by a lowering of cholesterol level. However, no effect of compliance on cholesterol level or outcome was observed in the placebo group.[76,77]

The Physicians' Health Study included a randomized fashion 22,000 United States male physicians 40–84 years old who were free of myocardial infarction and cerebral vascular disease.[78] This study analyzed the benefit of differing fre-

quencies of aspirin consumption on the prevention of myocardial infarction. In addition, the study identified factors associated with adherence and analyzed the relation of adherence with cardiovascular outcomes in the placebo group. Analysis showed an average compliance of 80% in the aspirin and placebo groups during the 60 months of follow-up.[78] Adherence during that trial was associated with several baseline characteristics in both the aspirin and placebo groups as follows. Trial participants with poor adherence (<50% compliance with pill consumption), relative to those with good adherence, were more likely to be younger than 50 years at randomization, to smoke cigarettes, to be overweight, not to exercise regularly, to have a parental history of myocardial infarction, and to have angina. These associations were statistically significant. In a multivariate logistic regression model, cigarette smoking, excess weight, and angina remained significant predictors of poor compliance. The strongest predictor of adherence during the trial was adherence during the run-in period. Baseline characteristics with little relation to adherence included regular alcohol consumption and a history of diabetes and hypertension.[78] Using intention-to-treat analysis, the aspirin group had a 41% lower risk of myocardial infarction compared with the placebo group. On subgroup analysis, participants reporting excellent (≥95%) adherence in the aspirin group had a significant, 51% reduction in the risk of first myocardial infarction relative to those with similar adherence in the placebo group. Lower adherence in the aspirin group was not associated with a statistically significant reduction in first myocardial infarction compared with excellent adherence in the placebo group. Excellent adherence in the aspirin group was associated with a 41% lower relative risk of myocardial infarction compared with low adherence in the aspirin group. Excellent adherence in the placebo group was not associated with a reduction in relative risk. The rate of stroke was different from that of myocardial infarction. On intention-to-treat analysis, the aspirin group had a nonsignificant, 22% increased rate of stroke compared with the placebo group. Participants with excellent adherence in the placebo group had a lower rate of strokes than participants in the aspirin or placebo groups with low (<50%) adherence. Excellent adherence in the placebo group was associated with a 29% lower risk of stroke compared with excellent adherence in the aspirin group.

Also analyzed in the above study, was the overall relation of adherence to aspirin therapy with cardiovascular risk when considered as a combined end point of all important cardiovascular events, including first fatal or nonfatal myocardial infarction or stroke or death resulting from cardiovascular disease with no previous myocardial infarction or stroke. On intention-to-treat analysis, there was an 18% decrease in the risk of all important cardiovascular events in the aspirin group compared with the placebo group. Participants with excellent adherence in the aspirin group had a 26% reduction in risk of a first major cardiovascular event compared with those with excellent adherence in the placebo group. However, participants in the aspirin group with low compliance had a 31% increased risk of a first cardiovascular event compared with those in the placebo group with excellent adherence. Within the placebo group, there was no association between level of adherence and risk of a first cardiovascular event. In the analysis of death resulting from any cause

in persons with a previous myocardial infarction or stroke, low adherence in both the aspirin group and the placebo group was associated with a fourfold increase in the risk of death. When the 91 deaths due to cardiovascular causes were studied, similar elevations in risk were found in both the placebo and aspirin groups with poor adherence compared with those in the placebo group with excellent adherence.

The Physicians' Health Study[78] found results similar to those of the Coronary Drug Project when all cause mortality and cardiovascular mortality were considered.[73] These relations remained strong when adjusted for potential confounding variables at baseline. The strong trend for higher death rates among participants with low adherence in both the aspirin and the placebo groups may be due to the tendency for subjects to decrease or discontinue study participation as their health declines to serious illness. Low adherence in the placebo group was not associated with an increased risk of acute events such as myocardial infarction. Thus placebo effects seem to vary depending on the outcome considered.

### *Miscellaneous*

Flaten conducted an experiment in which he told participants that they were receiving either a relaxant, stimulant, or an inactive agent, but in fact gave all of them the inactive agent. Patients who were told they were getting the relaxant showed reduced stress levels, while those who thought they were receiving the stimulant showed increased arousal levels. In another study, asthmatics that were told they were getting either a bronchodilator or bronchconstrictor and who actually received that particular therapy had more effective responses when the information received actually matched the drug effect.

Linde et al. evaluated the placebo effect of pacemaker implantation in 81 patients with obstructive hypertrophic cardiomyopathy.[79] The study design was a 3-month multicenter, double-blind, cross-over study. In the first study period 40 patients were assigned to inactive pacing, and were compared to 41 patients with active pacing. During inactive pacing, there was an improvement in chest pain, dyspnea, palpitations, and in the left ventricular outflow gradient. The change in the active pacing group for most parameters was greater.

## Clincial Trials and the Ethics of Using Placebo Controls

Since the 1962 amendments to the Food, Drug, and Cosmetic Act, the FDA has had to rely on the results of 'adequate and well-controlled' clinical trials to determine the efficacy of new pharmacologic therapies. Regulations govern pharmacologic testing and recognize several types of controls that may be used in clinical trials to assess the efficacy of new pharmacologic therapies. The controls include:

(1) placebo concurrent control, (2) dose-comparison concurrent control, (3) no-treatment concurrent control, (4) active-treatment concurrent control, and (5) historical control. Regulations, however, do not specify the circumstances for the use of these controls because there are various study designs that may be adequate in a given set of circumstances.[19]

There is ongoing debate concerning the ethics of using placebo controls in clinical trials of cardiac medications. The issue revolves around the administration of placebo in lieu of a proven therapy. Two articles, by Rothman and Michels[80] and Clark and Leaverton,[81] illustrate the debate.

Rothman and Michels[80] state that patients in clinical trials often receive placebo therapy instead of proven therapy for the patient's medical condition and assert that this practice is in direct violation of the Nuremberg Code and the World Medical Association's adaptation of this Code in the Declaration of Helsinki. The Nuremberg Code, a 10-point ethical code for experimentation in human beings, was formulated in response to the human experimentation atrocities that were recorded during the post-World War II trial of Nazi physicians in Nuremberg, Germany. According to Rothman and Michels,[80] violation occurs because the use of placebos as controls denies the patient and best proven therapeutic treatment. It occurs despite the establishment of regulatory agencies and institutional review boards, although these authors seem to ignore that informed consent is part of current practice, as certainly was not the case with the Nazi atrocities. However, a survey of federally funded grants found that despite the process of informed consent almost 25% of medical research subjects were unaware that they were part of a research project or that they were receiving investigational therapies. It should be noted, however, that this survey spanned 20 years, and did not include analysis for the more recent time period, when, most would agree, there has been more emphasis on informed consent.

One reason why placebo-controlled trials are approved by institutional review boards is that this type of trial is part of the FDA's general recommendation for demonstrating therapeutic efficacy before an investigational drug can be approved. That is, according to the FDA, when an investigational drug is found to be more beneficial by achieving statistical significance over placebo therapy, then therapeutic efficacy is proven.[82] As more drugs are found to be more effective than placebos in treating diseases, the inclusion of a placebo group is often questioned. However, this question ignores that in many cases drug efficacy in the past had been established by surrogate measures; and, as new and better measures of efficacy become available, additional study becomes warranted. Regarding surrogate past measures for example, the suppression of ventricular arrhythmia by antiarrhythmic therapy was later proven to be unrelated to survival; in fact, results with this therapy were worse than with placebo. Likewise, in studies of inotropic therapy for heart failure, exercise performance rather than survival was used as the measure of efficacy, and in fact a presumed efficacious therapy performed worse than placebo when survival was assessed. In the use of immediate short-acting dihydropyridine calcium antagonist therapy for the relief of symptoms of chronic stable angina pectoris, again a subject might have fared better had he or she been randomly assigned to placebo therapy.

Also important in the concept that established beneficial therapy should not necessarily prohibit the use of placebo in the evaluation of new therapies is that the natural history of a disease may change, and the effectiveness of so-called established therapies (e.g., antibiotic agents for treatment of infections) may diminish. When deciding on the use of an investigational drug in a clinical trial, the prevailing standard is that there should be enough confidence to risk exposure to a new drug, but enough doubt about the drug to risk exposure to placebo. Thus, in this situation, the use of a placebo control becomes warranted, particularly as long as other live-saving therapy is not discontinued.

The use of placebo-controlled trials may be advocated on the basis of a scientific argument. When pharmacologic therapy had been shown to be effective in previous placebo-controlled trials, conclusions made from trials without placebo controls may be misleading because the previous placebo-controlled trial becomes a historical control. These historical controls are the least reliable for demonstration of efficacy.[19] In active-controlled clinical trials, there is an assumption that the active control treatment is as effective under the new experimental conditions as it was in the previous placebo-controlled clinical trial. This assumption can result in misleading conclusions when results with an experimental therapy are found to be equivalent to those with active, proven therapy. This conclusion of equivalence can be magnified by conservative statistical methods, such as the use of the 'intent-to-treat' approach, an analysis of all randomized patients regardless of protocol deviations, and an attempt to minimize the potential for introduction of bias into the study. Concurrent placebo controls account for factors other than drug-effect differences between study groups. When instead of a placebo-control group an untreated control group is used, then blinding is lost and treatment-related bias may occur.[19,81]

Clark and Leaverton[81] and Rothman and Michels[80] agree that the use of placebo controls is ethical when there is no existing treatment to affect morbidity and mortality or survival favorably. Furthermore, there are chronic diseases for which treatment exists but does not favorably alter morbidity and mortality or survival. For example, no clinical trial has found the treatment of angina to increase a patient's survival. In contrast, treatment after a myocardial infarction with β-blocking agents has been convincingly proven to increase a patient's survival.[81]

However, Clark and Leaverton[81] disagree with Rothman and Michels[80] in asserting that for chronic disease, a placebo-controlled clinical trial of short duration is ethical because there is usually no alteration in long-term outcome for the patient. The short duration of the trial represents a small segment of the lifetime management of a chronic disease. For instance, the treatment of chronic symptomatic CHF and a low ejection fraction (<40%) with enalapril was shown to decrease mortality by 16%. This decrease in mortality was most marked in the first 24 months of follow-up, with an average follow-up period of 40 months. Therefore, only long-term compliance with pharmacologic therapy resulted in some decreased mortality. Another example of a chronic medical condition that requires long-term treatment and in which short-term placebo is probably not harmful is hypertension.[83] In some studies men and women with a history of myocardial infarction and with a ≥80%

compliance with treatment, including placebo therapy, had an increased survival. This increased survival was also described in patients in a 5-year study of the effects of lipid-influencing drugs on coronary heart disease.[73–75]

Therefore Rothman and Michels[80] and Clark and Leaverton[81] agree that a placebo should not be included in a trial when there exists a proven therapy that favorably affects morbidity and mortality, but they disagree when considering chronic cardiovascular diseases and short-term trials. Brief interruption of effective therapy has not been found to alter long-term outcome when the effective treatment is a long-term therapy. The claim that if a proven therapy exists the use of placebos in clinical trials violates the Nuremberg Code and the Declaration of Helsinki, does not account for all of the information currently available. The proven therapies for chronic CHF and hypertension are long-term therapies. The belief that patients receiving placebo are being harmed is not accurate because there is no adverse effect on morbidity and mortality or survival when proven, long-term therapy is withheld for a short duration.

A different argument for the ethical basis of using placebo controls relies on the informed consent process. Before a patient's participation in a clinical trial, the patient is asked to participate in the trial. The informed consent process includes a description of the use of placebos and other aspects of the trial. In this written agreement, the patient is responsible for notifying the physician of any medical problems and is informed of his or her right to withdraw from the study at any time, as described in the Nuremberg Code and the Declaration of Helsinki. During this disclosure, patients are presented with some new concepts and with risks and benefits to understand. On the basis of this information, a patient voluntarily decides to participate, knowing that he or she may receive a placebo or investigational medication.

However, despite physicians' efforts to inform the patient of research methods and the risks and benefits of trial participation, some patients agree to participate simply because of their trust in their physician. This situation may produce conflict between the physician-patient relationship and the physician's role as an investigator. A partial resolution of this conflict is the double-blind technique, in which neither the patient nor the physician knows which therapy a patient is receiving. This technique allows the doctor and patient to make medical decisions on the basis of clinical signs and symptoms. In addition, because of the requirement of informed consent, the decision about participation in a clinical trial is shifted to the patient rather than left solely with the physician. However, the patient's physician evaluates the suitability of the patient for a particular trial before asking the patient to participate.

For every pharmacologic therapy, there is an assumption made about patient compliance with the regimen. In clinical trials, investigators try to keep track of compliance by having patients bring their pill bottles to their appointments and counting the pills. Ultimately, the patient decides whether the beneficial effects of therapy outweigh the adverse effects. If a medication produces annoying and adverse side effects, then the patient may not continue to take the medication. Other factors affecting compliance are the number of pills taken per day or the frequency

of dosing. For instance, it is easier to take a medication once per day rather then three times per day. Furthermore, studies of patient compliance have found increased survival in patients with at least 80% rate of compliance with therapy, even when if is placebo therapy.[73–75]

All parties involved in research should be responsible for their research and accountable for its ethical. Clinical trials failing to comply with the Nuremberg Code and the Declaration of Helsinki should not be conducted and should not be accepted for publication. Yet, there is disagreement in determining which research methods are in compliance with the Nuremberg Code and Declaration of Helsinki. Scientific needs should not take precedence over ethical needs. Clinical trials need to be carefully designed to produce a high quality of trial performance. In addition, in experimentation involving human subjects, the Nuremberg Code and Declaration of Helsinki must be used as universal standards. The Declaration of Helsinki addresses the selection of appropriate controls by stating 'the benefits, risks, burdens, and effectiveness of a new method should be tested against the best current prophylactic, diagnostic, and therapeutic methods. This does not exclude the use of placebo, or of no treatment, in studies where no proven prophylactic, diagnostic, or therapeutic method exists.' Others have added that if the patient or subject is not likely to be harmed through exposure to placebo, and they can give voluntary informed consent, it is permissible to use placebo controls in some trials despite the existence of a know effective therapy.

## Conclusions

Until the mechanism of the placebo action is understood and can be controlled, a clinical trial that does not include a placebo group provides data that should be interpreted with caution. The absence of a placebo group makes it difficult to assess efficacy of a therapy. It is easy to attribute clinical improvement to a drug therapy if there is no control group. As was found with heart failure, chronic diseases have variable courses. Until the variability in chronic diseases is understood, placebo controls are needed to help explain it. In addition, because each clinical trial has a different setting and different study design within the context of the physician-patient relationship, a placebo group helps the investigator differentiate true drug effects from placebo effects.

More important than the inclusion of a placebo group is a careful study design that includes frequent review, by a data and safety monitoring board, of each patient's medical condition and trends affecting the patient's medical condition and trends affecting the patients' mortality and morbidity and survival. This monitoring is crucial to protect the study participants. To protect the participants, trials must include provisions that require a patient to be removed from a trial when the patient or doctor believes that removal is in the patient's best interest. The patient can then be treated with currently approved therapies.

Patients receiving placebo may report subjective clinical improvements, and demonstrate objective clinical improvement, for instance on exercise tolerance testing or Holter monitoring of ischemic events. Findings such as these dispel the implication that placebo therapy is the same as no therapy and may occur because many factors are involved in the physician-patient relationship such as the psychological state of the patient; the patient's expectations and conviction in the efficacy of the method of treatment' and the physician's biases, attitudes, expectations, and methods of communication.[2] An explanation of improvement in patients participating in trials is the close attention received by patients from the investigators. Baseline laboratory values are checked to ensure the safety of the patient and compliance with the study protocol. This beneficial response by the patient is called a positive placebo effect when found in control groups of patients receiving placebo therapy.[27,30,32,33,35,38,61,65,33,37,39,40,42,45,65,70]

Conversely, the condition of patients receiving placebos has also in some cases worsened. Every drug has side effects. These side effects are also found with placebo therapy and can be so great that they preclude the patient's continuation with the therapy. This phenomenon is always reported by patients in clinical trials receiving placebo.[15,36,45,65,84,85] Finally, placebos can act synergistically and antagonistically with other specific and nonspecific therapies. Therefore much is still to be discovered about the placebo effect.

## Summary

The effect of placebo on the clinical course of systemic hypertension, angina pectoris, silent myocardial ischemia, CHF, and ventricular tachyarrhythmia's has been well described. In the prevention of myocardial infarction, there appears to be a direct relation between compliance with placebo treatment and favorable clinical outcomes. The safety of short-term placebo-controlled trials has now been will documented in studies of drug treatment of angina pectoris. Although the ethical basis of performing placebo-controlled trials continues to be challenged in the evaluation of drugs for treating cardiovascular disease, as long as a life-saving treatment is not being denied it remains prudent to perform placebo-controlled studies for obtaining scientific information. The arguments for and against the use of placebo-controls is as follows.

The arguments in support of the use of placebo controls (placebo "orthodoxy") are numerous.[86] The word "orthodoxy" is from the Greek ortho ('right', 'correct') and doxa ('thought', 'teaching', 'glorification'). Orthodoxy is typically used to refer to the correct theological or doctrinal observance of religion, as determined by some overseeing body. The term did not conventionally exist with any degree of formality (in the sense in which it is now used) prior to the advent of Christianity in the Greek-speaking world, though the word does occasionally show up in ancient literature in other, somewhat similar contexts. Orthodoxy is opposed to heterodoxy ('other teaching'), heresy and schism. People who deviate from orthodoxy by

professing a doctrine considered to be false are most often called heretics. Some of the supporting arguments are that there are methodologic limitations of trials using active controls such as:

– Variable responses to drugs in some populations
– Unpredictable and small effects
– Spontaneous improvements

In addition, some believe that no drug should be approved unless it is clearly superior to placebo or no treatment, so that placebo is ethical if there is "no permanent adverse consequence" form its use; or, if there is "risk of only temporary discomfort, or if there "is no harm" consequent to its use. It should be noted that these latter two arguments are not equivalent; that is, patients may be harmed by temporary but reversible conditions, and that these criteria may in fact permit intolerable suffering. For example, in the 1990s several placebo-controlled trials of ondansetron for chemotherapy induced vomiting were performed when there were existent effective therapies (i.e. no permanent disability, but more than mere discomfort). Another example might be the use of placebo controlled trials of antidepressants, in which there might occur instances of depression-induced suicide.

Others argue for the use of active-controls (Active-control "Orthodoxy") in lieu of placebo controls. They argue that whenever an effective intervention for a condition exists, it must be used as the control group; that is, the clinically relevant question is not whether a new drug is better than nothing, but whether it is better than standard treatment. The supporters of the use of active controls point to the most recent "Declaration of Helsinki" which states; "the benefits, risks, burdens, and effectiveness of a new method should be tested against those of the most current prophylactic, diagnostic, or therapeutic methods. This does not exclude the use of placebo, or no treatment, in studies where no proven prophylactic, diagnostic or therapeutic method exists."

The problem with "Active-Control Orthodoxy" is that scientific validity constitutes a fundamental ethical protection, and that scientifically invalid research cannot be ethical no matter how safe the study participants are. Thus, the almost absolute prohibition of placebo in every case in which an effective treatment exists is too broad, and that patients exposed to placebo may be better off than the group exposed to a new intervention. These authors agree with Emmanual and Miller in support of a "middle ground" as discussed above.[86]

## References

1. Turner JA, Deyo RA, Loeser JD, Von Korff M, Fordyce WE. The importance of placebo effects in pain treatment and research. *JAMA*. May 25, 1994; 271(20):1609–1614.
2. Benson H, Epstein MD. The placebo effect. A neglected asset in the care of patients. *JAMA*. June 23, 1975; 232(12):1225–1227.
3. Melmon K, Morrelli H, Hoffman B, Mierenberg D, eds. Melmon and Morrelli's Clinical Pharmacology: Basic Principles in Therapeutics. 3rd ed. New York: McGraw-Hill; 1992; pp. 896.

4. Packer M, Medina N, Yushak M. Hemodynamic changes mimicking a vasodilator drug response in the absence of drug therapy after right heart catheterization in patients with chronic heart failure. *Circulation*. Apr 1985; 71(4):761–766.
5. Chalmers TC. Prophylactic treatment of Wilson's disease. *N Engl J Med*. Apr 18, 1968; 278(16):910–911.
6. Garrison FH. *History of Medicine*. 4th ed. Philadelphia, PA: Saunders; 1929.
7. Rana JS, Mannam A, Donnell-Fink L, Gervino EV, Sellke FW, Laham RJ. Longevity of the placebo effect in the therapeutic angiogenesis and laser myocardial revascularization trials in patients with coronary heart disease. *Am J Cardiol*. June 15, 2005; 95(12):1456–1459.
8. Randolph E, ed. *Stedman's Medical Dictionary*. Baltimore, MD: Lippincott Williams & Wilkins; 1990.
9. White L, Tursky B, Schwartz G. *Placebo: Theory, Research, and Mechanisms*. New York: Guilford Press; 1985.
10. Shapiro AK. Factors contributing to the placebo effect: their implications for psychotherapy. *AM J Pschother*. 1961; 18:73–88.
11. Byerly H. Explaining and exploiting placebo effects. *Perspect Biol Med*. Spring 1976; 19(3):423–436.
12. Lind JA. *A treatise of the scurvy*. Edinburgh: Edinburgh University Press; 1753.
13. Hill AB. The clinical trial. *Br Med Bull*. 1951; 7(4):278–282.
14. Beecher HK. The powerful placebo. *J Am Med Assoc*. Dec 24, 1955; 159(17):1602–1606.
15. Lasagna L, Mosteller F, Von Felsinger JM, Beecher HK. A study of the placebo response. *Am J Med*. June 1954; 16(6):770–779.
16. Wolf S, Pinsky RH. Effects of placebo administration and occurrence of toxic reactions. *J Am Med Assoc*. May 22, 1954; 155(4):339–341.
17. Davis JM. Don't let placebos fool you. *Postgrad Med*. Sept 15, 1990; 88(4):21–24.
18. Nies A, Spielberg S. Principles of therapeutics. In: Hardman JG, Limbird LE, eds. *Goodman and Gilman's The Pharmacological Basis of Therapeutics*. 9th ed. New York: McGraw-Hill; 1996.
19. Makuch RW, Johnson MF. Dilemmas in the use of active control groups in clinical research. *IRB*. Jan–Feb 1989; 11(1):1–5.
20. Galton F. Regression towards mediocrity in hereditary stature. *J Anthropol Inst*. 1886; 15:246–263.
21. Ederer F. Serum cholesterol changes: effects of diet and regression toward the mean. *J Chronic Dis*. May 1972; 25(5):277–289.
22. Davis CE. The effect of regression to the mean in epidemiologic and clinical studies. *Am J Epidemiol*. Nov 1976; 104(5):493–498.
23. The National Diet-Heart Study Final Report. *Circulation*. Mar 1968; 37(3 Suppl):I1–428.
24. Yudkin PL, Stratton IM. How to deal with regression to the mean in intervention studies. *Lancet*. Jan 27, 1996; 347(8996):241–243.
25. Asmar R, Safar M, Queneau P. Evaluation of the placebo effect and reproducibility of blood pressure measurement in hypertension. *Am J Hypertens*. June 2001; 14(6 Pt 1):546–552.
26. Oh VMS. Magic or medicine? Clinical pharmacological basis of placebo medication. *Ann Acad Med (Singapore)*. 1991; 20:31–37.
27. Kelly JP. Anatomical organization of the nervous system. In: Kandel ER, Schwartz JH, Jessel TM, eds. *Principles of Neural Science*. 3rd ed. New York: Elsevier; 1991; pp. 276–292.
28. Voudouris NJ, Peck CL, Coleman G. The role of conditioning and verbal expectancy in the placebo response. *Pain*. Oct 1990; 43(1):121–128.
29. Levine JD, Gordon NC, Bornstein JC, Fields HL. Role of pain in placebo analgesia. *Proc Natl Acad Sci USA*. July 1979; 76(7):3528–3531.
30. Hersh EV, Ochs H, Quinn P, MacAfee K, Cooper SA, Barasch A. Narcotic receptor blockade and its effect on the analgesic response to placebo and ibuprofen after oral surgery. *Oral Surg Oral Med Oral Pathol*. May 1993; 75(5):539–546.
31. Kojo I. The mechanism of the psychophysiological effects of placebo. *Med Hypotheses*. Dec 1988; 27(4):261–264.

32. Egbert LD, Battit GE, Welch CE, Bartlett MK. Reduction of postoperative pain by encouragement and instruction of patients. A study of doctor-patient rapport. *N Engl J Med*. Apr 16, 1964; 270:825–827.
33. Amsterdam EA, Wolfson S, Gorlin R. New aspects of the placebo response in angina pectoris. *Am J Cardiol*. Sept 1969; 24(3):305–306.
34. Glasser SP, Clark PI, Lipicky RJ, Hubbard JM, Yusuf S. Exposing patients with chronic, stable, exertional angina to placebo periods in drug trials. *JAMA*. Mar 27, 1991; 265(12):1550–1554.
35. Lipicky R, DeFelice A, Gordon M, et al. Placebo in Hypertension Adverse Reaction Meta-Analysis(PHARM). *Circulation*. 2003; 17(Supplement):IV–452.
36. Boissel JP, Philippon AM, Gauthier E, Schbath J, Destors JM. Time course of long-term placebo therapy effects in angina pectoris. *Eur Heart J*. Dec 1986; 7(12):1030–1036.
37. McGraw BF, Hemberger JA, Smith AL, Schroeder JS. Variability of exercise performance during long-term placebo treatment. *Clin Pharmacol Ther*. Sept 1981; 30(3):321–327.
38. Acute and chronic antianginal efficacy of continuous twenty-four-hour application of transdermal nitroglycerin. Steering committee, transdermal nitroglycerin cooperative study. *Am J Cardiol*. Nov 15, 1991; 68(13):1263–1273.
39. Beecher HK. Surgery as placebo. A quantitative study of bias. *JAMA*. July 1, 1961; 176:1102–1107.
40. Diamond EG, Kittle CF, Crockett JE. Evaluation of internal mammary artery ligation and sham procedures in angina pectoris. *Circulation*. 1958; 18:712–713.
41. Diamond EG, Kittle CF, Crockett JE. Comparison of internal mammary artery ligation and sham operation for angina pectoris. *Am J Cardiol*. 1960; 5:484–486.
42. Cobb LA. Evaluation of internal mammary artery ligation by double-blind technic. *N Engl J Med*. 1989; 260:1115–1118.
43. Carver JR, Samuels F. Sham therapy in coronary artery disease and atherosclerosis. *Pract. Cardiol*. 1988; 14:81–86.
44. van Rij AM, Solomon C, Packer SG, Hopkins WG. Chelation therapy for intermittent claudication. A double-blind, randomized, controlled trial. *Circulation*. Sept 1994; 90(3): 1194–1199.
45. Packer M. The placebo effect in heart failure. *Am Heart J*. Dec 1990; 120(6 Pt 2): 1579–1582.
46. Archer TP, Leier CV. Placebo treatment in congestive heart failure. *Cardiology*. 1992; 81(2–3):125–133.
47. Randomised controlled trial of treatment for mild hypertension: design and pilot trial. Report of medical research council working party on mild to moderate hypertension. *Br Med J*. June 4, 1977; 1(6074):1437–1440.
48. Gould BA, Mann S, Davies AB, Altman DG, Raftery EB. Does placebo lower blood-pressure? *Lancet*. Dec 19–26, 1981; 2(8260–8261):1377–1381.
49. Martin MA, Phillips CA, Smith AJ. Acebutolol in hypertension–double-blind trial against placebo. *Br J Clin Pharmacol*. Oct 1978; 6(4):351–356.
50. Moutsos SE, Sapira JD, Scheib ET, Shapiro AP. An analysis of the placebo effect in hospitalized hypertensive patients. *Clin Pharmacol Ther*. Sept–Oct 1967; 8(5):676–683.
51. Myers MG, Lewis GR, Steiner J, Dollery CT. Atenolol in essential hypertension. *Clin Pharmacol Ther*. May 1976; 19(5 Pt 1):502–507.
52. Pugsley DJ, Nassim M, Armstrong BK, Beilin L. A controlled trial of labetalol (Trandate), propranolol and placebo in the management of mild to moderate hypertension. *Br J Clin Pharmacol*. Jan 1979; 7(1):63–68.
53. A DOUBLE blind control study of antihypertensive agents. I. Comparative effectiveness of reserpine, reserpine and hydralazine, and three ganglionic blocking agents, chlorisondamine, mecamyamine, and pentolinium tartrate. *Arch Intern Med*. July 1960; 106:81–96.
54. Effects of treatment on morbidity in hypertension: results in patients with diastolic blood pressures averaging 115 through 119 mmHg by Veterans Administration cooperative study group on antihypertensive agents. *JAMA*. 1967; 202:116–122.

55. Effects of treatment on morbidity in hypertension. II. Results in patients with diastolic blood pressure averaging 90 through 114 mm Hg. *JAMA*. Aug 17, 1970; 213(7):1143–1152.
56. Hansson L, Aberg H, Karlberg BE, Westerlund A. Controlled study of atenolol in treatment of hypertension. *Br Med J.* May 17, 1975; 2(5967):367–370.
57. Wilkinson PR, Raftery EB. A comparative trial of clonidine, propranolol and placebo in the treatment of moderate hypertension. *Br J Clin Pharmacol*. June 1977; 4(3):289–294.
58. Chasis H, Goldring W, Schreiner GE, Smith HW. Reassurance in the management of benign hypertensive disease. *Circulation*. Aug 1956; 14(2):260–264.
59. Raftery EB, Gould BA. The effect of placebo on indirect and direct blood pressure measurements. *J Hypertens Suppl*. Dec 1990; 8(6):S93–100.
60. Mutti E, Trazzi S, Omboni S, Parati G, Mancia G. Effect of placebo on 24-h non-invasive ambulatory blood pressure. *J Hypertens*. Apr 1991; 9(4):361–364.
61. Dupont AG, Van der Niepen P, Six RO. Placebo does not lower ambulatory blood pressure. *Br J Clin Pharmacol*. July 1987; 24(1):106–109.
62. O'Brien E, Cox JP, O'Malley K. Ambulatory blood pressure measurement in the evaluation of blood pressure lowering drugs. *J Hypertens*. Apr 1989; 7(4):243–247.
63. Casadei R, Parati G, Pomidossi G, et al. 24-hour blood pressure monitoring: evaluation of Spacelabs 5300 monitor by comparison with intra-arterial blood pressure recording in ambulant subjects. *J Hypertens*. Oct 1988; 6(10):797–803.
64. Portaluppi F, Strozzi C, degli Uberti E, et al. Does placebo lower blood pressure in hypertensive patients? A noninvasive chronobiological study. *Jpn Heart J*. Mar 1988; 29(2):189–197.
65. Sassano P, Chatellier G, Corvol P, Menard J. Influence of observer's expectation on the placebo effect in blood pressure trials. *Curr Ther Res*. 1987; 41:304–312.
66. Prevention of stroke by antihypertensive drug treatment in older persons with isolated systolic hypertension. Final results of the Systolic Hypertension in the Elderly Program (SHEP). SHEP Cooperative Research Group. *JAMA*. June 26, 1991; 265(24):3255–3264.
67. Davis BR, Wittes J, Pressel S, et al. Statistical considerations in monitoring the Systolic Hypertension in the Elderly Program (SHEP). *Control Clin Trials*. Oct 1993; 14(5):350–361.
68. Al-Khatib SM, Califf RM, Hasselblad V, Alexander JH, McCrory DC, Sugarman J. Medicine. Placebo-controls in short-term clinical trials of hypertension. *Science*. June 15, 2001; 292(5524):2013–2015.
69. Michelson EL, Morganroth J. Spontaneous variability of complex ventricular arrhythmias detected by long-term electrocardiographic recording. *Circulation*. Apr 1980; 61(4):690–695.
70. Morganroth J, Borland M, Chao G. Application of a frequency definition of ventricular proarrhythmia. *Am J Cardiol*. Jan 1, 1987; 59(1):97–99.
71. Preliminary report: effect of encainide and flecainide on mortality in a randomized trial of arrhythmia suppression after myocardial infarction. The Cardiac Arrhythmia Suppression Trial (CAST) Investigators. *N Engl J Med*. Aug 10, 1989; 321(6):406–412.
72. Capone RJ, Pawitan Y, el-Sherif N, et al. Events in the cardiac arrhythmia suppression trial: baseline predictors of mortality in placebo-treated patients. *J Am Coll Cardiol*. Nov 15, 1991; 18(6):1434–1438.
73. Influence of adherence to treatment and response of cholesterol on mortality in the coronary drug project. *N Engl J Med*. Oct 30, 1980; 303(18):1038–1041.
74. Horwitz RI, Viscoli CM, Berkman L, et al. Treatment adherence and risk of death after a myocardial infarction. *Lancet*. Sept 1, 1990; 336(8714):542–545.
75. Gallagher EJ, Viscoli CM, Horwitz RI. The relationship of treatment adherence to the risk of death after myocardial infarction in women. *JAMA*. Aug 11, 1993; 270(6):742–744.
76. The Lipid Research Clinics Coronary Primary Prevention Trial results. II. The relationship of reduction in incidence of coronary heart disease to cholesterol lowering. *JAMA*. Jan 20, 1984; 251(3):365–374.
77. Sackett DL, Haynes RB, Gibson E, Johnson A. The problem of compliance with antihypertensive therapy. *Pract. Cardiol*. 1976; 2:35–39.

78. Glynn RJ, Buring JE, Manson JE, LaMotte F, Hennekens CH. Adherence to aspirin in the prevention of myocardial infarction. The Physicians' Health Study. *Arch Intern Med*. Dec 12–26, 1994; 154(23):2649–2657.
79. Linde C, Gadler F, Kappenberger L, Ryden L. Placebo effect of pacemaker implantation in obstructive hypertrophic cardiomyopathy. PIC Study Group. Pacing In Cardiomyopathy. *Am J Cardiol*. Mar 15, 1999; 83(6):903–907.
80. Rothman KJ, Michels KB. The continuing unethical use of placebo controls. *N Engl J Med*. Aug 11, 1994; 331(6):394–398.
81. Clark PI, Leaverton PE. Scientific and ethical issues in the use of placebo controls in clinical trials. *Annu Rev Public Health*. 1994; 15:19–38.
82. Schechter C. The use of placebo controls. *N Engl J Med*. Jan 5 1995; 332(1):60; author reply 62.
83. Alderman MH. Blood pressure management: individualized treatment based on absolute risk and the potential for benefit. *Ann Intern Med*. Aug 15, 1993; 119(4):329–335.
84. Drici MD, Raybaud F, De Lunardo C, Iacono P, Gustovic P. Influence of the behaviour pattern on the nocebo response of healthy volunteers. *Br J Clin Pharmacol*. Feb 1995; 39(2):204–206.
85. Roberts AH. The powerful placebo revisited: magnitude of nonspecific effects. *Mind/Body Medicine*. 1995; 1:35–43.
86. Emanuel E, Miller F. The ethics of placebo-controlled trials – a middle ground. NEJM. 2001; 345:915–918.

# Chapter 8
# Recruitment and Retention

**Stephen P. Glasser**

**Abstract** Nothing is more important to a clinical research study than recruiting and then retaining subjects in a study. In addition, losses to follow-up and destroy a study. This chapter will address such issues as to why people participate in clinical research, what strategies can be employed to recruit and then retain subjects in a study, issues involved with minority recruitment, and HIPAA; and, will include some real examples chosen to highlight the retention of potential drop-outs.

## Introduction

Nothing is more important to a clinical research study than recruiting and then retaining subjects in a study. However, many studies fail to recruit their planned number of participants. Studies that recruit too few patients might miss clinically important effects. The scale of the problem has been assessed; and, in one study that consisted of a multicenter cohort trial, only 37% of the trials met their planned recruitment goals.[1] Easterbrook et al. also studied the issue of recruitment in 487 research protocols submitted to the Central Oxford Research Ethics Committee, and found that 10 never started, and 16 reported abandonment of the study, because of recruitment difficulties.[2]

In addition, losses to follow-up can destroy a study (see Chapter 3). Recruitment and retention has become even more important in today's environment of scandals, IRB constraints, HIPPA, the ethics of reimbursing study participants, and skyrocketing costs. For example, one researcher demonstrated how not to do research as outlined in a USA Today article in 2000.[3] According to that newspaper article, the researcher put untoward recruitment pressure on the staff, ignored other co-morbid diseases in the recruited subjects, performed multiple simultaneous studies in the same subjects, fabricated and destroyed records, and ultimately blamed the study coordinators for all the errors found during an audit.

## Recruitment Process

The recruitment process involves a number of important steps and the trial enrollment process is being increasingly addressed because of its importance to the studies ultimate generalizability.[4] An outline of the enrollment process is shown in

S.P. Glasser (ed.), *Essentials of Clinical Research*,

Fig. 8.1 which also introduces a number of variables and definitions which should be considered and perhaps reported in large trials.[5] Recall that sampling (see Chapter 3) is perhaps one of the most important considerations in clinical research. Also recall, that the target population is the population of potentially eligible subjects, and how this is defined can have significant impact on the studies generalizability. From the target population, a smaller number are actually recruited and then enrolled (eligibility fraction and enrollment fraction). The product of these two fractions represents the proportion of potential participants who are actually enrolled in the study (recruitment fraction).[5] An example of the use of these various fractions is taken from a study, in which we found that as defined according to standards recommended by Morton et al.,[6] the response rate (percent agreeing to be interviewed among known eligible candidates contacted n = 57,253) plus an adjustment for the estimated proportion eligible among those of unknown eligibility (n = 25,581) was 44.7% (36,983/82,834). The cooperation rate (the proportion of known eligible participants who agreed to be interviewed) was 64.6% (36,983/57,253) (unpublished data). This helps the reader to understand how representative the study population is. However, as Halpern has pointed out, "although more thorough reporting would certainly help identify trials with potentially limited generalizability, it would not help clinicians apply trial results to individual patients."[7] Latter author also points out that data on patients who chose not to participate would be important. There follows an interesting discussion of the pros and cons addressing this entire issue which is important for the interested reader. Beyond the importance of generalizability, details of the recruitment process might also demonstrate obstacles to the recruitment process.

## Failures in Recruitment

There are a number of reasons for failure of the recruitment process including: ethical considerations, delayed start-up, inadequate planning, insufficient effort & staff, and over-optimistic expectations. In addition recruitment to NIH studies adds an additional burden as the NIH expects adequate numbers of women, minorities and children (when appropriate) to be recruited into studies that they fund. The ethical considerations regarding recruitment are increasingly becoming an issue. Every researcher faces a critical weighing of the balance between informing patients about the benefits and risks of participating in a trial, against unacceptable encouragement to participate. IRBs are exerting increasingly more rigorous control about what is appropriate and inappropriate in this regard. This has been the subject of debate in the United Kingdom as well, and is particularly acute due to the fact that the National Health Service requires that investigators adhere to strict regulations.[8] In the UK (and to some extent in the USA), ethicists are insisting that researchers can only approach subjects who have responded positively to letters from their general practitioners or hospital clinician-the so-called 'opt in' approach. That is, under the opt-in system a subject is responsible for contacting their doctor and letting them know it is okay for a researcher to contact them. In an opt-out system,

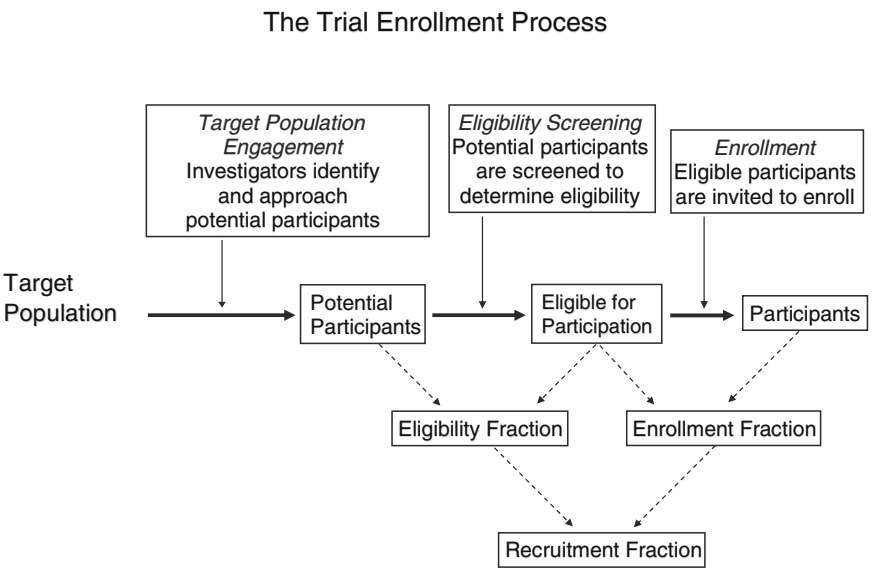



**Fig. 8.1** The trial enrollment process

the initial letter to the patient will explain that a researcher will be contacting them unless they tell their doctor that they wish not to be contacted. Hewison and Haines have argued that the public needs to be included in the debate about what is in the subject's best interests, before an ethicist can make a unilateral decision.[8] Hewison and Haines feel that 'research ethics requirements are compromising the scientific quality of health research', and that 'opt-in systems of recruitment are likely to increase response bias and reduce response rates'.[8] There is little data on the subject of opt-in vs. opt-out systems in regards to the concerns expressed above, but the potential for bias and reduced recruitment is certainly hard to argue.

The above considerations just apply to the method of contacting potential subjects. Other issues regarding recruitment are also becoming increasingly important as more studies (particularly Industry supported studies) matriculated out of academic centers and into private offices, where the investigator and staff might not have experience in clinical research. This privatization of clinical research began in the 1990s predominantly due to the inefficiencies of working with academia, including protracted contractual and budget negotiations, bureaucratic and slow moving IRBs, and higher costs.[9] Today, only 1/3 of all industry-funded clinical trials are placed within academic centers. Now, as NIH funding is dwindling and other federal funding changes are occurring, many within academic centers are again viewing the potential of industry supported research studies.

## Differences in Dealing with Clinical Trial Patients

There are differences in the handling of clinical patients in contrast to research subjects (although arguably this could be challenged). But at the least, research subjects are seen more frequently, have more testing performed, missed appointments result in protocol deviations, and patients lost to follow-up can erode the studies validity. In addition many research subjects are in studies not necessarily for their own health, but to help others. Thus, the expense of travel to the site, the expense of parking, less than helpful staff, and waiting to be seen may be even less tolerable than it is to clinical patients. Thus, the provisions for on site child care, a single contact person, flexible appointment times, telephone and letter reminders, and the provision of study calendars with study appointment dates are important for the continuity of follow-up. In addition, at a minimum, payment for travel and time (payments to research subjects are a controversial issue) need to be considered, but not at such a high rate that the payment becomes coercive.[10] The use of financial compensation as a recruiting tool in research is quite controversial, with one major concern that such compensation will unduly influence potential subjects to enroll in a study, and perhaps even to falsify information to be eligible.[11] In addition, financial incentives would likely result in an overrepresentation of the poor in clinical trials. Also, these days, it is important that study sites maintain records of patients that might be potential candidates for trials as funding agencies are more frequently asking for documentation that there will be adequate numbers of subjects available for study. Inflating the potential for recruitment is never wise as the modified cliché goes, 'you are only as good as your last study'. Failure to adequately recruit for a study will significantly hamper efforts to be competitive for the next trial. Demonstrating to funding agencies that there is adequate staff, and facilities, and maintaining records of prior studies is also key.

## Why People Participate in Clinical Research

There have not been many studies delving into why subjects participate in clinical research. In a study by Jenkins et al. the reasons for participating and declining to participate were evaluated (see Table 8.1).[12] This was also evaluated by West et al. and both found that a high proportion of participants enrolled in studies to help others.[13] West et al. performed a cross sectional survey with a questionnaire mailed to 836 participants and a response rate of 31% (n = 259). Responses were open-ended and an *a priori* category scale was used and evaluated by two research co-coordinators with a 10% random sample assessed by a third independent party in order to determine inter-reader reliability (Table 8.2).

Few studies have attempted to objectively quantify the effects of commonly used strategies aimed at improving recruitment and retention in research studies. One

**Table 8.1** Why do patients participate in clinical research?

| Advantages | Disadvantages |
|---|---|
| – Close observation 50% | – Inconvenience 31% |
| – Self knowledge 40% | – ADEs 10% |
| – Helping others 32% | – Sx worsening 9% |
| – New Rxs 27% | – Blinding 7% |
| – Free care 25% | – Rx withdrawal 1.6% |
| – Improve their Dz 23% | – |

**Table 8.2** Why do people participate?[12]

| Top reasons for accepting trial entry | n (%) |
|---|---|
| n = 138 (nine missing cases) | |
| I feel that others with my illness will benefit from results of the trial | 34 (23.1) |
| I trusted the doctor treating me | 31 (21.1) |
| I thought the trial offered the best treatment available | 24 (16.3) |
| Top reasons for declining trial entry n = 47 (four missing cases) | n (%) |
| I trusted the doctor treating me | 11 (21.6) |
| The idea of randomization worried me | 10 (19.6) |
| I wanted the doctor to choose my treatment rather than be randomized by computer | 9 (17.6) |

that did evaluate five common strategies, assessed the effect of notifying potential participants prior to being approached; providing potential research subjects with additional information about the study; changes in the consent process; changes in the study design (such as not having a placebo arm); and; the use of incentives. The author's conclusions were that it is not possible to predict the effect of most interventions on recruitment.[14]

## Types of Recruitment

There are a number of additional considerations one has to make for site recruitment and retention. For example, before the study starts consideration as to how the subjects will be recruited (i.e. from a data-base, colleague referral, advertising-print, television, radio, etc.) and once the study starts there needs to be weekly targets established and reports generated, The nature of the recruitment population also needs to be considered, For example, Gilliss et al studied the one-year attrition rates by the way in which they were recruited and ethnicity.[15] They found that responses to and subsequent 1 year attrition rates, differed between broadcast media, printed matter, face-to face recruitment, direct referral, and the use of the Internet; and, differed between American, African American and Mexican

American. For example, the response to broadcast media resulted in 58%, 62% and 68% being either not eligible or refusing to participate; and, attrition rates were 13%, 17% and 10% comparing American, Mexican American and African Americans respectively. In contrast, face to face recuritment resulted in lower refusal (21%, 28%, and 27%) and attrition rates 4%, 4%, and 16%).

## Minority Recruitment

Due to the increased interest in enrolling minorities into clinical research trails, this has become a subject of greater emphasis. This is because ethnicity-specific analyses have been generally inadequate for determining subgroup effects. In 1993, the National Institutes of Health Revitalization Act mandated minority inclusion in RCTs, and defined underrepresented minorities as African Americans, Latinos, and American Indians. Subsequently, review criteria have formally required minority recruitment plans or scientific justification for their exclusion. Yancey et al.,[16] evaluated the literature on minority recruitment and retention and identified 10 major themes or factors that emerged as influencing minority recruitment. Further, they noted that if addressed appropriately it: facilitated recruitment: attitudes towards perceptions of the scientific and medical community; sampling approach; study design; disease specific knowledge and perceptions of prospective participants; prospective participants psychosocial issues; community involvement; study incentives and logistics; sociodemographic characteristics of prospective participants; participant beliefs such as religiosity; and cultural adaptations or targeting. In general, most of the barriers to minority participation were similar for non-minorities except for the greater mistrust by African Americans toward participation (particularly into interventional trials), likely as a result of past problems such as the Tuskegee Syphilis Study.[17] Some of the authors conclusions based upon their review of the literature included: mass mailing is effective; population-based sampling is unlikely to produce sufficient numbers of ethnic minorities; community involvement is critical; survey response rates are likely to be improved by telephone follow-up.

## HIPPA

A final word about recruitment relates to HIPAA (the Health Insurance Portability and Accountability Act). Issued in 1996, the impetus of HIPPA was to protect patient privacy. However, many have redefined HIPAA as ‘How Is it Possible to Accomplish Anything’. As it applies to research subjects it is particularly confusing. The term protected health information (PHI) includes what physicians and other health care professionals typically regard as a patient’s personal health information.

PHI also includes identifiable health information about subjects of clinical research gathered by a researcher. Irrespective of HIPAA, the safeguarding of a patient's personal medical records should go without saying; and, failure of this aforementioned safeguarding has resulted in problems for some researchers. As it affects patient recruitment, however, HIPAA is problematic in that the researcher's ability to contact patients for a research study, particularly patients of another health care provider, becomes more problematic. In addition, in clinical research, the investigator is often in a dual role as it regards a patient-that of a treating physician and that of a researcher. Long standing ethical rules apply to researchers, but in regard to HIPAA, a researcher is not a 'covered entity' (defined as belonging to a health plan, health care clearinghouse, or health care provider that transmits health information electronically). However, complicating the issue is when the researcher is also a health care provider, or employees or other workforce members are a covered entity. The role and scope of HIPAA, as it applies to clinical research is beyond the intention (or comprehension) of this author and therefore will not be further discussed.

## Summary

During my over 35 years of clinical research experience I have developed a number of strategies aimed at retaining participants, and some examples are outlined below.

- A participant in a 4 year outcome trial discontinued study drug over 1 year ago (during the second year of the trial) due to vague complaints of weakness and fatigue, however, the participant did agree to continue to attend study visits. At one of the follow up visits, we asked the participant if they would be willing to try the study drug again, and in so doing were able to re-establish the participant in the trial. Recall that based upon the intention-to-treat principle (see Chapter 3) they would have been counted as having received their assigned therapy anyway, and in terms of the outcome it is still better that they received the therapy for 3 of the 4 years, than for less than that.
- Another participant reported a loss of interest in the study and stopped his study drug. Upon questioning it was determined that he had read newspaper articles about recent studies involved with the study drug you are testing, and felt there is nothing to gain from continuing in the study. We explained how this study differs from those reported in the newspaper, using a fact based approach, and the subject was willing to participate once again.
- A participant following up on the advice of his primary care doctor (PCP) decided he would like to know what study drug he was receiving when the PCP noted a BP of 150/90 mmHg. Further, the PCP had convinced the patient to discontinue blinded study therapy. You receive a call form the patient stating they

no longer wish to participate in the study. One way of preventing this in the first place is to involve the patients PCP from the beginning. However, in this case, the patient had transferred to a new PCP and had not informed us. As a result, we called the PCP and communicated the importance of the study and assured the PCP that better BP control is expected and that we would be carefully monitoring his BP.

In summary, a frank open discussion with the patient as to what happened and why he/she wants to discontinue is important, as well as preserving rapport with the patient and their PCP is the key to subject retention. It is also critical that the principal investigator (PI) maintain frequent contact (and thereby solidify rapport) with the patient, given that in many studies the study coordinator and not the PI may see the patient on most occasions. I remember asking one study coordinator if they knew the definition of PI and the immediate response was 'yes-practically invisible!'

## References

1. Charlson ME, Horwitz RI. Applying results of randomised trials to clinical practice: impact of losses before randomisation. *Br Med J (Clin Res Ed)*. Nov 10, 1984; 289(6454):1281–1284.
2. Easterbrook PJ, Matthews DR. Fate of research studies. *J R Soc Med*. Feb 1992; 85(2):71–76.
3. A case study in how not to conduct a clinical trial. *USA Today*, 2000.
4. Wright JR, Bouma S, Dayes I, et al. The importance of reporting patient recruitment details in phase III trials. *J Clin Oncol*. Feb 20, 2006; 24(6):843–845.
5. Gross CP, Mallory R, Heiat A, Krumholz HM. Reporting the recruitment process in clinical trials: who are these patients and how did they get there? *Ann Intern Med*. July 2, 2002; 137(1):10–16.
6. Morton LM, Cahill J, Hartge P. Reporting participation in epidemiologic studies: a survey of practice. *Am J Epidemiol*. Feb 1, 2006; 163(3):197–203.
7. Halpern SD. Reporting enrollment in clinical trials. *Ann Intern Med*. Dec 17, 2002; 137(12):1007–1008; author reply 1007–1008.
8. Hewison J, Haines A. Overcoming barriers to recruitment in health research. *BMJ*. Aug 5, 2006; 333(7562):300–302.
9. Getz K. Industry trials poised to win back academia after parting ways in the late 90s. *Appl Clin Trials*. Apr 1, 2007; 2007.
10. Giuffrida A, Torgerson DJ. Should we pay the patient? Review of financial incentives to enhance patient compliance. *BMJ*. Sept 20, 1997; 315(7110):703–707.
11. Dunn LB, Gordon NE. Improving informed consent and enhancing recruitment for research by understanding economic behavior. *JAMA*. Feb 2, 2005; 293(5):609–612.
12. Jenkins V, Fallowfield L. Reasons for accepting or declining to participate in randomized clinical trials for cancer therapy. *Br J Cancer*. June 2000; 82(11):1783–1788.
13. Hawkins C, West T, Ferzola N, Preismeyer C, Arnett D, Glasser S. Why do patients participate in clinical research? *Associates of Clinical Pharmacology 1993 Annual Meeting*; 1993.
14. Mapstone J, Elbourne DR, Roberts I. Strategies to improve recruitment in research studies; 2002.

15. Gilliss C, Lee K, Gutierrez Y, et al. Recruitment and Retention of Healthy Minority Women into Community-Based Longitudinal Research. *J Womens Health Gender-Based Med*. 2001; 10:77–85.
16. Yancy AK, Ortega AN, Kumanyika SK. Effective recruitment and retention of minority research participants. *Annu Rev Public Health*. 2006; 27:1–28.
17. Tuskegee Syphilis Study. http://www.tuskegee.edu/Global/Story.asp?s = 1207598.

# Chapter 9
# Data Safety and Monitoring Boards (DSMBs)

**Stephen P. Glasser and O. Dale Williams**

**Abstract** Data Safety and Monitoring Boards were introduced as a mechanism for monitoring interim data in clinical trials as a way to ensure the safety of participating subjects. Procedures for and experience with DSMBs has expanded considerably over recent years and they are now required by the NIH for almost any interventional and for some observational trials. A DSMB's primary role is to evaluate adverse events and to determine the relationship of the adverse event to the therapy (or device). Interim analyses and early termination of studies are two aspects of DSMBs that are particularly difficult. This chapter will discuss the role of DSMBs and address the aforementioned issues.

Data Safety and Monitoring Boards (DSMBs), which have various names including Data Safety and Monitoring Committees and Data Monitoring Committees, were born in 1967, a result of a NIH sponsored task force report known as the Greenberg Report.[1] Initially the responsibilities now assigned to a DSMB were a component of those of a Policy Advisory Board. From this emerged a subcommittee to focus on monitoring clinical trial safety and efficacy. More specifically, the DSMB was introduced as a mechanism for monitoring interim data in clinical trials as a way to ensure the safety of participating subjects. Procedures for and experience with DSMBs has expanded considerably over recent years, and several key publications relevant to their operations are now available.[2–4] In general, NIH now requires DSMBs for all clinical trials (including some Phase I and II trials) trials, and recently added device trials to this mandate.[5] DSMBs are now an established interface between good science and good social values. For example, NHLBI at the NIH[6] requires the following:

– For Phase III clinical trials, a Data and Safety Monitoring Board (DSMB) is required. This can be a DSMB convened by the NHLBI, or by the local institution, depending on the study, the level of risk and the funding mechanism.
– For a Phase II trial, a DSMB may be established depending on the study, but in most cases a DSMB appointed by the funded institution may suffice.
– For a Phase I trial, monitoring by the PI and the local IRB usually suffices. However, a novel drug, device or therapy with a high or unknown safety profile may require a DSMB.

S.P. Glasser (ed.), *Essentials of Clinical Research*,

– For an Observational Study, a Monitoring Board (OSMB) may be established for large or complex observational studies. This would be determined on a case-by-case basis by NHLBI.

NHLBI also requires that each DSMB operate under an approved Charter,[3] with the expectation that this Charter will delineate the primary function of the DSMB being to ensure patient safety, as well as to ensure that patients are adequately informed of the risk in study participation. The DSMB Charter requires a formal manual of operations (MOOP) and the DSMB and sponsor must agree on all the terms set forth in the MOOP (this is sometimes referred to a Data Safety and Monitoring Plan – DSMP). This includes such things as the DSMBs responsibility, its membership, meeting format and frequency, specifics about the decision making process, report preparation, whether the DSMB will be blinded or not to the treatment arms, and the statistical guidelines that will be utilized by the DSMB to determine whether early termination of the study is warranted. In addition, DSMBs assure that the rate of enrollment is sufficient to achieve adequate numbers of outcomes, develop guidelines for early study termination, and to evaluate the overall quality of the study to include accuracy, timeliness, data flow, etc.

The DSMB is charged with assessing the progress of clinical trials and to recommend whether the trail should continue, be modified, or discontinued. More specifically, the DSMB approves the protocol, has face-to-face meetings, usually every 6 months (these are supplemented with conference calls), they may have subgroup meetings for special topics and are on call for crises; and, DSMBs review interim analyses (generally required for NIH studies). An interim analysis is one performed prior to study completion.

Members of DSMBs are to be expert in areas relevant to the study, approved by the sponsor, and without conflicts of interest relative to the study to be monitored. The members should not be affiliated with the sponsor, and should be independent from any direct involvement in the performance of the clinical trial. Members of the DSMB tend to include clinical experts, statisticians, ethicists, and community representatives. Thus, the DSMB's overarching objectives are to ensure the safety of participants, oversee the validity of data, and to provide a mechanism for the early termination of studies.

## Early Study Termination

A DSMB's primary role is to evaluate adverse events and to determine the relationship of the adverse event to the therapy (or device). As the DSMB periodically reviews study results, evaluates the treatments for excess adverse effects, determines whether basic trial assumptions remain valid, and judges whether the overall integrity of the study remains acceptable, it ultimately makes recommendations to the funding agency. For NHLBI sponsored studies, this recommendation goes directly to the Institute Director, who has the responsibility to accept, reject, or modify DSMB recommendations.

The issue of terminating a study early or of altering the course of its conduct is a critically important decision. On the surface, when to terminate a study can be obvious such as when the benefit of intervention is so clear that continuing the study would be inappropriate, or conversely when harm is clearly evident. That is, a study should be stopped early if bad is happening, good is happening, or nothing is happening (and the prospects are poor that if the study continues there will be benefit). Finally, the DSMB can recommend early termination if there are forces external to the study that warrant its early discontinuation (e.g. a new life saving modality is approved during the course of the study). More frequently, however, it is difficult to sort out this balance of risk vs. benefit, and judgment is the key. As Williams so aptly put it 'stopping too early is to soon and too late is not soon enough i.e. no one is going to be happy in either case.[7] That is, stopping to early leads to results that may not be judged to be convincing, might impact other ongoing studies, or that endpoints not yet adjudicated may affect the results of the study. Finally, the DSMB must be concerned with the potential for operational chaos that may ensue, and unnecessary costs may be realized when a study is terminated ahead of schedule; however, stopping a trial to late may be harmful to patients. In addition one may keep society waiting for potentially beneficial therapy.

Another dilemma faced by early stopping is if the trial is in its beginning phases, and an excess of adverse events, for example, is already suggested. The DSMB is then faced with the question of whether this observation is just a 'blip' which will not be evident for the rest of the trial and stopping at this point would hamper if not cause cessation of a drugs development. If, on the other hand it is in the middle of the trial and efficacy, for example, is not yet fully demonstrated, the question faced by the DSMB is whether there can still be a turnaround such that the results will show benefit. Finally, if it is late in a trial, and there has been no severe harm demonstrated, but apparent efficacy is minimal, the question is whether it is worth the cost and confusion to stop the trial before completion, when it will be ending shortly anyway. In the final analysis, it is generally agreed that the recommendation to modify or terminate a trial should not solely be based upon statistical grounds. Rather, 'no statistical decision, rule, or procedure can take the place of the well reasoned consideration of all aspects of the data by a group of concerned, competent, and experienced persons with a wide range of scientific backgrounds and points of view'.[8]

## Interim Analysis

Interim analyses may occur under two general circumstances; based on accrual – e.g. one interim analysis after half of the patients have been evaluated for efficacy (this to some degree depends on the observation time for efficacy), or based on time – e.g. annual reviews. Often the information fraction (the number of events that have occurred compared to those expected) provides a frame of reference.[9]

Stopping rules for studies, as mentioned before, are dependent upon both known science and judgment. For example, in a superiority trial if one treatment arm is demonstrating 'unequivocal' benefit over another, study termination can be recommended. However there are problems with this approach. For example one of the study arms may show superiority at the end of year 1, but may then lose any advantage over the ensuing time period of the study. A way of dealing with this at the time of the interim analysis is to assess futility. That is, given the recruitment goals, if at the time of the study, an interim analysis suggests that there is no demonstrable difference in the treatment arms, and one can show that it would be unlikely (futile) that with the remaining patients a difference is likely to occur, the study can be stopped. [10]

Finally, an issue with interim analysis is the multiple comparisons problem (see Chapter 3). In other words, with each interim analysis, sometimes called a 'look,' one 'spends' some of the overall alpha level. Alpha is, of course, the overall significance level (usually 0.05). Statisticians have developed various rules to deal with the multiple comparison problem that arises with interim data analyses, One approach is to stop trials early only when there is overwhelming evidence of efficacy. Peto has suggested that overwhelming evidence is when $p < 0.001$ for a test that focuses on the primary outcome.[11] Perhaps the simplest method to understand is the Bonferroini adjustment, which divides the overall alpha level by the number of tests to be conducted to obtain the alpha level to use for each test. As discussed in Hulley et al.[12] that means that if five tests are done and the overall alpha is 0.05, then for statistical significance for stopping a $p < 0.01$ or less, for each individual test is needed. This latter approach is typically conservative in that the actual overall alpha level may be well below 0.05, unnecessarily so.

There often are compelling reasons to make it more difficult to cross a stopping boundary early rather than later in the study. Hence, another approach is to have a stopping boundary which changes as the trial moves closer to its predetermined end, with higher boundaries earlier and lower ones later. The rationale is that early in the study, the number of endpoints is typically quite small and thus trends are subject to high variability. This makes it more likely that there is a more extreme difference between the treatment arms early that will settle down later. Also, as the end of the trial nears, a less stringent p value is required to indicate significance, since the results are less likely to change (there will be fewer additional patients added to the trial compared to earlier in its conduct).[9]

The three most commonly used methods for setting boundaries, sometime referred to as group sequential boundaries, as a frame of reference for early termination decisions are: the Haybittle-Peto,[11,13] Pocock,[14] and Obrien-Fleming[15] methods. The Haybittle-Peto and Pocock methods do not provide higher boundaries early in the study, whereas, the Obrien-Fleming, and the Lan-Demets[9] modification do. Fig. 9.1 shows how these compare to each other for situations whereby five looks are expected for the trial.[16,17] Thus, interim safety reports pose well recognized statistical problems related to the multiplicity of statistical tests conducted on the accumulating set of data. The basic problem is well known and is referred to as 'sampling to a foregone conclusion,[16] or the problem of repeated significance tests.[18,19]

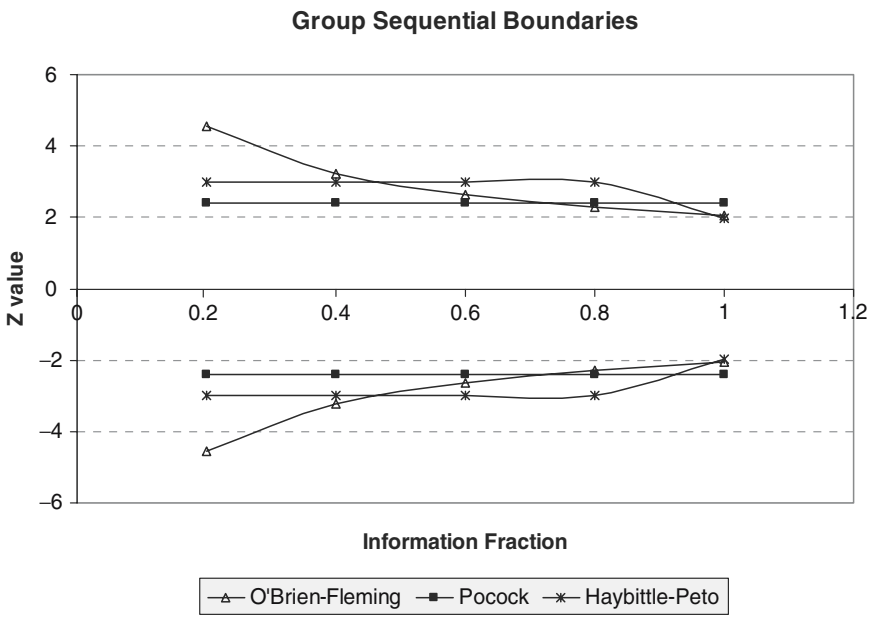



**Fig. 9.1** Group sequential boundaries

Califf et al. outlined the questions that should be asked by a DSMB before altering a trial.[20] Califf et al. also point out that statistical rules are not absolute but provide guidance only. Some additional issues discussed in their review include the role of the DSMB in event-driven (i.e. the trial continues until the pre-specified number of events has been accrued) vs. fixed-sample, fixed-duration trials; how the DSMB should approach equivalence vs. noninferiority trials; the role of a Bayesian approach to DSMB concerns; the use of asymmetric vs. symmetric boundaries (the threshold for declaring that a trial should be stopped should be less stringent for safety issues than it is when a therapy shows a positive result); and, perhaps most importantly, the philosophy of early stopping-that is, where does the committees primary ethical obligation lie, and what degree of certainty is required before a trial can be altered.[19]

The disadvantages of stopping a trial early are numerous. These include the fact that the trial might have been terminated on a random 'high'; the reduction in the credibility of the trial when the number of patients studied will have been less than planned; and, the greater imprecision regarding the outcome of interest as the smaller sample size will have resulted in wider confidence limits. Montori et al. performed a systematic review of randomized trials stopped early as a result of their demonstrating benefit at the time of an interim analysis.[21] They noted that 'taking the point estimate of the treatment effect at face value will be misleading if the decision to stop the trial resulted from catching the apparent benefit of treatment at a 'random high'. When this latter situation occurs, data from future trials will yield a more conservative estimate of the treatment effect, the so called regression to the

truth effect.' Montori's findings suggested that there were an increasing number of RCTs reported to have stopped early for benefit; and, that the early stopping occurred with (on average) 64% of the planned sample having been entered. More importantly, they concluded that information regarding the decision to stop early was inadequately reported, and overall such studies demonstrated an implausibly large treatment effect and they then suggest that the results of such studies should be viewed with skepticism.[20] One example of early stopping for harm was the ILLUMINATE trial which was terminated early by the DSMB because the trial drug, Pfizer's *torcetrapib*, had more events than placebo.[22] The questions addressed but not able to be answered were: Why this occurred? Was it the drug itself or the dose of the drug? What was the mechanism of adverse events, etc.

Finally, as discussed in Chapter 3 the duration of the clinical trial can be an important consideration in the DSMB deliberations. Some studies may show early lack of benefit and have a delayed beneficial effect. The DSMB should carefully follow the curves elucidating the study endpoints in order to identify the potential for a delayed effect. Thus, the DSMB might not be only involved in early stopping, but might suggest a longer duration of the RCT than originally planned.

## Observational Study and Monitoring Boards (OSMBs)

OSMBs are a more recent development and are not as often necessary as they are with interventional trials.[23] Thus, a main question is when should an OSMB be established? It is the policy of the NHLBI to establish OSMBs for Institute-sponsored observational studies and registries when an independent group is needed to evaluate the data on an ongoing basis to ensure participant safety and/or study integrity. The decision to establish an OSMB is made by the relevant Division Director with the concurrence of the Director, NHLBI. As a general rule, the NHLBI appoints OSMBs for:

- All large, long-term Institute-initiated and selected investigator-initiated observational studies, whether multiple or single center in nature and
- Selected smaller Institute-initiated and selected investigator-initiated observational studies or registries to help assure the integrity of the study by closely monitoring data acquisition for comprehensiveness, accuracy, and timeliness; and monitoring other concerns such as participant confidentiality

The role of the OSMBs is similar to that of the DSMB, that is to monitor study progress and to make recommendations regarding appropriate protocol and operational changes. They also address safety issues such as those involving radiation exposure or other possible risks associated procedures or measurements that are study components. Decisions to modify the protocol or change study operations in a major way may have substantial effects upon the ultimate interpretation of the study or affect the study's funding. Thus, OSMBs play an essential role in assuring quality research. The principal role of the OSMB is to monitor regularly the data

from the observational study, review and assess the performance of its operations, and make recommendations, as appropriate with respect to:

- The performance of individual centers (including possible recommendations on actions to be taken regarding any center that performs unsatisfactorily)
- Issues related to participant safety and informed consent, including notification of and referral for abnormal findings
- Adequacy of study progress in terms of recruitment, quality control, data analysis and publications
- Issues pertaining to participant burden
- Impact of proposed ancillary studies and substudies on participant burden and overall achievement of the main study goals and
- Overall scientific directions of the study

Thus, the OSMB must provide a multidisciplinary and objective perspective, with expert attention to all of these factors during the course of the study, and considerable judgment.

The responsibilities of the OSMBs are summarized in a document that can be found on the NHLBI web site.[23]

## References

1. Heart Special Project Committee. Organization, Review, and Administration of Cooperative Studies. Greenberg Report: A Report from Heart Special Project Committee to the National Advisory Heart Council. *Control Clin Trials*. 1967; 9:137–148.
2. DeMets D, Furberg C, Friedman L. *Data Monitoring in Clinical Trials. A Case Studies Approach*. New York: Springer; 2006.
3. Ellenberg SS, Fleming TR, DeMets D. *Data Monitoring Committees in Clinical Trials. A Practical Perspective*. West Sussex: Wiley; 2002.
4. Friedman L, Furberg C, DeMets D. *Fundamentals of Clinical Trials*. 3rd ed. New York: Springer; 1998.
5. Further Guidance on a Data and Safety Monitoring for Phase I and Phase II Trials. http://grants.nih.gov/grants/guide/notice-files?NOT-OD-99038.html. Accessed 6/14, 2007.
6. Guidelines for NIH Intramural Investigators and Institutional Review Boards on Data and Safety Monitoring. http://ohsr.od.nih.gov/info/pdf/InfoSheet18.pdf. Accessed 6/14, 2007.
7. *Fundamentals of Clinical Research (Power Point Lecture*; 2005.
8. The Coronary Drug Project Research Group. Practical aspects of decision making in clinical trials: the coronary drug project as a case study. *Control Clin Trials*. May 1981; 1(4):363–376.
9. DeMets DL, Lan KK. Interim analysis: the alpha spending function approach. *Stat Med*. July 15–30, 1994; 13(13–14):1341–1352; discussion 1353–1346.
10. DeMets DL, Pocock SJ, Julian DG. The agonising negative trend in monitoring of clinical trials. *Lancet*. Dec 4, 1999; 354(9194):1983–1988.
11. Peto R, Pike MC, Armitage P, et al. Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I. Introduction and design. *Br J Cancer*. Dec 1976; 34(6):585–612.
12. Hulley S, Cummings S, Browner Wea. *Designing Clinical Research*. 2nd ed. Philidelphia, PA: Lippincott Williams & Wilkins; 2000.

13. Haybittle JL. Repeated assessment of results in clinical trials of cancer treatment. *Br J Radiol*. Oct 1971; 44(526):793–797.
14. Pocock SJ. When to stop a clinical trial. *BMJ*. July 25, 1992; 305(6847):235–240.
15. O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. *Biometrics*. Sept 1979; 35(3):549–556.
16. Cornfield J. Sequential trials, sequential analysis and the likelihood principle. *Am Stat*. 1966; 20:18–23.
17. Jennison C, Turnbull BW. *Group Sequential Methods with Applications to Clinical Trials*. Boca Raton, FL: Chapman & Hall; 2000.
18. Armitage P, McPherson CK, Rowe BC. Repeated significance tests on accumulating data. *J Roy St Soc A*. 1969; 132:235–244.
19. McPherson K. The problem of examining accumulating data more than once. *New Eng J Med*. 1975; 290(501–502).
20. Caiff RM, Ellenberg SS. Statistical approaches and policies for the operations of data and safety monitoring committees. *Am Heart J.* 2000; 141:301–305.
21. Montori VM, Devereaux PJ, Adhikari NK, et al. Randomized trials stopped early for benefit: a systematic review. *JAMA*. Nov 2, 2005; 294(17):2203–2209.
22. National Heart L, and Blood Institute, National Institutes of Health. Monitoring Boards for Data and Safety. http://public.nhibi.nih.gov/ocr/home/GetPolicy.aspx?id = 8. Accessed 6/14, 2007.
23. Responsibilities of OSMBs appointed by the NHLBI. National Heart, Lung and Blood Institute. National Institutes of Health. http://www.nhlbi.nih.gov/funding/policies/dsmb_inst.htm

# Chapter 10
# Meta-Analysis

**Stephen P. Glasser and Sue Duval**

**Abstract** Meta-analysis refers to methods for the systematic review of a set of individual studies (either from the aggregate data or the individual patient data) with the aim to quantitatively combine their results. This has become a popular approach to attempt to answer questions when the results from individual studies have not been definitive. This chapter will discuss meta-analyses and highlight issues that need critical assessment before the results of the meta-analysis are accepted. Some of these critical issues include: publication bias, sampling bias, and study heterogeneity.

## Introduction

Meta- is from Latin meaning among, with, or after; occurring in succession to, situated behind or beyond, more comprehensive, or transcending. This has lead some to question if meta-analysis is to analysis as metaphysics is to physics (metaphysics refers to the abstract or supernatural), as a number of article titles would attest to, such as: "is a meta-analysis science or religion?"[1]; "have meta-analyses become a tool or a weapon?"[2]; "meta-statistics: help or hinderance?"[3]; and, "have you ever meta-analysis you didn't like?"[4] 'Overviews, systematic reviews, pooled analyses, quantitative reviews and quantitative analyses are other terms that have been used synonymously with meta-analysis, but some distinguish between them. For example, pooled analyses might not necessarily use the true meta-analytic statistical methods, and quantitative reviews might similarly be different than a meta-analysis. Compared to traditional reviews, meta-analyses are often more narrowly focused, usually examine one clinical question, and necessarily have a strong quantitative component. Meta-analysis can be literature based and these are essentially, studies of studies. The majority of meta-analyses rely on published reports, however more recently, meta-analyses of individual patient data (IPD) have appeared.

The earliest meta-analysis may have been that of Karl Pearson in 1904, which he applied in an attempt to overcome the problem of reduced statistical power in studies with small sample sizes.[5] The first meta-analysis of medical treatment is probably that of Henry K Beecher on the powerful effects of placebo, published in

S.P. Glasser (ed.), *Essentials of Clinical Research*,

1955.[6] But, the term meta-analysis is credited to Gene Glass in 1976.[7] Only 4 meta-analyses could be found before 1970, 13 were published in the 1970s and fewer than 100 in the 1980s. Since the 1980s more than 5,000 meta-analyses have been published.

## Definition

Meta-analysis refers to methods for the systematic review of a set of individual studies or patients (subjects) within each study, with the aim to quantitatively combine their results. Meta-analysis has become popular for many reasons, some of which include:

- The adoption of evidence based medicine which requires that all reliable information is considered
- The desire to avoid narrative reviews which are often misleading or inconclusive
- The desire to interpret the large number of studies that may have been conducted about a specific intervention
- The desire to increase the statistical power of the results by combining many smaller sized studies

Some definitions of a meta-analysis include:

- An observational study in which the units of observation are individual trial results or the combined results of individual patients (subjects) aggregated from those trials.
- A scientific review of original studies in a specific area aimed at statistically combining the separate results into a single estimation.
- A type of literature review that is quantitative.
- A statistical analysis involving data from two or more trials of the same treatment and performed for the purpose of drawing a global conclusion concerning the safety and efficacy of that treatment.

One should view meta-analyses the same way as one views a clinical trial (unless one is performing an exploratory meta-analysis), except that most meta-analyses are retrospective. Beyond that, a meta-analysis is like a clinical trial except that the units of observation may be individual subjects or individual trial results. Thus, all the considerations given to the strengths and limitations of clinical trials should be applied to meta-analyses (e.g. a clearly stated hypothesis, a predefined protocol, considerations regarding selection bias, etc.)

The reasons one performs a meta-analysis is to ‘force’ one to review all pertinent evidence, to provide quantitative summaries, to integrate results across studies, and to provide for an overall interpretation of these studies. This allows for a more rigorous review of the literature, and it increases sample size and thereby potentially enhances statistical power. That is to say, that the primary aim of a meta-analysis is

to provide a more precise estimate of an outcome (say a medical therapy in reducing mortality or morbidity) based upon a weighted average of the results from the studies included in the meta-analysis. The concept of a 'weighted average' is an important one. In the most basic approach, the weight given to each study is the inverse of the variance of the effect; that is, on average, the smaller the variance, and the larger the study, the greater the weight one places on the results of that study. Because the results from different studies investigating different but hopefully similar variables are often measured on different scales, the dependent variable in a meta-analysis is typically some standardized measure of effect size. In addition, meta-analyses may enhance the statistical significance of subgroup analysis, and enhance the scientific credibility of certain observations.

Finally, meta-analyses may identify new research directions or help put into focus the results of a controversial study. As such, meta-analyses may resolve uncertainty when reports disagree, improve estimates of effect size, and answer questions that were not posed at the start of individual trials, but are now suggested by the trial results. Thus, when the results from several studies disagree with regard to the magnitude or direction of effect, or when sample size of individual studies are too small to detect an effect, or when a large trial is too costly and/or to time consuming to perform, a meta-analysis should be considered.

## Weaknesses

As is true for any analytical technique, meta-analyses have weaknesses. For example, they are sometimes viewed as more authoritative than is justified. After all, meta-analyses are retrospective repeat analyses of prior published data. Rather, meta-analyses should be viewed as nearly equivalent (if performed properly under rigid study design characteristics) to a large, multi-center study. In fact, meta-analyses are really studies in which the 'observations' are not under the control of the meta-investigator (because they have already been performed by the investigators of the original studies); the included studies have not been obtained through a randomized and blinded technique; and, one must assume that the original studies have certain statistical properties they may not, in fact, have. In addition, one must rely on reported rather than directly observed values only, unless an IPD meta-analysis is undertaken.

There are at least nine important considerations in performing or reading about a meta-analysis:

1. They are sometimes performed to confirm an observed trend (this is equivalent to testing before hypothesis generation)
2. Sampling problems
3. Publication bias
4. Difficulty in pooling across different study designs
5. Dissimilarities of control treatment
6. Differences in the outcome variables

7. Studies are reported in different formats with different information available
8. The issues surrounding the choice of fixed versus random modeling of effects
9. Alternative weights for analysis

1. Meta-analyses are sometimes performed to confirm observed trends (i.e. testing before hypothesis generation)

Frequently in meta-analyses, the conduct of the analysis is to confirm observed 'trends' in sets of studies; and, this is equivalent to examining data to select which tests should be performed. This is well known to introduce spurious findings. It is important to be hypothesis driven – i.e. to perform planning steps in the correct order (if possible).

In planning the meta-analysis, the same principles apply as planning any other study. That is, one forms a hypothesis, defines eligibility, collects data, tests the hypothesis, and reports the results. But, just like other hypothesis testing, the key is to avoid spurious findings by keeping these steps in the correct order, and this is frequently *NOT* the case for meta-analyses. For example, frequently the 'trend' in the data is already known; in fact, most meta-analyses are performed because of a suggestive trend. In Petitti's steps in planning a meta-analysis she suggests first addressing the objectives (i.e. state the main objectives, specify secondary objectives); perform a review; information retrieval; specify MEDLINE search criteria; and explain approaches to capture 'fugitive' reports (those not listed in MEDLINE or other search engines and therefore not readily available).[8]

2. When sampling from the universe the samples are not replicable

Repeat samples of the universe do not produce replicable populations. In identifying studies to be considered in meta-analyses one is in essence, defining the 'sampling frame' for the meta-analysis. The overall goal is to include all pertinent studies; and, several approaches are possible. With Approach 1: 'I am familiar with the literature and will include the important studies', there may be a tendency to be aware of only certain types of studies and selection will therefore be biased. With Approach 2, one uses well-defined criteria for inclusion and an objective screening tool is also utilized such as MEDLINE. But, clearly defined keywords, clearly defined years of interest, and a clear description of what you did must be included in a report. Also, the impact of the 'Search Engine' on identifying papers is often not adequately considered. Surprising to some is that there may be problems with MEDLINE screening for articles. Other searches can be done with EMBASE or PUBMED and seeking the help of a trained Biomedical Librarian may be advisable. In addition, not all journals are included in these search engines and there is dependence on keywords assigned by authors, they do not include fugitive or grey literature, government reports, book chapters, proceedings of conferences, published dissertations, etc. One of the authors once searched the web for: **Interferons in Multiple Sclerosis**. The first search yielded about 11,700 'hits' and the search took 0.27 seconds. When subsequently repeated, the search took 0.25 seconds and returned 206,000 hits.

As previously stated, the included studies in a meta-analysis have not been obtained through a randomized and blinded technique, so that selection bias becomes an issue. Selection bias occurs because studies are 'preferentially'

included and excluded and these decisions are influenced by the meta-investigators prior beliefs as well as the fact that studies are included based upon recognized 'authorities'. That is, investigator bias occurs because the investigators who conducted the individual studies included in the meta-analysis may have introduced their own bias.

Thus, it is necessary for a complete meta-analysis to go to supplemental sources for studies, such as studies of which authors are personally aware, studies referenced in articles retrieved by MEDLINE, and searches of Dissertation Abstracts etc. The biggest limitation, however, is how to search for unpublished and unreported studies. This latter issue is clearly the most challenging (impossible?), and opens the possibility for publication bias and the file-drawer problem.

3. Publication bias (and the file-drawer problem)

Publication bias is one of the major limitations of meta-analysis as it derives from the fact that for the most part, studies that are published have positive results, so that negative studies are underrepresented. Publication bias results from selective publication of studies based on the direction and magnitude of their results. The pooling of results of published studies alone leads to an overestimation of the effectiveness of the intervention, and the magnitude of this bias tends to be greater for observational studies compared to RCTs. In fact, positive studies are three times more likely to be published than negative ones and this ratio is even greater for observational studies. Thus, investigators tend not submit negative studies (this is frequently referred to as the 'file-drawer' problem), journals do not publish negative studies as readily, funding sources may discourage publication of negative studies, and Medline and other electronic data bases may be inadequate, as negative studies that do get published are published in lower impact journals some of which might not be indexed in Medline or other databases. One also has to be wary of overrepresentation of positive studies because duplicate publication can occur. The scenario resulting in publication bias goes something like this: one thinks of an exciting hypothesis, examines the possibility in existing data, if significant, publishes findings, but if non-significant loses interest and buries the results (i.e. sticks them in a file drawer). Even if one is 'honorable' and attempts to publish a non-statistically significant study, usually the editor/reviewer will bury the result for you, since negative results are difficult to get published. One then continues on to the next idea and forgets that the analysis was ever performed. The obvious result of this is that the literature is then more likely to include mostly positive findings and thereby is biased toward benefit. Publication bias is equivalent to performing a screen to select patients who only respond positively to a treatment before performing a clinical trial to examine the efficacy of that treatment.

To moderate the impact of publication bias, one attempts to obtain all published and unpublished data on the question at hand. Short of that there are other techniques, such as those that test for the presence of publication bias, methods used to estimate the impact of publication bias and adjust for it, or to limit meta-analysis to major RCTs. It should be noted that publication bias is a much greater factor in epidemiological studies than clinical trials, because it is difficult to perform a major RCT and not publish the results, while this is not nearly so true for epidemiologic studies.

As mentioned, there are ways that one can determine the likelihood that publication bias is influencing the meta-analysis. One of the simplest methods is to construct a funnel plot, which is a scatter plot of individual study effects against a measure of precision within each study. In the absence of bias, the funnel plot should depict a 'funnel' shape centered around the true overall mean which the meta-analysis is trying to estimate. This is because we expect a wider spread of effects among the smaller studies. If the funnel appears truncated, it is likely that a group of studies is missing from the analysis set. It should be kept in mind however that publication bias is but one potential reason for this 'funnel plot asymmetry', and for this reason, current practice is to consider other mechanisms for the missing studies, such as English language bias, clinical heterogeneity, and location bias to name a few.

There are a number of relatively simple quantitative methods for detecting publication bias in the literature, including the rank correlation test of Begg[9] and the regression-based test of Egger et al.[10] The Trim and Fill method[10] can be used to estimate the number of missing studies and to provide an estimate of the treatment effect after adjustment for this bias. The mechanics of this approach are displayed in Fig. 10.1a, using a meta-analysis of the effect of gangliosides and mortality from acute ischemic stroke.[11] Although the effect size is not great, the striking thing about the plot is that it appears that there are no negative effects of therapy. The question is whether that observation is true or if this is an example of publication bias where the negative studies are not represented. Figure 10.1b shows what happens when the asymmetric studies are 'trimmed' to generate a symmetric plot to allow estimation of the true pooled effect (in this example, the five rightmost studies are trimmed). These trimmed studies are then returned, along with their imputed or 'filled' symmetric counterparts. An adjusted pooled estimate and corresponding confidence interval are then calculated based on the now complete dataset (bottom panel). The authors of this method stress that the main goal of such an analysis is to allow a 'what if' approach, that is to allow sensitivity analyses to the missing studies, rather than actually finding the values of those studies per se. Another sensitivity analysis approach to estimate the impact of publication bias on the conclusions of the review is called Rosenthal's file drawer number.[12] It purports to do this by estimating the number of unpublished neutral trials that would be needed to reverse the statistical significance of a pooled estimate. This is not usually recommended by these authors and should be considered nothing more than a crude guide.

Perhaps the best approach to avoiding publication bias is to have a registry of all trials at their inception, that is before results are available, thereby eliminating the possibility that the study results would influence inclusion into the meta-analysis. After a period of apathy, this concept is taking hold.

The effect of publication bias on meta-analytical outcomes was demonstrated by Glass et al. in 1979.[13] They reported on 12 meta-analyses and in every instance where it could be determined found that the average experimental effect from studies published in journals was larger than the corresponding effect estimated from unpublished work (mostly from theses and dissertations). This accounted for

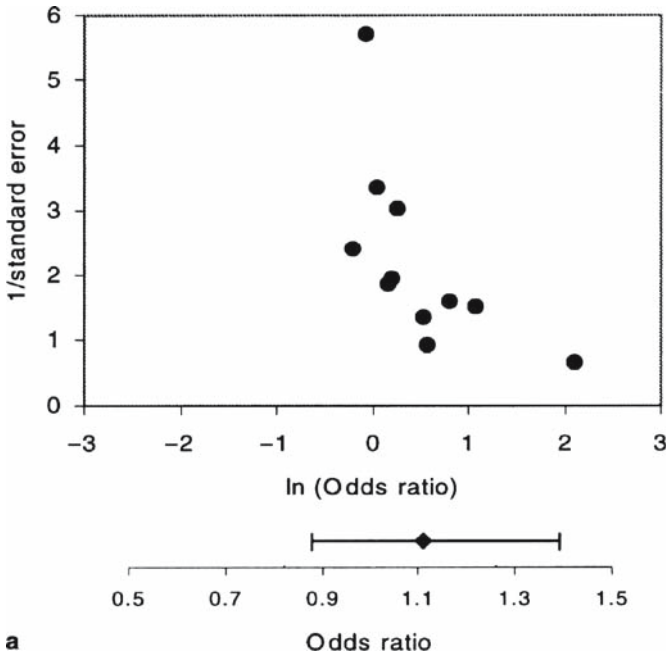



**Fig. 10.1a** A graphical portrayal of the studies included in the meta-analysis

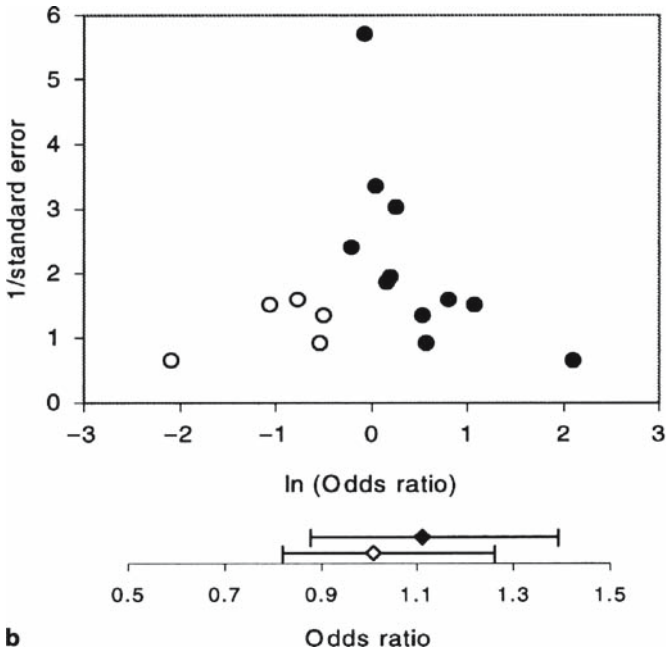



**Fig. 10.1b** Filled "presumed" negative studies

almost a 33% bias in favour of the benefit. As a result of this, some have suggested that a complete meta-analysis should include attempts to contact experts in the field as well as authors of referenced articles for access to unpublished data. Indeed guidelines for reporting meta-analyses of RCTs and observational studies have been published. More recent estimates have suggested that the effect of publication bias accounts for 5–15% in favour of benefit.

4. The difficulty in pooling across a set of individual studies and heterogeneity

One of the reasons that it is difficult to pool studies is selection bias. Selection bias occurs because studies are 'preferentially' included and excluded and these are influenced by the meta-investigators prior beliefs as well as the fact that studies are included based upon recognized 'authorities'. That is this type of investigator bias occurs because the investigators who conducted the individual studies included in the meta-analysis may have introduced their own bias. In addition, there is always a certain level of heterogeneity of study characteristics included in the meta-analysis so that as the cliché goes 'by mixing apples and oranges with an occasional lemon, ones ends up with an artificial product.' Glass argued this point rather eloquently as follows:

> *…Of course it mixes apples and oranges; in the study of fruit nothing else is sensible; comparing apples and oranges is the only endeavor worthy of true scientists; comparing apples to apples is trivial.'…*
> *The same persons arguing that no two studies should be compared unless they were studies of the 'same thing' are blithely comparing persons within studies i.e. no two things can be compared unless they are the same…but if they are the same then they are not two things.'*
> Glass went on to use the classic paradox of Theseus's ship, which set sail on a 5 year journey. After nearly 5 years, every plank had been replaced. The question then i*s*
> *'are Theseus and his men still sailing the ship that was launched 5 years earlier? What if as each plank was removed, it was taken ashore and repositioned exactly as it had been on the waters so that at the end of 5 years, there exists a ship on shore, every plank of which once stood exactly as it had been 5 years before. Is this new ship Thesues's ship.. or is it the one still sailing? The answer depends on what we understand the concept of 'same' to mean.*
> *Glass goes on to consider the problem of the persistence of personal identity when he asks the question 'how do I know that I am the same person who I was yesterday, or last year…*?

Glass also notes that probably there are no cells that are in common between the current organism called Gene Glass and the organism 40 years ago by the same name.[14]

Recall that a number of possible outcomes and interpretations of clinical trials is possible. When 1 trial is performed, the outcome may be significant, and one concludes that a treatment is beneficial, or the results may be inconclusive leading one to say that there is not convincing statistical evidence to support a treatment benefit. But when multiple trials are performed other considerations present themselves. For example, when 'most' studies are significant and in the same direction one can conclude a treatment is beneficial, but when 'most' studies are significant in different directions one might question whether there are differences in the population studied or methods that warrant further consideration. The question that may

then be raised is 'Could we learn anything by combining the studies?' It is this latter question that is the underlying basis for meta-analysis. Thus, when there is some treatment or exposure under consideration we assume that there is a 'true' treatment effect that is shared by all studies, and that the average has lower variance than the data itself. We then consider each of the individual studies as one data point in a 'mega-study' and presume that the best (most precise) estimate of this 'true' treatment effect is provided by 'averaging' across studies. But, when is it even reasonable to combine studies? The answer to this latter question is that studies must share characteristics, including similar 'experimental' treatment or exposure, similar 'standard' treatment or lack of exposure, similar follow-up protocol, outcome(s) and patient populations.

It is difficult to pool across different studies, even when there is an apparent similarity of treatments. This leads to heterogeneity when one performs any meta-analysis. The causes of study heterogeneity are numerous. Some of them are:

- Differences in inclusion/exclusion criteria of the individual studies making up the meta-analysis.
- Different control or treatment interventions (dose, timing, brand), outcome measures and definition, and different follow-up times were likely to be present in each individual study.
- The reasons for withdrawals, drop-outs, cross-over's will likely differ between individual studies, as will the baseline status of the patients and the settings for each study.
- Finally, the quality of the study design and its execution will likely differ.

Heterogeneity of the studies included in the meta-analysis can be tested. For example, Cochran's Q is a test of homogeneity that evaluates the extent to which differences among the results of individual studies are greater than one would expect if all studies were measuring the same underlying effect and the observed differences between them were due only to chance. A measure of the proportion of variation in individual study estimates that is due to heterogeneity rather than sampling error (known as $I^2$), is available and is the preferred method of describing heterogeneity.[15] This index does not rely on the number of studies, the type of outcome data or the choice of treatment effect. $I^2$ is related to Cochran's Q statistic and lies between 0% and 100%, making it useful for comparison across meta-analyses. Most reviewers consider that an $I^2$ greater than 50% indicates heterogeneity between the component studies. It is possible to weight studies based upon their methodological quality (although this is rarely done), rather sensitivity analysis to differences in study quality is more common. Sensitivity analysis describes the robustness of the results by excluding some studies such as those of poorer quality and/or smaller studies.

5. Dissimilarities in control groups

Just as important as the similarity in treatment groups is that one needs to take great caution to ensure that control groups between studies included in the meta-analysis are equivalent. For example, one study in a meta-analysis may have a statin drug vs. placebo, while another study compares a statin drug plus active risk factor

management (smoking cessation, hypertension control, etc.) compared to placebo plus active risk factor management. Certainly, one could argue that the between study control groups are not similar (clearly they are not identical), and one can only surmise the degree of bias that would be introduced by including both in the meta-analysis.

6. Heterogeneity in outcome

One might expect that the choice of an outcome to be evaluated in a meta-analysis is a simple choice. In many meta-analysis, it is not as simple as one would think,.. For example, consider a meta-analysis shown in Table 10.1. The range of effect has a risk differential from an approximately 60% decrease to 127% increase. One should reasonably ask whether the studies included in the meta-analysis should demonstrate approximately consistent results. Does it make sense to combine studies that are significant in different directions? If studies provide remarkably different estimates of treatment effect, what does an average mean? This particular scenario is used to further illustrate the use of sensitivity analyses in meta-analysis. A so-called 'influence analysis' is derived in which the meta-analysis is re-estimated after omitting each study in turn. It may be reasonable to consider excluding particular studies, or to present the results with one or two studies included and then excluded. Many analyses start out with the intention of producing quantitative syntheses, and fall short of this goal. If the reasons are well argued, this can often be the most reasonable outcome.

7. Studies are reported in different formats with different information available

Since studies are reported in different formats with different information available, the abstraction of data becomes problematic. There is no reason to anticipate that investigators will report data in a consistent manner. Frequently, differences in measures of association (odds ratio versus regression coefficients versus risk ratios, etc.) are presented in different reports which then forces the abstractor to try to reconstruct the same measure of association across studies. When abstracting information for meta-analyses, one must go through each study and attempt to collect

**Table 10.1** Meta-analysis of stroke as a result of an intervention

| | Similarities in outcomes | |
|---|---|---|
| Study | Estimate (95% CI) | |
| 1 | 1.12 (0.79–1.57) | Fatal and nonfatal first stroke |
| 2 | 1.19 (0.67–2.13) | Hospitalized F/NF stroke |
| 3 | 1.16 (0.75–1.77) | Occlusive stroke |
| 4 | 0.64 (0.06–6.52) | Fatal SAH |
| 5 | 2.27 (1.22–4.23) | Fatal and nonfatal stroke or TIA |
| 6 | 0.40 (0.01–3.07) | Fatal stroke |
| 7 | 0.97 (0.50–1.90) | Fatal and nonfatal first stroke |
| 8 | 0.63 (0.40–0.97) | Fatal occlusive disease |
| 9 | 0.97 (0.65–1.45) | Fatal and nonfatal stroke |
| *10* | *0.65 (0.45–0.95)* | Fatal and nonfatal first stroke |
| OVERALL | 0.96 (0.82–1.13) | |

the information in the same format. That is, one needs either a measure of association (e.g. an odds ratio) with some measure of dispersion (e.g. variance, standard deviation, confidence interval), or cell frequencies in 2 × 2 tables. If one wants to present a meta-analysis of subgroup outcomes, pooling may be even more problematic than pooling primary outcomes. This is because subgroups of interest are frequently not presented in a distinct manner.

The issue of consistency in reporting of studies is a particular problem for epidemiological studies where confounders are a major issue. Although confounders are easily addressed by multivariable models, there is no reason to assume that authors will use the same models in adjusting for confounders.

Another related problem is the possibility that there are multiple publications from a single population, and it is not always clear that this is happening. For example, let's say that there is a publication reporting results in 109 patients. Three years later a report from the same or similar authors reports the results of a similar intervention in 500 patients. The question is were the 500 patients all new, or did the first report of 109 patients get included in the 500 now being reported?

8. The use of random vs. fixed analysis approaches

By far, the most common approach to weighting the results in meta-analyses is to calculate a 'weighted average' of the effects (e.g. odds ratios, risk ratios) across the studies. This has the overall goal of:

– Calculating an 'weighted average' measure of effect and
– Performing a test to see if this estimated effect it is different from the null hypothesis of no effect

In considering whether to use the fixed effects or random effects modeling approach, the 'fixed' approach assumes that studies included in the meta-analysis are the only studies that could exist, while the 'random' approach assumes that the studies are a random sample of studies that may have occurred. The fixed effects model weights the studies by their 'precision'. Precision is largely driven by the sample size and reflected by the widths of the 95% confidence limits about the study-specific estimates. In general, when weights are assigned by the precision of the estimates they are proportional to (1/var(study)). This is the 'statistician's' approach, and as such is completely rational: the only problem is that it assigns a bigger weight to a big and poorly-done study than it does to a small and well-done study. Thus, a meta-analysis that includes one or two large studies is largely a report of just those studies. Random effects models estimate a between study variance and incorporates that into the model. This effectively makes the contributions of individual studies to the overall estimate more uniform. It also increases the width of the confidence interval. The random approach is likely more representative of the underlying statistical framework and the use of the 'fixed' approach can provide an underestimate of the true variance and may falsely inflate power to see effects. Most older studies have taken the 'fixed' approach, many newer studies are taking the 'random' approach since it is more representative of the 'real' world. Many meta-analysts argue that if some test of heterogeneity is significant, then one should use random effects. A reasonable approach is to present the results from both.

9. Assignment of weights

Alternative weighting schemes have been suggested such as weighting by the quality of the study,[16] with points given for whether a number of variables. The problem with weighting is that we started our meta-analysis in order to have an objective method to combine studies to provide an overall summary, and with weighting we are subjectively assigning weights to factors so that we can the objectively calculate a summary measure. However, this aforementioned weighting is but one scheme. Others reported in the literature are Jadad, and Newcaste-Ottawa which is probably currently more prevalent in the literature.[17]

## Statistical and Graphical Approaches

### *Forest Plot*

The Forest Plot is a common graphical way of portraying the data in a meta-analysis. In this plot, the point is the estimate of the effect, the size of the point is related to the size of the study, and the confidence intervals around that point estimate are displayed (for example, an odds ratio of 1 means the outcome is not affected by the intervention under study). In Fig. 10.2, a hypothetical forest plot of log hazard ratios for each study, ordered by the size of the effect within each study is shown. At the bottom, a diamond shows the combined estimate from the meta-analysis.

## Discussion

An example of some of these aforementioned principles is demonstrated in a theoretical meta-analysis of six studies. For this 'artificial' meta-analysis, only multicenter randomized trials were included, and the outcome is total mortality. Tables 10.2–10.4 present the raw data, mortality rates and odds ratios, and Fig. 10.3 presents a Forest Plot of the odds ratios with confidence intervals.

The fundamental statistical approach in meta-analysis is similar to that of an RCT in that the hypothesis is conceived to uphold the null. According to the Mantel-Haenszel-Peto method, a technique commonly used when events are sparse, a $2 \times 2$ table is constructed for each study to be included, and the observed number for the outcome of interest is computed.[18] From that computation one subtracts the expected outcome had no intervention been given. If the intervention of interest has no effect the observed minus the expected should be about zero; if the intervention is favorable (with the measure of association being the odds ratio) the OR will be greater than 1 (as will its confidence limits). The magnitude of effect can be measured in meta-analyses using a number of measures of association, such as the odds ratio (OR), relative risk (RR), risk difference (RD), and/or the number

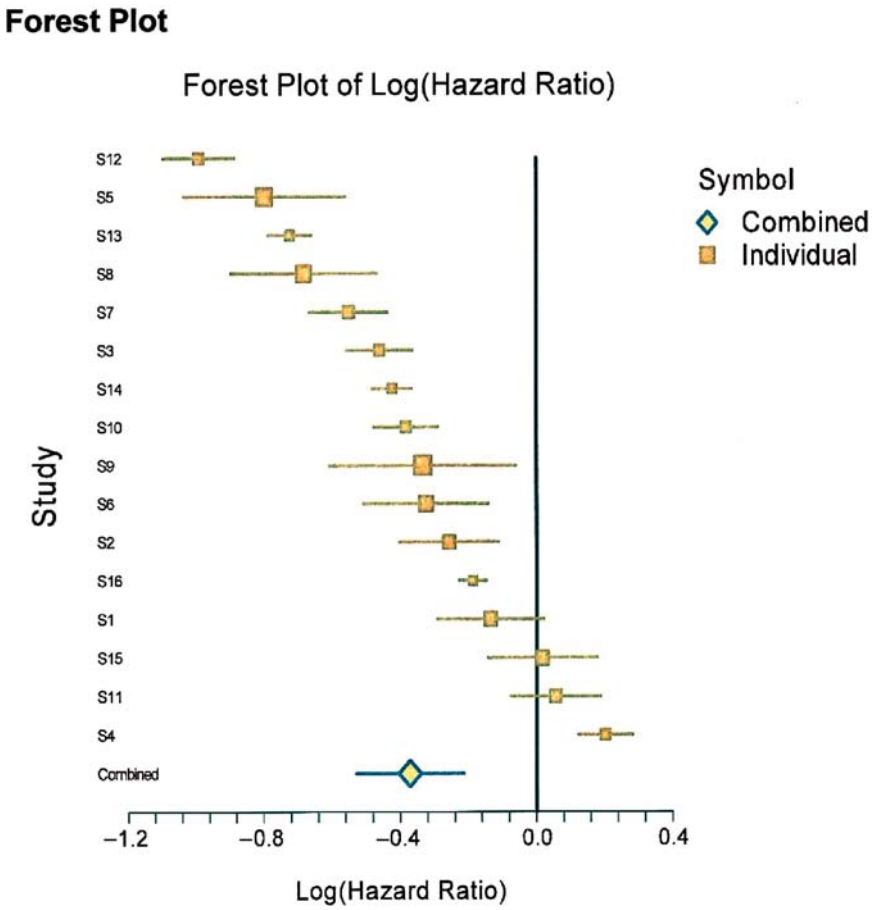



**Fig. 10.2** Forest plot

needed to treat (or harm), NNT (or NNH), to name a few The choice is, to a great degree, subjective as discussed in the Chapter 16, and briefly in number 7 above.

One limited type of meta-analysis, and a way to overcome some of the limitations of meta-analysis in general, is to preplan them with the prospective registration of studies, as has been done with some drug developments. Berlin and Colditz present the potential uses of meta-analyses (primarily of RCTs) in the approval and postmarketing evaluation of approved drugs.[19] If a sponsor of a new drug has a program to conduct a number of clinical trials, and the trials are planned as a series with prospective registration of studies at their inception, one has a focused question (drug efficacy say for lowering the total cholesterol), all patients are included (so no publication bias occurs), one then has the elements of a well planned meta analysis. In Table 10.5, Berlin and Colditz present their comparison of trials as they relate to four key elements of several types of clinical trials.

**Table 10.2** The raw data from the six studies included in the meta-analysis

| Raw data | | | | | | |
|---|---|---|---|---|---|---|
| | Treatment A | | | PLACEBO | | |
| Study | Total no. of patients | No. dead | No. alive | Total no. of patients | No. dead | No. alive |
| 1 | 615 | 49 | 566 | 624 | 67 | 557 |
| 2 | 758 | 44 | 714 | 771 | 64 | 707 |
| 3 | 317 | 27 | 290 | 309 | 32 | 277 |
| 4 | 832 | 102 | 730 | 850 | 126 | 724 |
| 5 | 810 | 85 | 725 | 406 | 52 | 354 |
| 6 | 2,267 | 246 | 2,021 | 2,257 | 219 | 2,038 |
| Total | 5,599 | 553 | 5,046 | 5,217 | 560 | 4,657 |

**Table 10.3** The individual mortality rates the six studies included in the meta-analysis

| Mortality rates | | | | | |
|---|---|---|---|---|---|
| Individual mortality rates and risk differences for the six trials | | | | | |
| | Treatment A | PLACEBO | Treatment-placebo | | |
| Study | Mortality rate | Mortality rate | Diff | SE of diff | P-value |
| 1 | 0.0797 | 0.1074 | –0.0277 | 0.0165 | 0.047 |
| 2 | 0.0580 | 0.0830 | –0.0250 | 0.0131 | 0.028 |
| 3 | 0.0852 | 0.1036 | –0.0184 | 0.0234 | 0.216 |
| 4 | 0.1226 | 0.1482 | –0.0256 | 0.0167 | 0.062 |
| 5 | 0.1049 | 0.1281 | –0.0231 | 0.0198 | 0.129 |
| 6 | 0.1085 | 0.0970 | 0.0115 | 0.0090 | 0.898 |

**Table 10.4** The data from the six studies included in the meta-analysis converted to odds ratios

| Odds ratios | | | | |
|---|---|---|---|---|
| Odds ratios for the six trials | | | | |
| Study | Log odds ratio | SE [log OR] | Odds ratio | Cl on OR |
| 1 | –0.33 | 0.197 | 0.72 | [0.49, 1.06] |
| 2 | –0.38 | 0.203 | 0.68 | [0.46, 1.02] |
| 3 | –0.22 | 0.275 | 0.81 | [0.47, 1.38] |
| 4 | –0.22 | 0.143 | 0.80 | [0.61, 1.06] |
| 5 | –0.23 | 0.188 | 0.80 | [0.55, 1.15] |
| 6 | 0.12 | 0.098 | 1.13 | [0.93, 1.37] |

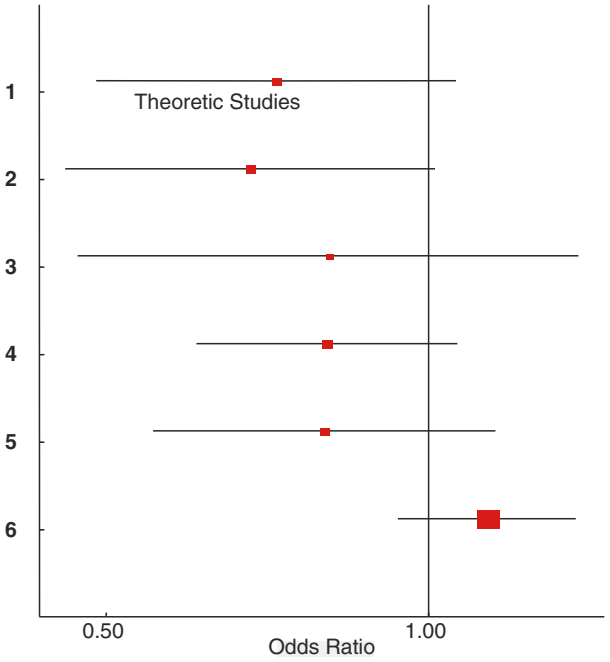



**Fig. 10.3** Example: theoretic meta-analysis

**Table 10.5** Variables relating to publication bias, generalizability, and validity with different study approaches

| | Generalizes | | | |
|---|---|---|---|---|
| Approach | Avoids publication bias | Across protocols | Across centers | Validity |
| Pre-planned meta-analysis | ++ | ++ | ++ | + |
| Large simple trial | + | − | ++ | + |
| Retrospective meta-analysis | − | ++ | ++ | + |
| 2 RCTs | − | + | + | + |
| 1 RCT | − | − | − | + |

## Conclusion

In designing a meta-analysis (or reading one in the literature) one should be certain that a number of details are included so the validity of the results can be weighed. Some of the considerations are: listing the trials included and excluded in the meta-analysis and the reasons for doing so; clearly defining the treatment assignment in each of the trials; describing the ranges of patient characteristics, diagnoses, and

treatments; and, addressing what criteria were used to decide that the studies analyzed were similar enough to be pooled.

## Evidence Based Medicine

*'It ain't so much what we don't know that gets us into trouble as what we do know that ain't so' (Will Rogers) (http://humrep.oxfordjournals.org)*

Meta-analysis and evidence based medicine (EBM) arose together as a result of the fact that the traditional way of learning (the Historic Paradigm i.e. 'evidence' is determined by the leading authorities in the field-from textbooks, review articles, seminars, and consensus conferences) was based upon the assumption that experts represented infallible and comprehensive knowledge. Numerous examples of the fallibility of that paradigm are present in the literature e.g.:

- Prenatal steroids for mothers to minimize risk of RDS
- Treatment of eclampsia with Magnesium sulfate vs. diazepam
- NTG use in suspected MI
- The use of diuretics for pre-eclampsia

In 1979 Cochrane stated 'It is surely a great criticism of our profession that we have not organised a critical summary, by specialty or sub-specialty, updated periodically, of all relevant randomized controlled trials'.[20] The idea of EBM then was to devise answerable questions, track down the best evidence to answer them, critically appraise the validity and usefulness of the evidence, apply the appraisal to clinical practice, and to evaluate one's performance after applying the evidence into practice (http://library.uchc.edu/lippub/fall99.PDF). As such, EBM called for the integration of individual clinical expertise with the best available external evidence from systematic research (i.e. meta-analysis). One definition of EBM is the conscientious, explicit judicious use of current best available evidence in making decisions about the care of individual patients with the use of RCTs, wherever possible, as the gold standard.[21] EBM also incorporates the need to encourage patterns of care that does more good than harm.

Someone said, it is not that we are reluctant to use evidence based approaches, it is that we may not agree on what the evidence is, so why shift to an EBM approach? The answers are many, but include the fact that the volume of new evidence can be overwhelming (this remains the clinicians biggest challenge), that the time necessary to keep up is not available, that up-to-date knowledge and clinical performance deteriorates with time, and that traditional CME has not been shown to improve clinical performance.

The necessary skills for EBM include the ability to precisely define a patient problem, ascertain what information is required to resolve the problem, the ability to conduct an efficient search of the literature with the selection of the most relevant articles, the ability to determine a study's validity, extract the clinical message and

apply it to the patient's problem. (http://hsa.usuhs.mil/2002ms2) There are, of course criticisms of the EBM approach. For example, some feel that evidence is never enough i.e. evidence alone can never guide our clinical actions and that there is a shortage of coherent, consistent scientific evidence. Also, the unique biological attributes of the individual patient renders the use of EBM to that individual, at best, limited. For many, the use of EBM requires that new skills be developed in an era of limited clinician time and technical resources. Finally, who is to say what the evidence is that evidence based medicine works? Some have asked," are those who do not practice EBM practicing 'non-evidence based medicine'? Karl Popper perhaps summarized this best when he noted that there are all kinds of sources of our knowledge but none has authority.[22]

EBM is perhaps a good term to the extent that it advocates more reliance on clinical research than on personal experience or intuition. But, Medicine has always been taught and practiced based on available scientific evidence and scientific interpretation and the question can be asked whether the results of a clinical trial hardly deserve the title *evidence* as questions arise about the statistical and design aspects, and data analysis, presentation, and interpretation contain many subjective elements as we have discussed in prior chapters. Thus, even if we observe consistency in the results and interpretation (a rare occurrence in science) how many times should a successful trial be replicated to claim proof? That is, whose evidence is *the evidence in evidence based medicine*?

In summary, the term EBM has been linked to three potentially false premises; that evidence has a purely objective meaning in biomedical science, that one can distinguish between what is evidence and what is lack of evidence, and that there is evidence based, and non-evidence based medicine. As long as it is remembered that the term evidence, while delivering forceful promises of truth, is limited in the sense that scientific work can never prove anything but only serves to falsify, the term has some usefulness. Finally, EBM does rely upon the ability to perform systematic reviews (meta-analyses) of the available literature, with all the attendant limitations of meta-analyses discussed above.

In a "tongue and cheek article, Smith and Pell addressed many of the above issues in an article entitled "*Parachute use to prevent death and major trauma related to gravitational challenge: systematic review of randomized control trials*".[23] In their Results Section, they note that they were unable to find any RCTs of "parachute intervention". They conclude that:

> *only two options exist. The first is that we accept that under exceptional circumstances, common sense might be applied when considering the potential risks and benefits of interventions. The second is that we continue our quest for the holy grail of exclusively evidence based interventions and preclude parachute use outside of a properly conducted trial. The dependency we have created in our population may make recruitment of the unenlightened masses to such a trial difficult. If so, we feel assured that those who advocate evidence based medicine and criticize use of interventions that lack evidence base will not hesitate to demonstrate their commitment by volunteering for a double blind, randomized, placebo controlled, crossover trail*. (See Fig. 10.4)


**Fig. 10.4** Parachutes reduce the risk of injury after gravitational challenge, but their effectiveness has not been proved with randomised controlled trials

## References

1. Meinert CL. Meta-analysis: science or religion? *Control Clin Trials*. Dec 1989; 10(4 Suppl):257S–263S.
2. Boden WE. Meta-analysis in clinical trials reporting: has a tool become a weapon? *Am J Cardiol*. Mar 1, 1992; 69(6):681–686.
3. Oxman AD. Meta-statistics: help or hindrance? *ACP J Club*. 1993.
4. Goodman SN. Have you ever meta-analysis you didn't like? *Ann Intern Med*. Feb 1, 1991; 114(3):244–246.
5. Pearson K. Report on certain enteric fever inoculation statistics. *Bri Med J*. 1904; 3:1243–1246.
6. Beecher HK. The powerful placebo. *J Am Med Assoc*. Dec 24, 1955; 159(17):1602–1606.
7. Glass G. Primary, secondary and meta-analysis of research. *Educ Res*. 1976; 5:3–8.
8. Petitti DB. Approaches to heterogeneity in meta-analysis. *Stat Med*. Dec 15, 2001; 20(23):3625–3633.
9. Begg CB, Mazumdar M. Operating characteristics of a rank correlation test for publication bias. *Biometrics*. Dec 1994; 50(4):1088–1101.
10. Egger M, Smith DG, Altman DG. *Systematic Reviews in Health Care: Meta-Analysis in context*. London: BMJ Books; 2000.
11. Candelise L, Ciccone A. Gangliosides for acute ischaemic stroke. *Cochrane Database Syst Rev*. 2001(4):CD000094.
12. Rosanthal R. File drawer problem and tolerance for the null results. *Psychol Bull*. 1979; 86:638–641.
13. Smith ML. Publication Bias and Meta-Analysis. *Eval Educ*. 1980; 4:22–24.
14. Glass G. *Meta-Analysis at 25*. 2000.

15. Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ*. Sept 6, 2003; 327(7414):557–560.
16. Chalmers I, Hedges LV, Cooper H. A brief history of research synthesis. *Eval Health Prof*. Mar 2002; 25(1):12–37.
17. Wells G, Shea B, O'Connell D, et al. The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses.
18. Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. *J Natl Cancer Inst*. Apr 1959; 22(4):719–748.
19. Berlin JA, Colditz GA. The role of meta-analysis in the regulatory process for foods, drugs, and devices. *JAMA*. Mar 3, 1999; 281(9):830–834.
20. *The Cochrane Library, issue 2*. Chichester: Wiley; 2007.
21. Panda A, Dorairajan L, Kumar S. Application of evidence-based urology in improving quality of care. *Indian J Urol* 2007; 23(2):91–96.
22. The Problem of Induction (1953, 1974). http://dieoff.org/page126.htm
23. Smith GC, Pell JP. Parachute use to prevent death and major trauma related to gravitational challenge: systematic review of randomised controlled trials. *BMJ*. Dec 20, 2003; 327(7429):1459–1461.

# Part II

This Part deals with some of the newer approaches in clinical research, specifically research methods for genetic studies, diagnostic testing studies, and pharmacoepidemiology studies. This Part concludes with a chapter that addresses a newer field of Implementation Research – that is, how to implement the research findings that are published, into everyday practice.

> *On being asked to talk on the principles of research, my first thought was to arise after the chairman's introduction, to say, 'Be careful!', and then to sit down.*
>
> J Cornfield, Am J Ment Def. 1959; 64:240

# Chapter 11
# Research Methods for Genetic Studies

**Sadeep Shrestha and Donna K. Arnett**

**Abstract** This chapter introduces the basic concepts of genes and genetic studies to clinicians. Some of the relevant methods and issues in genetic epidemiology studies are briefly discussed with an emphasis on single nucleotide polymorphism based association studies which are currently the main focus of clinical and translational genetics.

Genetics is the fundamental basis of any organism so understanding of genetics will provide a powerful means to discover hereditary elements in disease etiology. In recent years, genetic studies have shifted from disorders caused by a single gene (e.g. Huntington's disease) to common multi-factorial disorders (e.g. hypertension) that result from the interactions between inherited gene variants and environmental factors, including chemical, physical, biological, social, infectious, behavioral or nutritional factors.

A new field of science, Genetic Epidemiology emerged in the 1960s as a hybrid of genetics, biostatistics, epidemiology and molecular biology, which has been the major tool in establishing whether a phenotype (any morphologic, biochemical, physiologic or behavioral characteristic of an organism) has a genetic component. A second goal of genetic epidemiology is to measure the relative size of that genetic effect in relation to environmental effects. Morton and Chung defined genetic epidemiology as "a science that deals with the etiology, distribution, and control of disease in groups of relatives, and with inherited causes of disease in populations".[1] In the era of a known human genome sequence, genetic epidemiology methods have been instrumental in identifying the contribution of genes, the environment and their interactions to better understanding disease processes.

Genomic scientists have predicted that comprehensive, genomic-based care will become the norm, with individualized preventive medicine, early detection of illnesses and tailoring of specific treatments to genetic profile. Practicing physicians and health professionals need to be knowledgeable in the principles, applications, and limitations of genetics to understand, prevent, and treat any biological disorders

S.P. Glasser (ed.), *Essentials of Clinical Research*,

in their everyday practice. The primary objective of any genetic research is to translate information from individual laboratory specimen and build inferences about the human genome and its influence on the risk of disease. This chapter will focus on the fundamental concepts and principles of genetic epidemiology that are important to help clinicians understand genetic studies.

## Important Principles of Genetics

In the 19th century, long before DNA was known, an Augustinian clergyman, Gregory Mendel, described genes as the fundamental unit that transmits traits from parents to offspring. Based on the observations from his cross-breeding experiments in his garden, Mendel developed some basic concepts on genetic information which still provides the framework upon which all subsequent work in human genetics has been based. Mendel's first law is referred to as the independent assortment of alleles (alternate forms of the gene or sequence at a particular location of the chromosome), which states that two genetic factors are transmitted independently of each other. His second law is referred to as the independent segregation of genes which basically states that alleles at one of the parent's genes segregate independently of the alleles at another locus. However, Mendel's law is not always true and loci physically closer in the same chromosomes tend to transmit together; this deviation lays the foundation for genetic epidemiology studies as described in the next section.

All human cells except the red blood cells (RBC) have a nucleus that carries the individual's genetic information organized in chromosomes. Given the diploid nature, each human inherits one copy of the chromosome from the father and the other from the mother. Humans have 22 pairs of autosomal chromosomes and 2 sex-specific chromosomes (X and Y). Chromosomes are composed of molecules called deoxyribonucleic acid (DNA) which contain the basic instructions needed to construct proteins and other cellular molecules.

At the molecular level, DNA is a linear strand of alternating sugars (deoxyribose) and phosphate residues with one of four types of bases attached to the sugar. All information necessary to maintain and propagate life is contained within these four simple bases: adenine (A), guanine (G), thymine (T), and cytosine (C). In addition to this structure of a single strand, the two strands of the DNA molecule are connected by a hydrogen bond between two opposing bases of the two strands (T always bonds with A and C always bonds with G) forming a slightly twisted ladder. It was not until 1953 that James Watson and Francis Creek described this structure of DNA which became the foundation for our understanding of genes and disease.

With the knowledge of underlying molecular biology, gene is defined as the part of the DNA segment that encodes a protein which forms the functional unit of the "hereditary" factor. The basic length unit of the DNA is one nucleotide, or one basepair (bp) which refers to the two bases that connect the two strands. In total,

the human DNA contains about 3.3 billion bp and any two DNA fragments differ only with respect to the order of their bases. Three base units, together with the sugar and phosphate component (referred to as **codons**) translate into amino acids. According to the central dogma of molecular biology, DNA is copied into single stranded ribonucleic acid (RNA) in a process called transcription, which is subsequently translated into proteins. These proteins make intermediate phenotypes which regulate the biology of all diseases, so any difference in the DNA could change the disease phenotype. In many species, only a small fraction of the total sequence of the genome encodes protein. For example, only about 1.5% of the human genome consists of protein-coding exons (about 30,000–40,000), with over 50% of human DNA consisting of non-coding repetitive sequences. We are still in the infancy of understanding the significance of the rest of the non-coding DNA sequence; however the sequence could have structural purposes, or be involved in regulating the use of functional genetic information.

## Units of Genetic Measure

Different genetic markers, which are a segment of DNA with a known physical location on a chromosome with identifiable inheritance, can be used as measures for genetic studies. A marker can be a gene, structural polymorphisms (e.g. insertion/deletion) or it can be some section of DNA such as short tandem repeat (STR) and single nucleotide polymorphism (SNP). Recent advancements in molecular technology have resulted in the discovery of numerous DNA markers and the database is increasing by day. Polymorphism (poly = many and morphism = form) is a sequence variation at any locus (any point in the genome) in the population that has existed for some time and observed in at least 1% of the population, whereas a mutation is recent and the frequency in populations is less than 1%. The terms mutation and polymorphism are often used interchangeably. Variants within coding regions may change the protein function (missense) or predict premature protein truncation (non-sense) and as a result can have effects ranging from beneficial to mutual to deleterious. Likewise, although introns (intragenic regions between coding sequences) do not encode for proteins, polymorphisms can affect intron splicing or expression regulation of adjacent genes. To understand the role of genetic factors it is important to understand these sequence variations within (population) and between (family) generations. We briefly describe the significant ones commonly used for genetic testing.

**STRs**: STRs are tandemly repeated simple DNA sequence motifs of —two to seven bases in length that are arranged head-to-tail and are well distributed throughout the human genome, primarily in the intragenic regions. They are abundant in essentially all ethnically and geographically defined populations and are characterized by simple Mendelian inheritance. STR polymorphisms originate due to mutations caused by slipped-strand mispairing during DNA replication that results from either the gain or loss of repeat units. Mutation rates typically range from $10^{-3}$ to

$10^{-5}$ events per gamete per generation, compared to single nucleotide rates of mutation of $10^{-7}$ to $10^{-9}$. In humans, STR markers are routinely used in gene mapping, paternity testing and forensic analysis, linkage and association studies, along with evolutionary and other family studies. STRs have served as valuable tool for linkage studies of monogenic diseases in pedigrees, but have limited utility for candidate gene association studies.

**SNPs**: SNPs are the variations that occur at a single nucleotide of the sequence. Ninety percent of the polymorphisms in the genome are single nucleotide polymorphisms (SNPs). The human genome contains more than 5.3 million SNPs with a frequency of 10–50% and about 10 million with frequency >1%. SNPs are the markers of choice for association studies because of their high frequency, low mutation rates and the availability of high-throughput detection methods. Most SNPs are found in the non-coding region and have no distinct biological function, but may be surrogate markers or be involved in gene expression and splicing. With few exceptions, the majority of the SNPs are bi-allelic and the genotypes (genetic makeup at both chromosomes) can be heterozygote (different allele in each chromosome) or homozygote (same allele in both chromosomes) for either allele (Fig. 11.1).

Recently, it has been found that SNPs alone cannot explain the complete genetic variations and other structural polymorphisms have been found in higher frequency in the human genome. It is estimated that 5% of the human genome consists of struc-

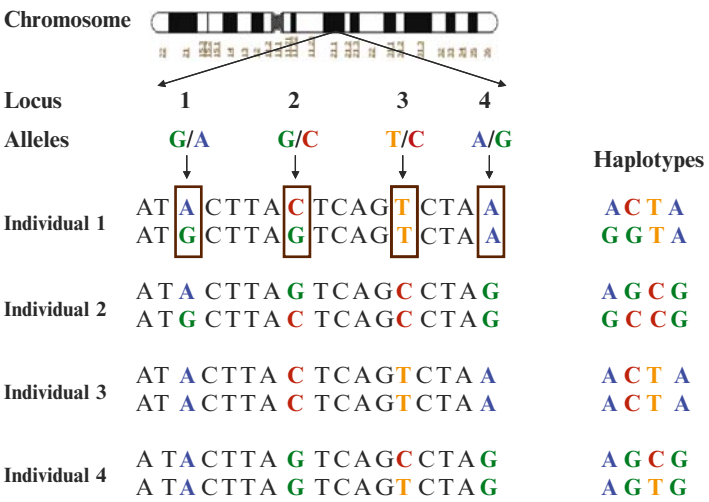



**Fig. 11.1** Alleles and genotypes determined for bi-allelic single nucleotide polymorphisms at four different loci and the corresponding haplotypes. At locus 1, G and A are the alleles; individuals 1 and 2 have AG heterozygote genotype and individuals 3 and 4 have AA homozygote genotype. If the phase is known as shown above, the haplotypes for individual 1 would be ACTA and GGTA. However, in most cases, the variant loci are not physically close and the assays may not be able to partition the phase, thus haplotypes are usually estimated with various methods

tural variants which include deletions, duplications, inversions, and rearrangements of genomic segments. It is estimated that 5% of the human genome is structurally variable.

**Copy number polymorphism**: Recent studies have also focused on copy number variants (CNVs), composed of segmental duplications, large insertion/deletion and inversion of DNA segments 1 kb or larger across the human genome.[2] CNVs are more common in the human genome than originally thought and can have dramatic phenotypic consequences as a result of altering gene dosage, disrupting coding sequences, or perturbing long-range gene regulation.

Although there are different genetic markers (as described above), SNPs are the most frequent variant in the genome and are widely used in genetic studies, so we will refer to SNP polymorphisms to explain the basic concepts in epidemiology especially in the context of association studies.

## Terms and Basic Concepts in Genetic Epidemiology

**Hardy-Weinberg Equilibrium (HWE)**: HWE is one of the key concepts of population genetics that can be used to determine whether a genetic variant could be a valid marker in genetic epidemiology studies. In HWE, allele and genotype frequencies are related through the Hardy-Weinberg law which states that if two alleles, A and a at any locus with frequencies p and q, respectively, are in equilibrium in a population, the proportions of the genotypes, AA homozyogtes, Aa heterozygotes and aa homozygotes will be $p^2$, 2pq, and $q^2$ respectively as a consequence of random mating in the absence of mutation, migration, natural selection, or random drift. One of the implications of HWE is that the allele frequencies and the genotype frequencies remain constant from generation to generation maintaining genetic variation. Extensions of this approach can also be used with multi-allelic and X-linked loci. Deviation from these proportions could indicate (a) genotyping error (b) presence of non-random mating, thus bias in the control selection (c) existence of population stratification (as described later) or (d) recent mutation, migration or genetic drift that has not reached equilibrium. Cases are more likely to represent the tail of a distribution of disease, and any putative genetic variant for that disease may not be in HWE; therefore, it is recommended to assess HWE only in the control groups.

**Linkage and Linkage Disequilibrium (LD)**: Linkage and linkage disequilibrium (LD) are the *sine qua non* of genetic epidemiology. While genes in different chromosomes segregate, Thomas Hunt Morgan and his co-workers observed that genes physically linked to one another on chromosomes of drosophila tended to be transmitted together. This phenomenon where two genetic loci are transmitted together from parent to offspring more often than expected under independent inheritance is termed linkage. Linkage was first demonstrated in humans by Julia Bell and J.B.S Haldane who showed that hemophilia and color blindness tended to

be inherited together in some families. Two loci are linked if recombination (exchange of genetic information between two homologous chromosomes during meiosis) occurs between them with a probability of less than 50%. Recombination is inversely related to the physical distance between loci. However, after several generations, successive recombinations may lead to complete independence even between loci that are very close together.

LD is defined as the extent of non-random association between two genetic loci such that the presence of one allele at a locus provides information about the allele of the other loci. LD occurs in populations as a result of mutation, random genetic drift, selection, and population admixture. Many different measures of LD have been proposed in the literature, most of which capture the strength of association between pairs of SNPs. Although concepts of LD date to early 1900s, the first commonly used LD measure, D was developed by Richard Lewontin in 1964.[3] D measures the departure from allelic equilibrium between separate loci on the same chromosome that is due to the genetic linkage between them. The other two important pairwise measures of LD used in association studies are Lewontin's D′ and $r^2$ also denoted as $\Delta^2$.

For two loci with alleles A/a at the first locus and B/b at the second allele, D is estimated as follows:

$$D = p_{AB} - p_A p_B \tag{1}$$

The disadvantage of D is that the range of possible value depends greatly on the marginal allele frequency. D′ is a standardized D coefficient and is estimated as follows:

$$D' = \frac{D}{D_{\max}} \tag{2}$$

If $D > 0$, $D_{max} = \min [P_A(1\text{-}P_B), P_B(1\text{-}P_A)]$
If $D < 0$, $D_{max} = \min[P_A P_B, (1\text{-}P_A)(1\text{-}P_B)]$
and $r^2$ (Ardlie et al.[4]) is the correlation between two loci and is estimated as follows:

$$r^2 = \frac{D^2}{p_A p_a p_B p_b} \tag{3}$$

Both D′ and $r^2$ range from 0 (no disequilibrium) to 1 (complete disequilibrium), but their interpretation is slightly different. In the case of true SNPs, D′ equals 1 if just two or three of the possible haplotypes are present and is <1 if all four possible haplotypes are present. On the other hand, $r^2$ is equal to 1 if only two haplotypes are present. Association is best estimated using the $r^2$ because it acts as a direct correlation to the allele at the other SNP. Additionally, there is a simple inverse relationship between $r^2$ and the sample size to detect association between susceptibility loci and SNPs.

**Haplotype**: Haplotype is a specific combination of alleles along a chromosome inheriting one from the mother and the other from the father (Fig. 11.1). Recent

studies have shown that the human genome can be parsed into discrete blocks of high LD interspersed by shorter regions of low or no LD. Only a small number of characteristic ("tag") SNPs are sufficient to capture most of the haplotype structure of the human genome in each block. Tag SNPs are loci that can serve as proxies for many other SNPs such that only a subset of loci needs to be genotyped to obtain the same information and power obtained from genotyping a larger number of SNPs. The SNPs within the same block show a strong LD pattern while those in different blocks generally show a weak LD pattern. This advantage, along with the relatively smaller number of haplotypes defined by tag SNPs in each block provides another way to resolve the complexity of haplotypes.

High LD between adjacent SNPs, also result in a much smaller number of haplotypes observed than the theoretical number of all possible haplotypes ($2^n$ haplotypes for n SNPs). There is also biological evidence that several linked variations in a single gene can cause several changes in the final protein product and the joint effect can have an influence on the function, expression and quantity of protein resulting in the phenotype variation. The most robust method to determine haplotypes is either pedigree analysis or DNA sequencing of cloned DNA. Both of these methods are limited by data collection of families or intensive laboratory procedures, but the phase (knowledge of the orientation of alleles on a particular transmitted chromosome) of the SNPs in each haplotype can be directly determined. Haplotypes can also be constructed statistically, although constructing haplotypes from unrelated individuals is challenging because the phase is inferred rather than directly measured. Unless all SNPs are homozygous or at most only one heterozygous SNP is observed per individual, haplotypes cannot be discerned. To account for ambiguous haplotypes, several statistical algorithms have been developed. Three common algorithmic approaches used in reconstructing population-based haplotypes are (i) a parsimony algorithm,[5] (ii) a Bayesian population genetic model that uses coalescent theory,[6] and (iii) a maximum likelihood approach that is based on expectation-maximization (EM) algorithm,[7] The details of these methods are beyond the scope of this book, but readers are referred to the book "Computational Methods for SNPs and Haplotype Inference"[8] for further discussion.

## Biological Specimen

Although the focus of this chapter is not on the laboratory methods of specimen collection, we briefly describe the samples used in clinical studies and their importance. Clinicians deal with different biological organs and tissues in their everyday practice. Most of these however may not be an efficient or convenient source for DNA, the most commonly used resource for genetic studies. Based on factors including cost, convenience for collection and storage, quantity and quality of the source, DNA is commonly extracted from four types of biological specimens: (1) dried blood spots collected in special filter paper (2) whole blood collected in ethylenediaminetetraacetic acid (EDTA) or other anticoagulants such as heparin

and acid citrate dextrose (ACD) (3) lymphocytes isolated from whole blood and EBV-transformed for unlimited source of DNA and (4) buccal epithelial cells collected from swabs or mouth-washes.

## Ethical, Legal and Social Implications (ELSI)

Even for well-intentioned research, one can raise legitimate concerns about the potential misuse of genetic data in regard to social status, employment, economic harm and other factors. A significant amount of work has been done on ethical, legal and social implications (ELSI) research of genetics and policies, but ethics remains an area of major concern. All research protocols can only be conducted upon approval from an institutional review board (IRB) with an appropriate informed consent. It is a routine practice to label the samples with unlinked coded identifiers rather than personal identifiers, so that the individual's identity is masked when linking to phenotypic, demographic, or other personal information. The confidentiality of the data needs to be maximized to protect individual privacy.

## Measurable Outcome and Phenotype

Phenotype is an observable and measurable trait which can be defined qualitatively or quantitatively and does not necessarily have to be related to a disease. Some traits or diseases, like the simple Mendelian traits, have a clear phenotype definition. However other illnesses, like the psychiatric disorders, are complex to define. The unclear classification of cases and controls can be a major problem in any study that can easily introduce biases and inconsistencies between studies. Phenotypes can be defined qualitatively or measured quantitatively. A qualitative trait can be categorized into two or more groups. For example, qualitative traits can be dichotomous (e.g. $HIV^+$ vs. $HIV^-$), ordinal (low, average and high blood pressure group) or nominal (green, black, blue eyes). On the other hand, quantitative measures can be used as continuous variables such as the height or cholesterol level. It may be difficult to examine the genetic effect of quantitative measures, however they can be transformed into meaningful qualitative values where the genetic effect can be more distinct, e.g. the extreme outliers such as dwarfism and hypercholesterolemia. Some diseases may also have an intermediate phenotype that can be measured with molecular markers, while others are strictly based on clinical diagnoses. For example, blood cholesterol levels which can be precisely measured may be a better outcome of cardiovascular disease than flu where the symptoms may be heterogenous in the population and has no intermediate measurement. In other cases environmental exposures such as detection of virus (e.g. HIV viral load) can define a phenotype better than the clinical symptoms since virally infected individuals could be asymptomatic for undefined period of time. Likewise, in one case everyone positive for

HIV could be defined as the outcome of interest while in another scenario clinical symptoms of HIV could define the outcome. Even the ones with the clinical diagnoses, some have distinct symptoms or signs whereas others do not have clear definitions. Some diseases, like Alzheimer's, can have phenotypic heterogeneity, where the same disease shows different features in different families or subgroups of patients. Like in any other clinical study, the key to a genetic study is a clear and consistent definition of the phenotype. The uniformity in phenotype is especially important in multi-center studies.

## General Methods in Clinical Genetic and Genetic Epidemiology Studies

Over the last two decades epidemiologic methods and approaches have been integrated with those of basic genetics to identify the role of genetic factors in disease occurrence in families and populations. Family studies examine the rates of diseases in the relatives of proband cases versus the relatives of carefully matched controls. For many common diseases, the risk to an individual is doubled if a first degree relative is affected. For rare Mendelian disorders, this risk is very high ($10^3$–$10^6$ fold) compared to the general population. For a quantitative trait, such as blood pressure, we can measure correlation of trait values among family members to derive estimates of heritability.

The first step in clinical or epidemiologic genetic studies is to determine whether a phenotype of interest is controlled by a genetic component. There are five key scientific questions that are addressed in sequence in genetic epidemiologic studies (Fig. 11.2): (1) Is there familial clustering? (2) Is there evidence of genetic effect? (3) Is there evidence for a particular genetic model? (4) Where is the disease gene? (5) How does this gene contribute to disease in the general population? The first three questions do not require DNA data and are referred as phenometric studies, but the latter two depend on DNA and referred as genometric studies.

(1) **Familial Aggregation**: The first step to determine whether a phenotype has a genetic component is to examine the clustering within families. Familial aggregation estimates the likelihood of a phenotype in close relatives of cases compared to the non-cases. If the phenotype is a binary trait, familial aggregation is often measured by the relative recurrence risk. The recurrence risk ratio is the ratio of prevalence of the phenotype in relatives of affected cases to the general population. Greater risk associated with closer degrees of relatedness could also indicate the genetic component. If the prevalence of the phenotype is higher in 1st degree relatives (father, mother, siblings) versus 2nd degree relatives (uncle aunt, cousins) it would suggest a genetic component since the 1st degree relatives share more genetic information than the 2nd degree relatives. For example, Kerber and O'Brien showed a distinctly higher RR for common cancers in the Utah genealogical and cancer registry of individuals born between 1870 and

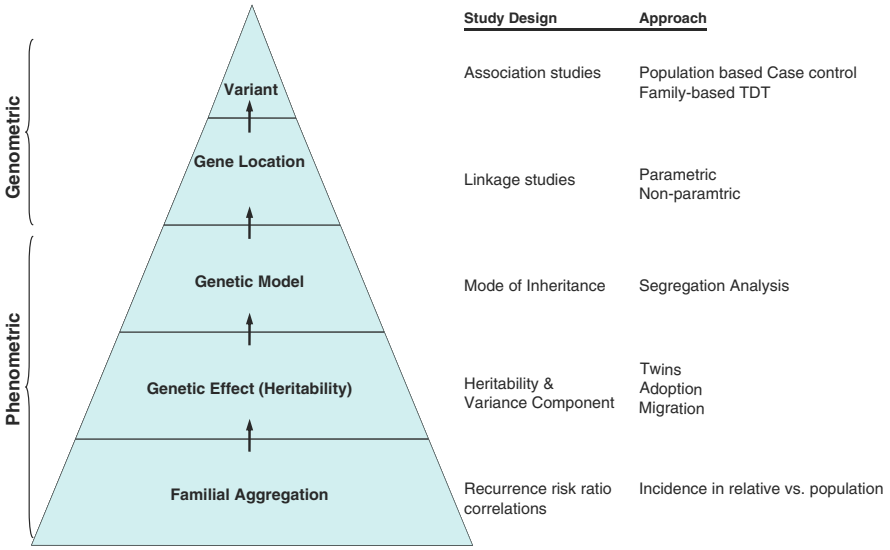



**Fig. 11.2** Systematic designs and approaches in genetic epidemiology studies to identify the genetic and non-genetic causal of disease

1984.[9] On the other hand, assessment of familial aggregation of a continuous trait, such as height can be estimated with a correlation or covariance-based measure such as intrafamily correlation coefficient (ICC). The ICC indicates the proportion of the total variability in a phenotype that can reasonably be attributed to real variability between families. Because an increased risk in family members does not directly indicate genetic inheritance because of the influence of the shared familial environment, familial aggregation is necessary but not sufficient evidence of genetic contribution. It is difficult to disentangle genetic effect from the environmental effect due to the shared physical environment. For example, obesity could be due to shared genes within the family or the eating or physical activity habits in the family.

(2) **Genetic Effect**: Once the familial aggregation is established, the next step is to distinguish between genetic and non-genetic familial effects and estimate the extent of genetic effect. Different variance component models such as heritability, which is defined as the proportion of variation directly attributable to genetic differences among relatives to the total variation in the population (both genetic and environmental) is traditionally used to estimate the genetic effect in familial aggregation. Heritability can be estimated both for qualitative and quantitative traits although it was developed for the latter phenotypes. Heritability is however population-specific and must be used with caution when comparing different populations. Other classical designs for distinguishing non-genetic family effects from genetic effects have been studies of twins, adoptees and migrants.

**Twin studies**: Studies of twins are useful in estimating the contribution to a phenotype through the comparison of monozygotic (MZ) pairs (who share all genes) with dizygotic (DZ) pairs (who share only half of their genes). If family upbringing acts equally on monozygotic twins as it does on dizygotic twins, then the greater similarity of phenotypes in MZ than DZ twins must be due to genetic factors. A standard measure of similarity used in twin studies is the concordance rate. For example, the concordance rate of diabetes Type I is 25–35% among MZ, 5–6% among DZ twins or siblings and 0.4% among the general population, clearly indicating a genetic component.

**Adoption studies**: This study design examines the similarity and differences in the phenotype in the biological parents and foster parents of adoptees, and in their biological and adopted siblings. The assumptions are that the similarity between an adopted child and biological parent is due only to genetic effects, while that between the adopted child and the adoptive parent or adoptive siblings is due to the shared environment.

**Migration studies**: A similar incidence in migrants compared to the aboriginal population's incidence suggests a strong environmental factor, whereas similar incidence to the original ethnic group or relatives in the original residence could suggest genetic effect. Genes do not change as easily as environment, so the variation in the phenotype after taking into account of all the common and new environmental factors could point to a genetic effect.

(3) **Genetic Model**: After the genetic basis is established, the next step is to find the mode of inheritance which has historically been done using segregation analyses, although these methods are not as common in the era of SNP association studies. Segregation analyses does not use DNA-based genetic data, but rather, the methods test whether or not the observed phenotype follows a Mendelian inheritance in the offspring in the pedigree. Mendelian diseases can be autosomal dominant, autosomal recessive, X-linked dominant, or X-linked recessive (usually with high penetrance and low frequency of risk alleles). Traditional segregation analysis has primarily studied simple Mendelian disorders where a single gene mutation is sufficient and necessary to cause a disorder. However, most common chronic diseases are regarded as complex where a large number of genetic variants along with environmental factors interact with each other (necessary or un-necessary but not sufficient) to affect the disease outcomes. These diseases usually cluster in families, but do not follow a traditional Mendelian inheritance pattern. While segregation analyses are powerful to test different modes of Mendelian inheritance in the family, it is not useful for complex traits. Linkage and association analysis, both of which utilize DNA, are more powerful to study genetic effects of complex diseases.

(4) **Disease Gene Location**:

**Linkage studies**: Linkage studies focus on concordant inheritance and are used to identify broad genomic regions that contain gene or genes associated with the phenotype, in the absence of previous biologically driven hypotheses. Major genes for

monogenic traits have been located by linkage analysis. Genetic linkage analysis tests whether the marker segregates with the disease in pedigrees with multiple affected according to a Mendelian mode of inheritance and relies entirely on the tendency for shorter haplotypes to be passed on to the next generation intact, without recombination events at meiosis. If a marker is passed down through family generation and occurs more commonly in cases than controls, then the marker can be used as a surrogate for the location of the gene. Genetic linkage analysis test is formulated as logarithm of the ratio $L(\theta)/L(\theta = 0.5)$ or lod score, i.e., the likelihood of observing the segregation pattern of the marker alleles at a given recombination frequency $\theta$ compared with the likelihood of the same segregation pattern in the absence of linkage.

Two types of linkage analysis can be performed: parametric and nonparametric analysis. Parametric or model-based linkage analysis by the lod score method requires a defined model specifying the relationship between the phenotype and the factors (environmental and genetic), which have an effect on phenotype expression. For example, such a model can be provided by complex segregation analysis. The objective of parametric linkage analysis is to estimate the recombination frequency ($\theta$) and to test whether $\theta$ is less than 0.5, which is the case when two loci are genetically linked. The nonparametric or model-free approach evaluates the statistical significance of excess allele sharing for specific markers among affected sibs and does not require information about the mode of disease inheritance. The genes contributing to the phenotypic variation have been successfully localized by linkage (cosegregation) analysis for Mendelian diseases that have a strong genetic effect and are relatively rare. For more complex diseases, usually fine mapping with association studies are carried out to narrow down the putative disease locus after initial linkage finding.

**Association studies**: Genetic association studies aim to correlate differences in allelic frequencies at any locus with differences in disease frequencies or trait levels. We would see a genetic association if the specific genetic variant is more frequent in the affected group than the non-affected group. Most association studies represent classical case-control approaches where the risk factor under investigation is the allele at the genetic marker (mostly with SNPs). SNP-based association studies can be performed in two ways: (i) direct testing of an exposure SNP with a known varying function such as altered protein structures and (ii) indirect testing of a SNP which is a surrogate marker for locating adjacent functional variant that contributes to the disease state (Fig. 11.3a). The first method requires the identification of all common variants in coding and regulatory regions of genes. The latter method avoids the need for cataloguing potential susceptibility variants by relying instead on association between disease and neutral polymorphisms marking haplotypes located near a risk-conferring variant. It exploits the phenomenon of linkage disequilibrium (LD) between alleles of closely linked loci forming haplotypes within the genomic regions.

Given the diallelic nature of SNPs, a disease locus may be difficult to find unless the marker is closely linked to the disease locus. Apart from a single SNP association strategy, a dense panel of SNPs from the coding and non-coding regions of the

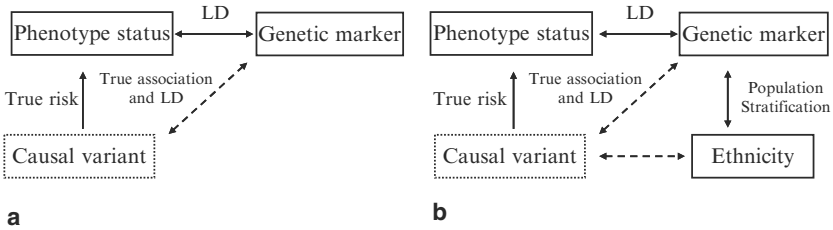



**Fig. 11.3** True association, LD and the effect of population stratification. (a) Genetic marker that is in LD with causal variant serves as a surrogate of the true association with the phenotype. (b) Population stratification is a confounder that leads to spurious association

gene that form haplotypes can also be tested in cases and controls. Initial studies have also demonstrated that the analysis of haplotypes rather than individual SNPs can detect association with complex diseases. It has been suggested that single SNP-based candidate gene studies may be statistically weak as true associations may be missed because of the incomplete information from individual SNPs. For example, haplotypes contain more heterozygosity than any of the individual markers that comprise them and also mark more of the variation in the gene than single SNPs. Several haplotype association studies in the case control design have shown the power of haplotypes over individual SNPs.

**Gene contribution**: Once the association of the genetic allele is discovered, it is important to assess the contribution of this variant to the phenotype. The public health relevance of a given polymorphism is addressed by estimating the proportion of diseased individuals in the population that could be prevented if the high-risk alleles were absent (known as attributable fraction, etiologic fraction, or population attributable risk percent). Accurate estimation of the population frequency of the high-risk variant (allele and/or genotype) is important because attributable fraction is a function of the frequency of the high-risk variant in the population and the penetrances (i.e., the likelihood that the trait will be expressed if the patient carries the high-risk variant). Attributable fractions can also be used to estimate the proportion of disease that is a result of the interaction of a genetic variant and an environmental exposure. Genetic variants are not usually modifiable within the longevity of an individual (although very possible evolutionarily over time); therefore the prevention of disease will depend on interventions that target environmental factors that interact with genetic susceptibility to influence the risk of disease.

## Candidate Gene vs. Genome-wide Association Studies

Candidate gene approaches examines polymorphisms in genes with functional significance related to the phenotype of interest. Some of the candidate genes are also based on physical location or sequence homology to a gene encoding protein that

is in the etiologic pathway. As attractive as this hypothesis-driven candidate gene approach is, it focuses exclusively on the relatively few known genes, ignoring many that have not yet been characterized. One major drawback of candidate gene approach is that *a priori* knowledge of the pathogenesis of the disease is required – when the molecular mechanism is poorly understood, it could lead to selection of the wrong genes. Even with the right genes within the pathway, the challenge is to find variants that influence the regulation of gene function. Candidate gene studies have proven to be more successful when used as a follow-up of linkage studies. For examples, APOE4, the most common genetic factor associated with Alzheimer's disease was primarily discovered by candidate gene approach following the linkage study which mapped to chromosome 19.

Alternatively, with assurance of adequate power, hypothesis-generating genome-wide association studies (GWAS) are also being widely used. The GWASs has the advantage that no *a priori* knowledge of the structure or function of susceptibility genes is required. Hence, this approach provides the possibility of identifying genes that modulate susceptibility to infectious diseases that had previously not been suspected of playing such a biological role. Upon completion of the human genome project and the draft of the human HapMap project (define genetic variation patterns in individuals from Nigeria (Yoruba), Japan, China and US with European ancestry), technological advances have led to the identification and cost-effective high-throughput genotyping arrays of common genetic variants making the GWASs more promising and attractive. A two step design is being embraced by researchers where common variation is first screened for association signals using cost-effective typing of tagging SNPs with GWASs approach followed by denser sets of SNPs in regions of potentially positive signals. If the sample size is large enough, a third stage of validation of association can also be conducted with proper power calculations. Although promising results have been found for different phenotypes with GWAS, analytical considerations are still underway to develop a robust strategy to interpret the findings especially for complex diseases with multiple gene-gene and gene-environmental interactions.

## Risk Quantification

**Gene-gene and gene-environment interaction**: A central theme of genetic epidemiology is that human disease is caused by interactions within and between genetic and non-genetic environmental factors. Thus in the design and analysis of epidemiologic studies, such interaction needs to be explicitly considered. A simple approach would be to create a classic $2 \times 2$ table with genotypes at the two loci classified as present or absent and compute odds ratios for all groups with one reference group. The extent of the joint effect of two loci can be compared with the effects for each locus independently. The same approach can be considered for gene-environmental interaction for qualitative measurements. However, as more genes are involved and the environmental exposure is quantitatively measured, the

analysis and interpretation of the interaction can be complicated, but various methods are being continuously developed. Large sample sizes are needed to observe true interactions, especially if they are small effects.

## Additional Applications of Genetic Studies

Most of the genetic studies (candidate or genome-wide) are focused on case-control design with the underlying goal of understanding the biological cause of the disease. Other time dependent studies can be performed to understand the genetic effect in the natural history or progression of the disease. The outcomes of these studies are helpful for providing counseling to individuals about their offspring (genetic screening) or the interaction between environmental factors. However, there are a growing number of genetic studies examining the differential response to drugs or vaccines. For instance, **pharmacogenetic** studies focus on genetic determinants of individual variation in response to drugs, including variation in the primary domain of drug action and variation in risk for rare or unexpected side effects of drugs. Likewise, **vaccinogenetic** studies examine the genetic determinants of differential vaccine response and side effects between individuals.

## Major Issues and Limitations in Genetic Studies

In most cases with complex diseases, the effect of any genetic variant is small and can only be observed in studies with larger sample size or the frequency of the allele is rare. There are very few common variants (>10% allele frequency) with a relative risk exceeding two (e.g. APOE and Alzheimer's disease). A major concern with respect to genetic association studies has been lack of replication studies, especially contradictory findings across studies. Replication of findings is very important before any causal inference can be drawn. The severity of this problem can be best exemplified in a comprehensive review conducted by Hirschhorn et al. where they surveyed 600 positive associations (166 associations studied three or more times) between gene variants and common diseases and showed that only six were replicated consistently. However, before jumping to the conclusion of the false positive results, several study design and statistical issues need to be seriously considered when conducting genetic studies which are briefly described below:

(1) **Genetic Heterogeneity**: There are several cases where multiple alleles at a locus are associated with the same disease. This phenomenon is known as **allelic heterogeneity** and can be observed with multi-allelic locus. This may thus explain why in some studies one allele is associated with the disease and

in other studies it is another allele. Likewise, locus heterogeneity may also exist where multiple genes influence the disease independently and thus a gene found to be associated in one study may not be replicated in the other but rather another gene may be associated.

(2) **Confounding**: One crucial consideration in genetic studies is the choice of an appropriate comparison group. In general, as in any well-designed epidemiological case-control studies, controls need to be sampled from the same source population as the cases. The use of convenient comparison groups without proper ascertainment criteria may lead to spurious findings as a result of confounding caused by unmeasured genetic and environmental factors. Population stratification can occur if cases and controls have different frequencies of ethnic groups or individuals have differential admixture (the proportions of the genome that have ancestry from each subpopulation), and when phenotypes of interest differ between ethnic groups (Fig. 11.3b). Although most genetic variation is inter-individual, there is also significant inter-ethnic variation irrespective of disease status. One classic example is reported by Knowler et al. (1988) who showed spurious inverse association between variants in the immunoglobulin haplotype Gm3;5,13,14 and non-insulin dependent diabetes mellitus among the Pima-Papago Indians (Knowler et al., 1988). Individuals with the haplotype had a higher prevalence of diabetes than those without it (29% vs. 8%). This haplotype, however measured the subjects' degree of Caucasian genetic heritage and when the analysis was stratified by degree of admixture, the association did not exist.

One way to overcome such issue of confounding by population stratification is to conduct family based designs with special statistical analyses such as transmission-disequilibrium test (TDT). Basically, in TDT, alleles of parents not transmitted to the patients are used as "virtual control" genotypes so any population-level allele frequency differences become irrelevant. Several other family-based and population-based methods have also been derived from TDT. While these methods are attractive because they correct false positives from population stratification, family-based samples are difficult to collect and might not be feasible for late-onset diseases where the parents might be deceased. Another approach is to use a "homogeneous" population. In recent years, there is growing interest to study genetically isolated populations such as Finland and Iceland. These populations have been isolated for several years and expanded from a small group of individuals called "**founder population**". Founder population limits the degree of genetic diversity making more or less a homogenous population. One major limitation of finding from such isolated population is the generalizability to other populations which may have different genetic make-ups.

Studies have shown that there is admixture even within such isolated populations. An alternate method to control for population stratification is to use unrelated markers from the non-functional region of the genome as indicators of the amount of background diversity in individuals. The first approach referred as "**genomic control**" adjusts the standard $\chi^2$ statistic in the case-control analysis by a scaling

factor based on the degree of stratification measured by the unlinked neutral markers. The second is the structured-association approach pioneered by Pritchard and colleagues, which uses Bayesian methods (using programs such as STRUCTURE) to cluster subjects into homogenous groups using **ancestry informative markers** (AIMS) and performing analysis within these groups. AIMs are identified based on the differences in sequence between the world's various populations (0.1% of the human genome).

(3) **Genotype Error and Misclassification**: For family-based studies (trio data for TDT), genotyping errors have been shown to increase type I and type II errors and for population-based (case-control) studies it can increase type II errors and thus decrease the power. Additionally, misclassification of genotypes can also bias LD measurements.

In general, genotyping errors could be a result of poor amplification, assay failure, DNA quality and quantity, genomic duplication or sample contamination. It is important that a quality-check be performed for each marker and the low-performance once be removed from the analysis before the results are interpreted. Several laboratory based methods such as (a) genotyping duplicate individuals (b) genotyping the same individuals for the same marker using different assay platforms or (c) genotyping in family pedigrees to check for Mendelian inconsistency, (i.e. the offspring should share the genetic makeup of the parents and any deviation could indicate genotype error) can be used to assure the quality of the genotypic data. Testing for HWE is also commonly used, however it is important to note that deviation from HWE does not necessarily indicate genotype error and could be due to any of the underlying causes as described earlier.

(IV) **Multiple Testing**: Regardless of whether each SNP is analyzed one at a time or as part of a haplotype, the number of individual tests can become very large and can lead to an inflated (false positive) type I error rate both in candidate gene approach and whole genome approach. If the selected SNPs are all independent, then adjustments to the conventional p-value of 0.05 with Bonferroni correction could account for the multiple testing. However, given the known LD pattern between SNPs, such adjustments would overcorrect for the inflated false-positive rate, resulting in a reduction in power. An alternate method would be to use the False Discovery Rate (FDR) approach which rather than correcting the p-value, corrects for fraction of false-positives with the significant p-value. When a well defined statistical test is performed (testing a null against an alternative hypothesis) multiple times, the FDR estimates the expected proportion of false positives from among the tests declared significant. For example, if 100 SNPs are said to be significantly associated with a trait at a false discovery rate of 5%, then on average 5 are expected to be false positives. However, the gold standard approach that is being appreciated more is the permutation testing where the groups status of the individuals are randomly permuted and the analysis repeated several times to get a distribution for the test statistics under the null hypothesis but this method can also be computationally intensive and time-consuming.

## Concluding Remarks

The completion of the Human Genome Project in 2003 has heightened expectations of the health benefits from genetic studies. Methods in genetic epidemiology are very powerful in examining and identifying the underlying genetic basis of any phenotype, if conducted properly. There are several study designs that can be used with a common goal of finding both the individual effects and interactions within and between genes and environmental exposures that causes the disease. With the availability of cost-effective high-throughput technologies, currently SNP-based case-control studies are the widely accepted approach, with some considerations for CNVs. Regardless of the approach, several design and methodological issues need to be seriously considered when conducting studies and interpreting the results (Table 11.1). Although these studies may find association of the phenotype with a genetic variant, the challenge is to meaningfully translate the findings. In most instances the alleles are in the non-coding region and the frequencies are rare but this the stepping stone in the process of understanding the complexity of common diseases. Very rarely can we find a conclusive evidence of genetic effect from a single study, so replication studies with larger samples size should be encouraged to provide insurance against the unknown confounders and biases. To ensure the biology of the variants, animal studies and gene expression studies can be conducted as follow-up studies. Clinicians need to be aware of the potential role of

**Table 11.1** Possible explanations to consider before interpreting the association study results

| Outcomes of association studies | Possible explanations to consider |
|---|---|
| Positive association | – True causal association |
| | – LD with causal variant |
| | – Confounding by population stratification |
| | – Hardy Weinberg disequilibrium |
| | – Multiple comparison (false positive) |
| Negative association | – No causal association |
| | – Small sample size |
| | – Phenotype misclassification |
| Multiple genes associated to the same phenotype | – Genetic heterogeneity |
| | – Interactions within and between genes and environmental factor |
| | – False positive |
| Multiple alleles at the same gene associated to the same phenotype | – Allelic heterogeneity |
| | – False positive |
| Same allele in the same gene associated with the same phenotype but in opposite direction | – Confounding by population stratification |
| | – Phenotype heterogeneity |
| | – False positive |

genetics in disease etiology and thus be familiar with methods and issues in conducting genetic epidemiology studies in order to conduct their own studies or assist other researchers.

## Recommended Readings

Hartl DL, Clark AG. *Principles of Population Genetics*. Sunderland: Sinauer Associates; 2007.
Khoury MJ, Beaty TH, Cohen BH. *Fundamentals of Genetic Epidemiology*. 4th ed. New York: Oxford University Press; 1993.
Khoury MJ, Burke W, Thomson, EJ (eds). *Genetics and Public Health in the 20th Century*. New York: Oxford University Press; 2000.
Knowler WC, Williams RC, Pettitt DJ, et al. Gm3;5, 13, 14 and type 2 diabetes mellitus: an association in American Indians with genetic admixture. AM J Hum Genetic. Oct 1988; 43(4): 520–526.
Morton, NE. *Outline of Genetic Epidemiology*. Basel: Karger; 1982.
Ziegler A & Konig IR. *Statistical Approach to Genetic Epidemiology: Concepts and Applications*. Weinheim: Wiley-VCH/Verlag/GmbH & Co. KGaA; 2006.

## References

1. Morton NE. *Genetic Epidemiology*. New York: Academic; 1978.
2. Redon R, Ishikawa S, Fitch KR, et al. Global variation in copy number in the human genome. *Nature*. Nov 23 2006; 444(7118):444–454.
3. Lewontin RC. The interaction of selection and linkage. I. General considerations; heterotic models. *Genetics*. 1964; 49:49–67.
4. Ardlie K, Kruglyak L, Seislstad. Patterns of linkage disequilibrium in the human genome. *Nat Genet*. 2002; 3:299–309.
5. Clark AG. Inference for haplotypes from PCR-amplified samples of diploid populations. *Mol Biol Evol*. 1990; 7:111–122.
6. Lin S, Cutler D, Zwick M, Chakravarti A. Haplotype inference in random population samples. *Am J Hum Genet*. 2002; 71:1129–1137.
7. Excoffier L, Slatkin M. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol Biol Evol*. 1995; 12:921–927.
8. Istrail S, Waterman M, Clark AG. *Computational Methods for SNPs and Haplotype Inference*. Berlin, Heidelberg: Springer; 2004.
9. Kerber RA, O'Brien E. A cohort study of cancer risk in relation to family histories of cancer in the Utah population database. *Cancer*. May 1 2005; 103(9):1906–1915.

# Chapter 12
# Research Methods for Pharmacoepidemiology Studies

**Maribel Salas and Bruno Stricker**

**Abstract** Pharmacoepidemiology (PE) applies epidemiologic concepts to clinical pharmacology. This discipline was born on 1960s and since then various methods and techniques have been developed to design and analyze medications' data.[1] This chapter will review the factors involved in the selection of the type of pharmacoepidemiologic study design, and advantages and disadvantages of these designs. Since other chapters describe randomized clinical trials in detail, we will focus on observational studies.

Pharmacoepidemiology (PE) is the discipline that studies the frequency and distribution of health and disease in human populations, as a result of the use and effects (beneficial and adverse) of drugs. PE uses methods similar to traditional epidemiologic investigation, but applies them to the area of clinical pharmacology.[1] In this chapter, we discussed general concepts of clinical research with emphasis on those related to PE.

In the last few years, PE has acquired relevance because of various drug withdrawals from the market; and, as a result of public scandals related to drug safety and regulatory issues. Some of these withdrawn and controversial drugs include troglitazone,[2–4] cisapride,[5,6] cerivastatin,[7–10] rofecoxib,[11–13] and valdecoxib.[13–15] One of the major allegations cited with each of these drug withdrawals were flaws in the study designs that were used to demonstrate drug efficacy or safety. Furthermore, the study designs involved with these withdrawn drugs were variable and reported conflicting results.[16] An example of the controversies surrounding drug withdrawals is the association of nonsteroidal antiinflamatory drugs (NSAID) with chronic renal disease.[17–21] The observation that one study may produce different results from another, presumably similar study (and certainly from studies of differing designs) is, of course, not unique to PE, as has been discussed in prior chapters.

This chapter will review the factors involved in the selection of the type of pharmacoepidemiologic study design, and advantages and disadvantages of these designs. Since other chapters describe randomized clinical trials in detail, we will focus on observational studies.

S.P. Glasser (ed.), *Essentials of Clinical Research*,

## Selection of Study Design

In PE as in any clinical research, investigators need to select the appropriate study design, to answer an appropriate research question that includes the objective and the purpose of the study. There is a consensus that an appropriate research question includes information about the exposure, outcome, and the population of interest. For example, an investigator might be interested in the question of whether there is an association of rosiglitazone with cardiac death in patients with type 2 diabetes mellitus. In this case, the exposure is the antidiabetic drug rosiglitazone, the outcome is cardiac death, and the population is a group of patients with type 2 diabetes. Although this may seem simplistic, it is surprising how many times it is unclear what the exact research question of a study is, and what the elements are which are under study.

The key elements for clearly stated objectives are keeping them SMART: Specific, Measurable, Appropriate, Realistic and Time-bound (SMART).[22] An objective is specific if it indicates the target; in other words, who and what is the focus of the research, and what outcomes are expected. By measurable, it is meant that the objective includes a quantitative measure. Appropriate, refers to an objective that is sensitive to target needs and societal norms, and realistic refers to an objective that includes a measure which can be reasonably achieved under the given conditions of the study. Finally, time-bound refers to an objective that clearly states the study duration. For example, a clearly stated objective might be: 'to estimate the risk of rosiglitazone used as monotherapy on cardiac death in patients with type 2 diabetes treated between the years 2000 to 2007.' In summary, in PE as in other areas of clinical research, clearly stated objectives are important in order to decide on the study design and analytic approach. That is, when a researcher has a clear idea about the research question and objective, it leads naturally to the optimal study design. Additionally, the investigator then takes into account the nature of the disease, the type of exposure, and available resources in order to complete the thought process involved in determining the optimal design and analysis approach. By the 'nature of the disease' it is meant that one is cognizant of the natural history of the disease from its inception to death. For example, a disease might be acute or chronic, and last from hours to years, and these considerations will determine whether the study needs to follow a cohort for weeks or for years in order to observe the outcome of interest. In PE research, the exposure usually refers to a drug or medication, and this could result in a study that could vary in duration (hours to years), frequency (constant or temporal) and strength (low vs. high dose). All of these aforementioned factors will have an impact on the selection of the design and the conduct of the study. In addition, a researcher might be interested in the effect of an exposure at one point in time (e.g. cross-sectional) vs. an exposure over long periods of time (e. g. cohort, case-control).

Since almost every research question can be approached using various designs, the investigator needs to consider both the strengths and weaknesses of each design in order to come to a final decision. For example, if an exposure is rare, the most

| | Prevalence or Incidence of Outcome | | |
|---|---|---|---|
| | | Not Rare | Rare |
| Drug Exposure | Not Rare | Cohort or clinical trial | Case-control |
| | Rare | Cohort | Case-Cohort |

**Fig. 12.1** Designs by frequency of exposure and outcome

efficient design is a cohort study (provided the outcome is common) but if the outcome is rare, the most efficient design is a case-control study (provided the exposure is common). If both the outcome and exposure are rare, a case-cohort design might be appropriate where odds ratio might be calculated with exposure data from a large reference cohort (Fig. 12.1).

## Study Designs Common in PE

Table 12.1 demonstrates the study designs frequently used in PE research. Observational designs are particularly useful to study unintended drug effects in the postmarketing phase of the drug cycle. It is also important to consider the comparative effectiveness trial that is used in postmarketing research (see Chapter 5).

Effectiveness trials can be randomized or not randomized, and they are characterized by the head-to-head comparison of alternative treatments in large heterogeneous populations, imitating clinical practice.[23–25] As it is mentioned in Chapter 3, randomized clinical trials provide the most robust evidence, but they have often limited utility in daily practice because of selective population, small sample size, low drug doses, short follow-up period, and highly controlled environment.[26]

## Descriptive Observational Studies

Recall that these are predominantly hypothesis generating studies where investigators try to recognize or to characterize a problem in a population. In PE research, for example, investigators might be interested in recognizing unknown adverse effects, in knowing how a drug is used by specific populations, or how many people

**Table 12.1** Type of designs used in PE research

| |
|---|
| I. Descriptive observational studies |
| A. Case report |
| B. Case series |
| C. Ecologic studies |
| D. Cross-sectional studies |
| II. Analytical studies |
| Observational studies |
| A. Case-control studies |
| B. Cross-sectional studies |
| C. Cohort studies |
| D. Hybrid studies |
| 1. Nested case-control studies |
| 2. Case-cohort studies |
| 3. Case-crossover studies |
| 4. Case-time studies |
| Interventional studies |
| A. Controlled clinical trials |
| B. Randomized, control clinical trials |
| C. N of trials |
| D. Simplified clinical trials |
| E. Community trial |

might be at risk of an adverse drug event. As a consequence, these studies do not generally measure associations; rather, they use measures of frequency such as proportions, rate, risk and prevalence.

## Case Report

Case reports are descriptions of the history of a single patient who has been exposed to a medication and experiences a particular and unexpected effect, whether that effect is beneficial or harmful. In contrast to traditional research, in pharmacoepidemiologic research, case reports have a privileged place, because they can be the first signal of an adverse drug event, or the first indication for the use of a drug for conditions not previously approved (off-label indications by the regulatory agency e.g. Food and Drug Administration). As an example, case reports were used to communicate unintended adverse events such as phocomelia associated with the use of thalidomide.[27] Case reports also make up the key element for spontaneous reporting systems such as MedWatch, The FDA Safety Information and Adverse Event Reporting Program. The MedWatch program allows providers, consumers and manufacturers to report serious problems that they suspect are associated with the drugs and medical devices they prescribe, dispense, or use. By law, manufacturers, when they become aware of any adverse effect, must submit a case report form of serious unintended adverse events that have not been listed in the drug labeling within 15 calendar days.[28]

## Case Series

Case series is essentially a collection of 'case reports' that share some common characteristics such as being exposed to the same drug; and, in which same outcome is observed. Frequently, case series are part of phase IV postmarketing surveillance studies, and pharmaceutical companies may use them to obtain more information about the effect, beneficial or harmful, of a drug. For example, Humphries et al. reported a case series of cimetidine carried out in its postmarketing phase, in order to determine if cimetidine was associated with agranulocytosis.[29] The authors followed new cimetidine users, and ultimately found no association with agranulocytosis. Often, case series characterize a certain drug-disease association in order to obtain more insight into the clinicopathological pattern of an adverse effect; such as, hepatitis occurring as a result of exposure to nitrofurantoin.[30] The main limitation of case series is that they do not include a comparison group(s). The lack of a comparison group is critical, and the result is that is difficult to determine if the drug effect is greater, the same or less than the expected effect in a specific population (a situation that obviously complicates the determination of causality).

## Ecologic Studies

Ecologic studies evaluate secular trends and are studies where trends of drug-related outcomes are examined over time or across countries. In these studies, data from a single region can be analyzed to determine changes over time; or, data from a single time period can be analyzed to compare one region vs. another. Since ecologic studies do not provide data on individuals (rather they analyze data based on study groups), it is not only impossible to adjust for confounding variables; but, it does not reveal whether an individual with the disease of interest actually used the drug (this is termed the ecologic fallacy). In ecologic studies, sales, marketing, and claims databases are commonly used. For example, one study compared urban vs. the rural areas in Italy using drug sales data to assess for regional differences in the sales of tranquilizers.[31,32] For the reasons given above, ecologic studies are limited in their ability to associate a specific drug with an outcome; and, invariably there are usually other factors that could also explain the outcome.

## Cross-Sectional Studies

Cross-sectional studies are particularly useful in drug utilization studies and in prescribing studies, because they can present a picture of how a drug is actually used in a population or how providers are actually prescribing medications. Cross-sectional studies can be descriptive or analytical. Cross-sectional studies are considered

descriptive in nature when they describe the 'big' picture about the use of a drug in a population, and the information about the exposure and the outcome are obtained at the same point in time. Cross sectional designs are used in drug utilization studies because these studies are focused on prescription, dispensing, ingesting, marketing, and distribution; and, also address the use of drugs at a societal level, with special emphasis on the drugs resultant effect on medical, social, and economic consequences. Cross-sectional studies in PE are particularly important to determine how specific groups of patients, e.g. elderly, children, minorities, pregnant, etc. are using medications. As an example, Paulose-Ram et al. analyzed the U.S. National Health and Nutrition Examination Survey (NHANES) from 1988 to 1994 in order to estimate the frequency of analgesic use in a nationally representative sample from the U.S. From this study it was estimated that 147 million adults used analgesics monthly, women and Caucasians used more analgesics than men and other races, and more than 75% of the use was over the counter.[33]

## Analytical Studies

Analytic studies, by definition, have a comparison group and as such are more able to assess an association or a relationship between an exposure and an outcome. If the investigator is able to allocate the exposure, the analytical study is considered to be an interventional study; while if the investigator does not allocate the exposure; the study is considered observational or non-experimental (or non-interventional). Analytical observational pharmacoepidemiolgic studies quantify beneficial or adverse drug effects using measures of association such as rate, risk, odds ratios, rate ratios, or risk difference.

## Cross-Sectional Studies

Cross-sectional studies can be analytical if they are attempting to demonstrate an association between an exposure and an outcome. For example, Paulose-Ram et al. used the NHANES III data to estimate the frequency of psychotropic medication used among Americans between 1988 and 1994; and, to estimate if there was an association of sociodemographic characteristics with psychotropic medication use. They found that psychotropic medications were associated with low socioeconomic status, lack of high school education, and whether subjects were insured.[34] The problem with analytical cross-sectional studies is that it is often unknown whether the exposure really precedes the outcome because both are measured at the same point in time. This is obviously important since if the exposure does not precede the outcome, it can not be the cause of that outcome. This is especially in cases of chronic disease where it may be difficult to ascertain which drugs preceded the onset of that disease.

## Case-Control Studies (or Case-Referent Studies)

Case control and cohort studies are designs where participants are selected based on the outcome (case-control) or on the exposure (cohort) Fig. 12.2. In PE case-control studies, the odds of drug use among cases (the ratio exposed cases/unexposed cases) are compared to the odds of drug use among non cases (the ratio exposed controls/unexposed controls). The case-control design is particularly desirable when one wants to study multiple determinants of a single outcome.[35] The case-control design is a particularly efficient study when the outcomes are rare, since the design guarantees a sufficient number of cases. For example, Ibanez et al. designed a case-control study to estimate the association of non-steroidal anti-inflammatory drugs (NSAID) (common exposure) with end-stage renal disease (a rare outcome). In this study, the cases were patients entering a local dialysis program from 1995 to 1997 as a result of end-stage renal disease; while controls, were selected from the hospital where the case was first diagnosed (in addition, the controls did not have conditions associated with NSAID use). Information on previous use of NSAID drugs (exposure) was then obtained in face-to-face interviews (which, by the way, might introduce bias – this type of bias may be prevented if prospectively gathered prescription data are available, although for NSAIDs the over-the-counter use is almost never registered on an individual basis).

As implied above, case-control studies are vulnerable to selection, information and confounding bias. For example, selection bias can occur when the cases enrolled in the study have a drug use profile that is not representative of all cases. For instance, selection bias occurs if cases are identified from hospital data and if people with the medical condition of interest are more likely to be hospitalized if they used the drug (than if they did not). Selection bias may also occur by selective nonparticipation in the study, or when controls enrolled in a study have a drug use profile that differs from that of the 'sample study base' (Fig. 12.3). Selection bias can then be minimized if controls are selected from the same source population (study base) as the cases.[36,37]

Since the exposure information in case-control studies is frequently obtained retrospectively-through medical records, interviews, and self-administered questionnaires, case-control studies are often subject to information bias. Most information bias pertains to recall and measurement bias. Recall bias may occur, for example, when interviewed cases remember more details about drug use than non-cases. The use of electronic pharmacy databases, with complete information about

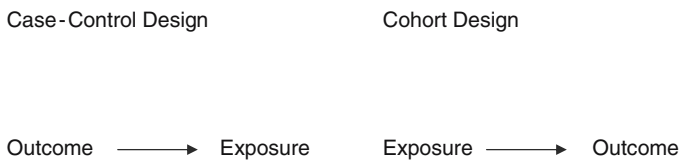



**Fig. 12.2** Direction of exposure and outcome in case-control and cohort designs

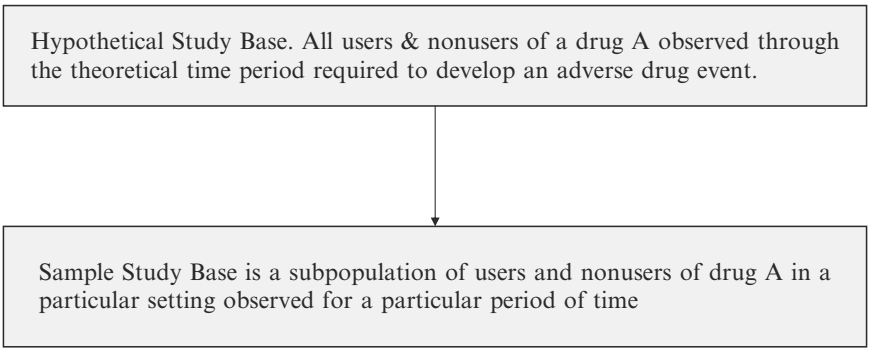



**Fig. 12.3** Study base and sample study base

drug exposure, could reduce this type of bias. Finally, an example of measurement or diagnostic bias occurs when researchers partly base the diagnosis of interpretation of the diagnosis on knowledge of the exposure status of the study subjects.

## Cohort Studies

Recall, that in cohort studies, participants are recruited based on the exposure and they are followed up over time while studying differences in their outcome. In PE cohort studies, users of a drug are compared to nonusers or users of other drugs with respect to rate or risk of an outcome. PE cohort studies are particularly efficient for rarely used drugs, or when there are multiple outcomes from a single exposure. The cohort study design then allows for establishing a temporal relationship between the exposure and the outcome because drug use precedes the onset of the outcome. In cohort studies, selection bias is generally less than in case-control designs. Selection bias is less likely, for example, when the drug use profile of the sample study base is similar to that of subjects enrolled in the study.

The disadvantages of cohort studies include the need for large number of subjects (unless the outcome is common, cohort studies are potentially uninformative for rare outcomes – especially those which require a long observation period); they are generally more expensive than other designs, particularly if active data collection is needed. In addition, they are vulnerable to bias if a high number of participants are lost during the follow-up (high drop-out rate). Finally, for some retrospective cohort studies, information about confounding factors might be limited or unavailable. With retrospective cohort studies, for example, the study population is frequently dynamic because the amount of time during which a subject is observed varies from subject to subject. PE retrospective cohort studies are frequently performed with information from automated databases with reimbursement or health care information (e.g. Veterans Administration database, Saskatchewan database, PHARMO database).

A special bias exists with cohort studies, the immortal time bias, which can occur when, as a result of the exposure definition, a subject, cannot incur the outcome event of interest during the follow up. For example, if an exposure is defined as the first prescription of drug 'A', and the outcome is death, the period of time from the calendar date to the first prescription where the outcome does not occur is the immortal time bias (red oval in Fig. 12.4). If during that period, the outcome occurs (e.g. death), then the subject won't be classified as part of the study group, rather, that subject will be part of the control group. This type of bias was described in the seventies when investigators compared the survival time of individuals receiving a heart transplant (study group) vs. those who were candidates but did not receive the transplant (control group). They found longer survival in the study group.[38,39] A reanalysis of data demonstrated that there was a waiting time from diagnosis of cardiac disease to the heart transplant, where patients were 'immortal' because if they died before the heart transplant, they were part of the control group.[40] This concept was adopted in pharmacoepidemiology research and since then, many published studies have been described with this type of bias.[41–46] (Fig. 12.4).

As prior mentioned, the consequence of this immortal time bias is the spurious appearance of a better outcome in the study group such as lower death rates. In other words, there is an underestimation of person-time without a drug treatment leading to an overestimation of a treatment effect.[47] One of the techniques to avoid immortal time bias is time-dependent drug exposure analysis.[48]

## Hybrid Studies

In PE research, hybrid designs are commonly used to study drug effects and drug safety. These designs combine several standard epidemiologic designs with resulting increased efficiency. In these studies, cases are selected on the basis of the outcome; and, drug use is compared with the drug use of several different types of

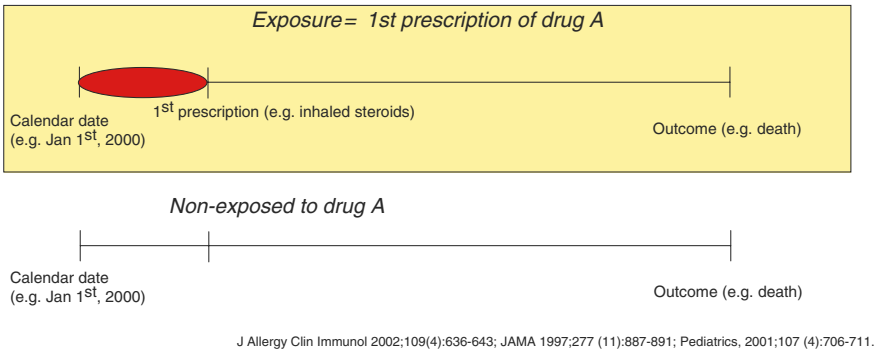



**Fig. 12.4** Immortal time bias in exposed (study) and non-exposed (control) groups

**Table 12.2** Differences in comparison groups for some of the PE hybrid designs

| Design | Control group |
|---|---|
| Nested case-control | Subjects in the same cohort, without the case condition |
| Case-cohort | A sample of the cohort at baseline (may include later cases) |
| Case-crossover | Cases, at an earlier time period |
| Case-time-control | Cases, at an earlier time period but time effect is considered |

comparison groups (see Table 12.2). These designs include: nested-case control studies, case-cohort design, case-crossover design and, case-time-control design.

## Nested Case-Control Studies

Recall that a nested case-control study refers to a case-control study which is nested in a cohort study or RCT. In PE, nested case-control studies, a defined population is followed for a period of time until a number of incident cases of a disease or an adverse drug reaction is identified. If the case-control study is nested in a cohort with prospectively gathered data on drug use, recall bias is no longer a problem. In PE as in other clinical research, nested case-control studies are used when the outcome is rare or the outcome has long induction time and latency. Frequently, this type of design is used when there is the need to use stored biological samples and additional information on drug use and confounders are needed. When it is inefficient to collect the aforementioned data for the complete cohort, (a common occurrence) a nested case-control study is desirable.

## Case-Cohort Studies

Recall that this type of study is similar to a nested case-control design, except the exposure and covariate information is collected from all cases, whereas controls are a random representative sample selected from the original cohort.[49,50] Case-cohort studies are recommended in the presence of rare outcomes or when the outcome has a long induction time and latency, but especially when the exposure is rare (if the exposure in controls is common, a case-control study is preferable). In PE case-cohort studies, the proportion of drug use in cases is compared to the proportion of drug use in the reference cohort (which may include cases). An example of the use of this design was to evaluate the association between immunosuppressive therapy (cyclophosphamide, azathioprine and methotrexate) and haematological changes in lung cancer, in patients with systemic lupus erythematosus (this was based on a lupus erythematosus cohort from centers in North America, Europe and Asia, where exposure and covariate information for all cases was collected). Cases were

defined as SLE, with invasive cancers discovered at each center after entry into the lupus cohort; and, the index time for each risk set was the date of the case's cancer occurrence. Controls were obtained from a random sample of the cohort (10% of the full cohort) and they represented cancer free patients up to the index time. Authors found that immunosuppressive therapy may contribute to an increased risk of hematological malignancies.[51]

## Case-Crossover Studies

Recall that the case-crossover design was proposed by Maclure, and in this design only cases that have experienced an outcome are considered. In that way, each case contributes one case window and one or more control windows at various time periods, and for the same patient. In other words, control subjects are the same as cases, just at an earlier time, so cases serve as own controls (see Chapter 4).[52,53] This type of design is particularly useful when a disease does not vary over time and when exposures are transient, brief and acute.[52,54] The case-crossover design contributes to the elimination of control selection bias and avoids difficulties in selecting and enrolling controls. However, case crossover designs are not suitable for studying chronic conditions.[55] In PE, case-crossover studies might compare the odds of drug use at a time close to onset of a medical condition compared with odds at an earlier time (Fig. 12.5).

Case-crossover designs have been used to assess the acute risks of vehicular accidents associated with the use of benzodiazepines[56] and also to study changes in medication use associated with epilepsy-related hospitalization. In this latter study, Handoko, et al. used the PHARMO database from 1998 to 2002. For each patient, changes in medication in a 28-day window before hospitalization, were compared with changes in four earlier 28-day windows; and, pattern of drug use, dosages, and interaction with medications were analyzed. Investigators found that patients starting with three or more new non antiepileptic drugs had a five times higher risk of epilepsy-related hospitalization.[57] In case-crossover designs, conditional logistic

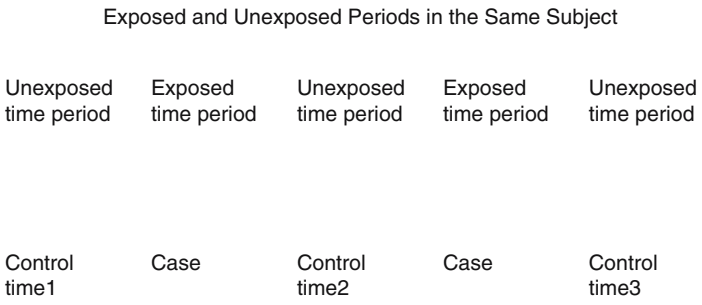



**Fig. 12.5** Case-crossover design

regression analysis is classically used to assess the association between events and exposure.[58,59]

## Case-Time-Control Studies

The case-time control design was proposed by Suissa[60] to control for confounding by indication. In this design subjects from a conventional case-control design are used as their own controls. This design is an extension of the case-crossover design but it takes into account the time effect, particularly the variation in the drug use over time. This type of design is recommended when an exposure varies over time and when there are two or more points measured at different times, and it is expected to be able to separate the drug effect from the disease severity. Something to consider is that the same precautions used in case-crossover designs should also be taken into account in case-time-control designs, and the exposures of control subjects must be measured at the same points in calendar time as their cases.

## Biases in PE

In PE, a special type of bias (confounding by indication) occurs when those who receive the drug have an inherently different prognosis from those who do not receive the drug. If the indication for treatment is an independent risk factor for the study outcome, the association of this indication with the prescribed drug may cause confounding by indication. A variant of confounding by indication (confounding by severity) may occur if a drug is prescribed selectively to patients with specific disease severity profiles.[61] Some hybrid designs and statistical techniques have been proposed to control for confounding by indication. In terms of statistical techniques, it has been proposed that one use multivariable model risk adjustment, propensity score risk adjustment, propensity-based matching and instrumental variable analysis to control for confounding by indication. Multivariable model risk adjustment is a conventional modeling approach that incorporates all known confounders into the model. Controlling for those covariates produces a risk-adjusted treatment effect and removes overt bias due to those factors.[62]

Propensity score risk adjustment is a technique used to adjust for nonrandom treatment assignment. It is a conditional probability of assignment to a particular treatment given a set of observed patient-level characteristics.[63,64] In this technique, a score is developed for each subject based on a prediction equation and the subject's value of each variable is included in the prediction equation,[65] and it is a scalar summary of all observed confounders. Within propensity score strata, covariates in treated and non-treated groups are similarly distributed, so the stratification using propensity score strata is claimed to remove more than 90% of the overt bias due to the covariates used to estimate the score.[66,67] Unknown biases can be partially

removed only if they are correlated with covariates already measured and included in the model to compute the score.[68–70]

Instrumental variable analysis is an econometric method used to remove the effects of hidden bias in observational studies.[71,72] Instrumental variables are highly correlated with treatment and they do not independently affect the outcome. Therefore, they are not associated with patient health status. Instrumental variable analysis compared groups of patients that differ in likelihood of receiving a drug.[73]

## Summary

In pharmacoepidemiology research as in for traditional research, the selection of an appropriate study design requires the consideration of various factors such as the frequency of the exposure and outcome, and the population under study. Investigators frequently need to weigh the choice of a study design with the quality of information collected along with its associated costs. In fact, new pharmacoepidemiologic designs are being developed to improve study efficiency.

Pharmacoepidemiology is not a new discipline, but it is currently recognized as one of the most challenging areas in research, and many techniques and methods are being tested to confront those challenges. Pharmacovigilance (see Chapter 5) as a part of pharmacoepidemiology is of great interest for decision makers, researchers, providers, manufacturers and the public, because of concerns about drug safety. Therefore, we should expect in the future, the development of new methods to assess the risk/benefit ratios of medications.

## References

1. Strom B, Kimmel S. Textbook of Pharmacoepidemiology. Hoboken, NJ: Wiley; 2006.
2. Miller, JL. Troglitazone withdrawn from market. Am J Health Syst Pharm. May 1, 2000; 57(9):834.
3. Gale EA. Lessons from the glitazones: a story of drug development. Lancet. June 9, 2001; 357(9271):1870–1875.
4. Scheen AJ. Thiazolidinediones and liver toxicity. Diabetes Metab. June 2001; 27(3):305–313.
5. Glessner MR, Heller DA. Changes in related drug class utilization after market withdrawal of cisapride. *Am J Manag Care*. Mar 2002; 8(3):243–250.
6. Griffin JP. Prepulsid withdrawn from UK & US markets. Adverse Drug React Toxicol Rev. Aug 2000; 19(3):177.
7. Graham DJ, Staffa JA, Shatin D, et al. Incidence of hospitalized rhabdomyolysis in patients treated with lipid-lowering drugs. *JAMA*. Dec 1, 2004; 292(21):2585–2590.
8. Piorkowski JD, Jr. Bayer's response to "potential for conflict of interest in the evaluation of suspected adverse drug reactions: use of cerivastatin and risk of rhabdomyolysis". *JAMA*. Dec 1, 2004; 292(21):2655–2657; discussion 2658–2659.

9. Strom BL. Potential for conflict of interest in the evaluation of suspected adverse drug reactions: a counterpoint. *JAMA*. Dec 1, 2004; 292(21):2643–2646.
10. Wooltorton E. Bayer pulls cerivastatin (Baycol) from market. *CMAJ*. Sept 4, 2001; 165(5):632.
11. Juni P, Nartey L, Reichenbach S, Sterchi R, Dieppe PA, Egger M. Risk of cardiovascular events and rofecoxib: cumulative meta-analysis. *Lancet*. Dec 4–10, 2004; 364(9450):2021–2029.
12. Sibbald B. Rofecoxib (Vioxx) voluntarily withdrawn from market. *CMAJ*. Oct 26, 2004; 171(9):1027–1028.
13. Wong M, Chowienczyk P, Kirkham B. Cardiovascular issues of COX-2 inhibitors and NSAIDs. *Aust Fam Physician*. Nov 2005; 34(11):945–948.
14. Antoniou K, Malamas M, Drosos AA. Clinical pharmacology of celecoxib, a COX-2 selective inhibitor. *Expert Opin Pharmacother*. Aug 2007; 8(11):1719–1732.
15. Sun SX, Lee KY, Bertram CT, Goldstein JL. Withdrawal of COX-2 selective inhibitors rofecoxib and valdecoxib: impact on NSAID and gastroprotective drug prescribing and utilization. *Curr Med Res Opin*. Aug 2007; 23(8):1859–1866.
16. Prentice RL, Langer R, Stefanick ML, et al. Combined postmenopausal hormone therapy and cardiovascular disease: toward resolving the discrepancy between observational studies and the Women's Health Initiative clinical trial. *Am J Epidemiol*. Sept 1, 2005; 162(5):404–414.
17. Dubach UC, Rosner B, Sturmer T. An epidemiologic study of abuse of analgesic drugs. Effects of phenacetin and salicylate on mortality and cardiovascular morbidity (1968 to 1987). *N Engl J Med*. Jan 17, 1991; 324(3):155–160.
18. Elseviers MM, De Broe ME. A long-term prospective controlled study of analgesic abuse in Belgium. *Kidney Int*. Dec 1995; 48(6):1912–1919.
19. Morlans M, Laporte JR, Vidal X, Cabeza D, Stolley PD. End-stage renal disease and nonnarcotic analgesics: a case-control study. *Br J Clin Pharmacol*. Nov 1990; 30(5):717–723.
20. Murray TG, Stolley PD, Anthony JC, Schinnar R, Hepler-Smith E, Jeffreys JL. Epidemiologic study of regular analgesic use and end-stage renal disease. *Arch Intern Med*. Sept 1983; 143(9):1687–1693.
21. Perneger TV, Whelton PK, Klag MJ. Risk of kidney failure associated with the use of acetaminophen, aspirin, and nonsteroidal antiinflammatory drugs. *N Engl J Med*. Dec 22, 1994; 331(25):1675–1679.
22. Piotrow PT, Kincaid DL, Rani M, Lewis G. *Communication for Social Change*. Baltimore, MD: The Rockefeller Foundation and Johns Hopkins Center for Communication Programs; 2002.
23. Major outcomes in high-risk hypertensive patients randomized to angiotensin-converting enzyme inhibitor or calcium channel blocker vs diuretic: the Antihypertensive and Lipid-Lowering Treatment to Prevent Heart Attack Trial (ALLHAT). *JAMA*. Dec 18, 2002; 288(23):2981–2997.
24. Pilote L, Abrahamowicz M, Rodrigues E, Eisenberg MJ, Rahme E. Mortality rates in elderly patients who take different angiotensin-converting enzyme inhibitors after acute myocardial infarction: a class effect? *Ann Intern Med*. July 20, 2004; 141(2):102–112.
25. Schneider LS, Tariot PN, Dagerman KS, et al. Effectiveness of atypical antipsychotic drugs in patients with Alzheimer's disease. *N Engl J Med*. Oct 12, 2006; 355(15):1525–1538.
26. Schneeweiss S. Developments in post-marketing comparative effectiveness research. *Clin Pharmacol Ther*. Aug 2007; 82(2):143–156.
27. Mellin GW, Katzenstein M. The saga of thalidomide. Neuropathy to embryopathy, with case reports of congenital anomalies. *N Engl J Med*. Dec 13, 1962; 267:1238–1244 concl.
28. Food and Drug Administration. Medwatch Website. http://www.fda/gov/medwatch. Accessed Aug 20, 2007.
29. Humphries TJ, Myerson RM, Gifford LM, et al. A unique postmarket outpatient surveillance program of cimetidine: report on phase II and final summary. *Am J Gastroenterol*. Aug 1984; 79(8):593–596.
30. Stricker BH, Blok AP, Claas FH, Van Parys GE, Desmet VJ. Hepatic injury associated with the use of nitrofurans: a clinicopathological study of 52 reported cases. *Hepatology*. May–June 1988; 8(3):599–606.

31. Martin A, Leslie D. Trends in psychotropic medication costs for children and adolescents, 1997–2000. *Arch Pediatr Adolesc Med*. Oct 2003; 157(10):997–1004.
32. Williams P, Bellantuono C, Fiorio R, Tansella M. Psychotropic drug use in Italy: national trends and regional differences. *Psychol Med*. Nov 1986; 16(4):841–850.
33. Paulose-Ram R, Hirsch R, Dillon C, Losonczy K, Cooper M, Ostchega Y. Prescription and non-prescription analgesic use among the US adult population: results from the third National Health and Nutrition Examination Survey (NHANES III). *Pharmacoepidemiol Drug Saf*. June 2003; 12(4):315–326.
34. Paulose-Ram R, Jonas BS, Orwig D, Safran MA. Prescription psychotropic medication use among the U.S. adult population: results from the third National Health and Nutrition Examination Survey, 1988–1994. *J Clin Epidemiol*. Mar 2004; 57(3):309–317.
35. Strom B. *Study Designs Available for Pharmacoepidemiology Studies. Pharmacoepidemiology*. 3rd. ed: Wiley; 2000.
36. Risks of agranulocytosis and aplastic anemia. A first report of their relation to drug use with special reference to analgesics. The International Agranulocytosis and Aplastic Anemia Study. *JAMA*. Oct 3, 1986; 256(13):1749–1757.
37. Wilcox AJ, Baird DD, Weinberg CR, Hornsby PP, Herbst AL. Fertility in men exposed prenatally to diethylstilbestrol. *N Engl J Med*. May 25, 1995; 332(21):1411–1416.
38. Clark DA, Stinson EB, Griepp RB, Schroeder JS, Shumway NE, Harrison DC. Cardiac transplantation in man. VI. Prognosis of patients selected for cardiac transplantation. *Ann Intern Med*. July 1971; 75(1):15–21.
39. Messmer BJ, Nora JJ, Leachman RD, Cooley DA. Survival-times after cardiac allografts. *Lancet*. May 10, 1969; 1(7602):954–956.
40. Gail MH. Does cardiac transplantation prolong life? A reassessment. *Ann Intern Med*. May 1972; 76(5):815–817.
41. Donahue JG, Weiss ST, Livingston JM, Goetsch MA, Greineder DK, Platt R. Inhaled steroids and the risk of hospitalization for asthma. *JAMA*. Mar 19, 1997; 277(11):887–891.
42. Fan VS, Bryson CL, Curtis JR, et al. Inhaled corticosteroids in chronic obstructive pulmonary disease and risk of death and hospitalization: time-dependent analysis. *Am J Respir Crit Care Med*. Dec 15, 2003; 168(12):1488–1494.
43. Kiri VA, Vestbo J, Pride NB, Soriano JB. Inhaled steroids and mortality in COPD: bias from unaccounted immortal time. *Eur Respir J*. July 2004; 24(1):190–191; author reply 191–192.
44. Mamdani M, Rochon P, Juurlink DN, et al. Effect of selective cyclooxygenase 2 inhibitors and naproxen on short-term risk of acute myocardial infarction in the elderly. *Arch Intern Med*. Feb 24, 2003; 163(4):481–486.
45. Suissa S. Observational studies of inhaled corticosteroids in chronic obstructive pulmonary disease: misconstrued immortal time bias. *Am J Respir Crit Care Med*. Feb 15, 2006; 173(4):464; author reply 464–465.
46. Suissa S. Immortal time bias in observational studies of drug effects. *Pharmacoepidemiol Drug Saf*. Mar 2007; 16(3):241–249.
47. Suissa S. Effectiveness of inhaled corticosteroids in chronic obstructive pulmonary disease: immortal time bias in observational studies. *Am J Respir Crit Care Med*. July 1, 2003; 168(1):49–53.
48. Clayton D, Hills M, eds. *Time-Varying Explanatory Variables. Statistical models in epidemiology*. Oxford: Oxford University Press; 1993:307–318.
49. Sato T. Risk ratio estimation in case-cohort studies. *Environ Health Perspect*. 1994; 102(8):53–56.
50. van der Klauw MM, Stricker BH, Herings RM, Cost WS, Valkenburg HA, Wilson JH. A population based case-cohort study of drug-induced anaphylaxis. *Br J Clin Pharmacol*. Apr 1993; 35(4):400–408.
51. Bernatsky S, Boivin JF, Joseph L, et al. The relationship between cancer and medication exposures in systemic lupus erythematosus: a case-cohort study. *Ann Rheum Dis*. June 1, 2007.
52. Maclure M. The case-crossover design: a method for studying transient effects on the risk of acute events. *Am J Epidemiol*. Jan 15, 1991; 133(2):144–153.

53. Maclure M, Mittleman MA. Should we use a case-crossover design? *Annu Rev Public Health*. 2000; 21:193–221.
54. Marshall RJ, Jackson RT. Analysis of case-crossover designs. *Stat Med*. Dec 30, 1993; 12(24):2333–2341.
55. Donnan PT, Wang J. The case-crossover and case-time-control designs in pharmacoepidemiology. *Pharmacoepidemiol Drug Saf*. May 2001; 10(3):259–262.
56. Barbone F, McMahon AD, Davey PG, et al. Association of road-traffic accidents with benzodiazepine use. *Lancet*. Oct 24, 1998; 352(9137):1331–1336.
57. Handoko KB, Zwart-van Rijkom JE, Hermens WA, Souverein PC, Egberts TC. Changes in medication associated with epilepsy-related hospitalisation: a case-crossover study. *Pharmacoepidemiol Drug Saf*. Feb 2007; 16(2):189–196.
58. Greenland S. A unified approach to the analysis of case-distribution (case-only) studies. *Stat Med*. Jan 15 1999; 18(1):1–15.
59. Scneeweiss S, Sturner TMM. Case-crossover and case = time-control designs as alternatives in pharmacoepidemiologic research. *Pharmacoepidemiol Drug Saf*. 1997; 6(suppl 3):S51–59.
60. Suissa S. The case-time-control design. *Epidemiology*. May 1995; 6(3):248–253.
61. Salas M, Hofman A, Stricker BH. Confounding by indication: an example of variation in the use of epidemiologic terminology. *Am J Epidemiol*. June 1, 1999; 149(11):981–983.
62. Stukel TA, Fisher ES, Wennberg DE, Alter DA, Gottlieb DJ, Vermeulen MJ. Analysis of observational studies in the presence of treatment selection bias: effects of invasive cardiac management on AMI survival using propensity score and instrumental variable methods. *JAMA*. Jan 17, 2007; 297(3):278–285.
63. D'Agostino RB, Jr. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Stat Med*. Oct 15, 1998; 17(19):2265–2281.
64. Morant SV, Pettitt D, MacDonald TM, Burke TA, Goldstein JL. Application of a propensity score to adjust for channelling bias with NSAIDs. *Pharmacoepidemiol Drug Saf*. June 2004; 13(6):345–353.
65. Ahmed A, Husain A, Love TE, et al. Heart failure, chronic diuretic use, and increase in mortality and hospitalization: an observational study using propensity score methods. *Eur Heart J*. June 2006; 27(12):1431–1439.
66. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983; 70(41–55).
67. Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. *J AM Stat Assoc*. 1984; 79:516–524.
68. Austin PC, Mamdani MM, Stukel TA, Anderson GM, Tu JV. The use of the propensity score for estimating treatment effects: administrative versus clinical data. *Stat Med*. May 30, 2005; 24(10):1563–1578.
69. Braitman LE, Rosenbaum PR. Rare outcomes, common treatments: analytic strategies using propensity scores. *Ann Intern Med*. Oct 15, 2002; 137(8):693–695.
70. Harrell FE. *Regression Modeling Strategies with Applications to Linear Models, Logistic Regression and Survival Analysis*. New York: Springer; 2001.
71. McClellan M, McNeil BJ, Newhouse JP. Does more intensive treatment of acute myocardial infarction in the elderly reduce mortality? Analysis using instrumental variables. *JAMA*. Sept 21, 1994; 272(11):859–866.
72. Newhouse JP, McClellan M. Econometrics in outcomes research: the use of instrumental variables. *Annu Rev Public Health*. 1998; 19:17–34.
73. Harris KM, Remler DK. Who is the marginal patient? Understanding instrumental variables estimates of treatment effects. *Health Serv Res*. Dec 1998; 33(5 Pt 1):1337–1360.

# Chapter 13
# Implementation Research: Beyond the Traditional Randomized Controlled Trial

**Amanda H. Salanitro, Carlos A. Estrada, and Jeroan J. Allison**

**Abstract** Implementation research is a new scientific discipline emerging from the recognition that the public does not derive sufficient or rapid benefit from advances in the health sciences.[1,2] One often-quoted estimate claims that it takes an average of 17 years for even well-established clinical knowledge to be fully adopted into routine practice.[3] In this chapter, we will discuss particular barriers to evidence implementation, present tools for implementation research, and provide a framework for designing implementation research studies, emphasizing the randomized trial. The reader is advised that this chapter only provides a basic introduction to several concepts for which new approaches are rapidly emerging. Therefore, our goal is to stimulate interest and promote additional in-depth learning for those who wish to develop new implementation research projects or better understand this exciting field.

## Introduction

### *Overview and Definition of Implementation Research*

Implementation research is a new scientific discipline emerging from the recognition that the public does not derive sufficient or rapid benefit from advances in the health sciences.[1,2] One often-quoted estimate claims that it takes an average of 17 years for even well-established clinical knowledge to be fully adopted into routine practice.[3] For example, in 2000, only one-third of patients with coronary artery disease received aspirin when no contraindications to its use were present.[2] In 2003, a landmark study by McGlynn et al. estimated that the American public was only receiving about 55% of recommended care.[4]

In this setting where adoption lags evidence Rubenstein and Pugh defined implementation research as:

S.P. Glasser (ed.), *Essentials of Clinical Research*,

> …scientific investigations that support movement of evidence-based, effective health care approaches (e.g., as embodied in guidelines) from the clinical knowledge base into routine use. These investigations form the basis for health care implementation science. Implementation science consists of a body of knowledge on methods to promote the systematic uptake of new or underused scientific findings into the usual activities of regional and national health care and community organizations, including individual practice sites.[5]

More recently, Kiefe et al. updated the definition of implementation research as:

> the scientific study of methods to promote the rapid uptake of research findings, and hence improve the health of individuals and populations.[6]

Finally, the definition of implementation research may be expanded to encompass work that promotes patient safety and eliminates racial and ethnic disparities in health care.

Forming an important core of implementation research, disparities research identifies and closes gaps in health care based on race/ethnicity and socioeconomic position through culturally-appropriate interventions for patients, clinicians, health care systems, and populations.[7–10] Under-represented populations make up a significant portion of the U.S. population, shoulder a disproportionate burden of disease, and receive inadequate care.[11] In addition, these groups have often been marginalized from traditional clinical research studies for several reasons. Researchers and participants often do not share common cultural perspectives, which may lead to lack of trust.[12] Lack of resources, such as low levels of income, education, health insurance, social integration, and health literacy, may preclude participation in research studies.[12]

Gaps in health care, such as those described above for vulnerable populations, may be classified as "errors of omission", or failure to provide necessary care.[13] In addition to addressing errors of omission, implementation research seeks to understand and resolve errors of commission, such as the delivery of unnecessary or inappropriate care which causes harm. In 1999, a landmark report from the Institute of Medicine drew attention to patient safety and the concept of preventable injury.[14] Studies of patient safety have focused on "medical error resulting in an inappropriate increased risk of iatrogenic adverse event(s) from receiving too much or hazardous treatment (overuse or misuse)".[13]

For example, inappropriate antibiotic use may promote microbial resistance and cause unnecessary adverse events. Therefore, an inter-governmental task force initiated a campaign in 1999 to promote appropriate prescribing of antibiotics for acute respiratory infections (ARIs).[15] In 1997, physicians prescribed antibiotics for 66% of patients diagnosed with acute bronchitis. In 2001, based on data from randomized controlled trials (RCTs) demonstrating no benefit, guidelines recommended against antibiotic use for acute bronchitis.[16,17] Although overall antibiotic use for ARIs declined between 1995–2002, use of broad-spectrum antibiotic prescriptions for ARIs increased.[18] A more recent implementation research project successfully used a multidimensional intervention in emergency departments to decrease antibiotic prescribing.[19]

In response to what may be perceived as overwhelming evidence that thousands of lives are lost each year from errors of omission and commission, there have been strong national calls for health systems, hospitals, and physicians to adopt new approaches for moving evidence into practice.[20,21] While many techniques have been promoted, such as computer-based order entry and performance-based reimbursement, rigorous supporting evidence is often lacking.

Even though our understanding of implementation science is incomplete, local clinicians and health systems must obviously strive to improve the quality of care for every patient. This practical consideration means that certain local decisions must be based on combinations of incomplete empiric evidence, personal experience, anecdotes, and supposition. As with the clinician caring for the individual patient, every decision about local implementation cannot be guided by data from a randomized trial.[23,22] However, a stronger evidence base is needed to inform widespread implementation efforts. Widespread implementation beyond evidence raises concern about unintended consequences and opportunity costs from public resources wrongly expended on ineffective interventions.[22]

To generate this evidence base, implementation researchers use a variety of techniques, ranging from qualitative exploration to the controlled, group-randomized trial. Brennan et al. described the need to better understand the 'basic science' of health care quality by applying methods from such fields as social, cognitive, and organizational psychology.[24] Recently, Berwick emphasized the importance of understanding the mechanism and context through which implementation techniques exert their potential effects within complex human systems.[25] Berwick cautioned that important lessons may be lost through aggregation and rigorous scientific experimentation, challenging the implementation research community to reconsider the basic concept of evidence, itself. Interventions for translating evidence into practice must operate in complex, poorly understood environments with multiple interacting components which may not be easily reducible to a clean, scientific formula. Therefore, we later present situational analysis as a framing device for implementation research. Nonetheless, in keeping with the theme of this book, we mainly focus on the randomized trial as one of the many critical tools for implementation research.

In summary, implementation research is an emerging body of scientific work seeking to close the gap between knowledge generated from the health sciences and routine practice, ultimately improving patient and population health outcomes. Implementation research, which encompasses the patient, clinician, health system, and community, may promote the use of needed services or the avoidance of unneeded services. Implementation research often focuses on patients who are vulnerable because of race/ethnicity or socioeconomic position. By its very nature implementation research is inter-disciplinary.

In this chapter, we will discuss particular barriers to evidence implementation, present tools for implementation research, and provide a framework for designing implementation research studies, emphasizing the randomized trial. The reader is advised that this chapter only provides a basic introduction to several concepts for

which new approaches are rapidly emerging. Therefore, our goal is to stimulate interest and promote additional in-depth learning for those who wish to develop new implementation research projects or better understand this exciting field.

## Overcoming Barriers to Evidence Implementation

Although the conceptual basis for moving evidence into practice has not been fully developed, a solid grounding in relevant theory may be useful to those designing new implementation research projects.[26] Many conceptual models have been developed in other settings and subsequently adapted for translating evidence into practice.[27] For example, implementation researchers frequently apply Roger's theory describing innovation diffusion. Rogers proposed three clusters of influence on the rapidity of innovation uptake: (1) perceived advantages of the innovation; (2) the classification of new technology users according to rapidity of uptake; and, (3) contextual factors.[28] First, potential users are unlikely to adopt an innovation that is perceived to be complex and inconsistent with their needs and cultural norms. Second, rapidity of innovation uptake often follows a sigmoid-shaped curve, with an initial period of slow uptake led by the 'innovators.' Next follows a more rapid period of uptake led by the early adopters, or 'opinion leaders.' During the last adoption phase, the rate of diffusion again slows as the few remaining 'laggards' or traditionalists adopt the innovation. Finally, contextual or environmental factors such as organizational culture exert a profound impact on innovation adoption, a concept which is explored in more detail in the following sections of this chapter.

Consistent with the model proposed by Rogers, multiple barriers often work synergistically to hinder the translation of evidence into practice.[29] Interventions often require significant time, money, and staffing. Implementation sites may experience difficulties in implementation from limited resources, competing demands, and entrenched practices. The intervention may have been developed and tested under circumstances that differ from those at the planned implementation site. The implementation team may not adequately understand the environmental characteristics postulated by Roger's diffusion theory as critical to the adoption of innovation. Because of such concerns a thorough environmental analysis is needed prior to widespread implementation efforts.[29]

Building upon models proposed by Sung et al.[30] and Rubenstein et al.,[5] Fig. 13.1 depicts the translational barriers implementation research seeks to overcome. The first translational roadblock lies between basic science knowledge and clinical trials. The second roadblock involves translation of knowledge gained from clinical trials into meaningful clinical guidance, which often takes the form of evidence-based guidelines.

The third roadblock occurs between current clinical knowledge and routine practice, carrying important implications for individual practitioners, health care systems, communities, and populations. Given the expansive nature of this third roadblock, a multifaceted armamentarium of tools is required. One tool, industrial-

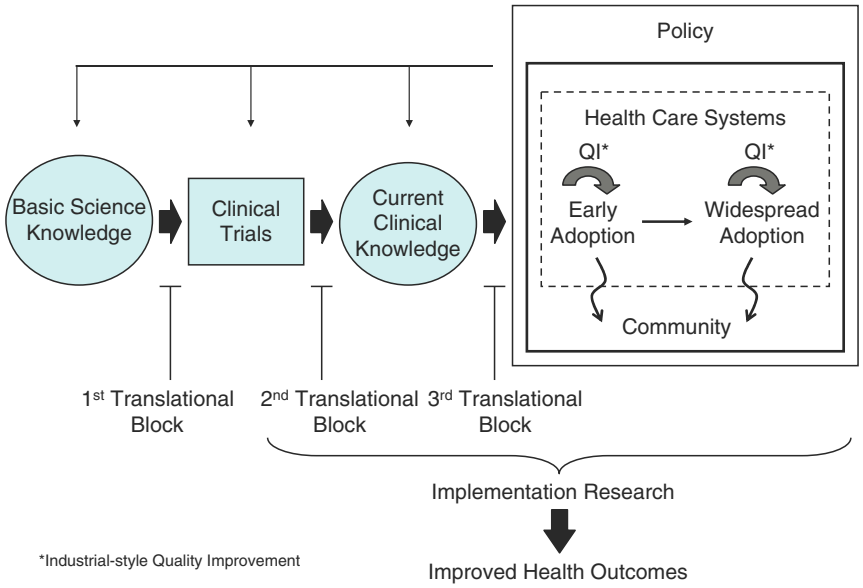



**Fig. 13.1** Translational blocks targeted by Implementation Research

style quality improvement, described below in more detail, operates at the level of the clinical microsystem, the smallest, front-line functional unit that actually delivers care to a patient.[31] Clinical microsystems consist of complex adaptive relationships among patients, providers, support staff, technology, and processes of care. To achieve sustainable success, researchers seeking to overcome this third translational barrier need to be effective advocates for changes in local and governmental health policy. Finally, implementation research may inform clinical trials and basic science.

To promote the spectrum of research depicted in Fig. 13.1, the 2003 NIH Roadmap acknowledges translational research as an important discipline.[32] In fact, several branches of the NIH now have open funding opportunities for implementation research. The integration of research findings from the molecular to the population level is a priority. The Roadmap seeks to join communities and interdisciplinary academic research centers to translate new discoveries into improved population health.[33]

## Implementation Research Tools

The tools used to translate clinical evidence into routine practice are varied, and no single tool or combination of tools has proven sufficient or completely effective. Furthermore, it may not be the tool itself but how it is implemented in a system that drives change.[34]

In fact, this lack of complete effectiveness spurs implementation research to develop innovative adaptations or combinations of currently available tools.[35]

Below, we provide an overview of available tools, which are intended as basic building blocks for future implementation research projects. Although different classification systems have been proposed,[36] we arranged these tools by their focus: on the patient, the community, the provider, and the healthcare organization. We acknowledge that this classification is somewhat arbitrary because several implementation tools overlap multiple categories.

## *Patient-Based Implementation Tools*

A growing body of evidence suggests that patients may be successfully 'activated' to improve their own care. For example, a medical assistant may review the medical record with the patient and encourage the patient to ask questions at an upcoming visit with the physician. Patients exposed to such programs had better health outcomes, such as improved glycemic control for those with diabetes.[37,38] In another study, a health maintenance reminder card presented by the patient to the physician at appointments significantly increased rates of influenza vaccination and cancer screening.[39]

Other interventions have taught disease-management and problem solving skills to improve chronic disease outcomes. Teaching patient self-management skills is more effective than passive patient education, and these skills have been shown to improve outcomes and reduce costs for patients with arthritis and asthma.[40] As part of the 'collaborative model,' self-management is encouraged through better interactions between the patient, physician, and health care team. The collaborative model includes: (1) identifying problems from the joint perspective of the patient and clinical care team; (2) targeting problems, setting appropriate goals, and developing action plans together; (3) continuing self-management training and support services for patients; (4) active follow up to reinforce the implementation of the care plan.[40]

## *Community-Based Implementation Tools*

The Community Health Advisor (CHA) model has been implemented throughout the world to deliver health messages, promote positive health behavior change, and facilitate access to the health care system.[41] Based on the CHA model, community members, usually without formal education in the health professions, undergo special training and certification. CHA interventions have been used to promote prevention and treatment for a large array of conditions, including cancer, asthma, cardiovascular disease, depression, and diabetes. CHA programs have also been developed to decrease youth violence and risky sexual behavior. CHA interventions

may be especially relevant for underserved populations and those living in rural areas. Although promising, CHA interventions often rely on volunteer workers who may be vulnerable to stress and burnout from work overload. Also, intense training and oversight is often required to assure the accuracy of the health messages being transmitted. A review by Swider found limited high-quality evidence that CHA interventions actually improve health outcomes. Swider also called for additional rigorous research on the effectiveness and underlying mechanisms through which CHA interventions work.[42] A more recent review commissioned by the Robert Wood Johnson Foundation found that specific CHA interventions may reduce health disparities, particularly for patients with hypertension and diabetes.[9]

## *Provider-Based Implementation Tools*

### Clinical Guidelines

Clinical guidelines have been defined as "systematically developed statements to assist practitioners' and patients' decisions about appropriate health care for specific clinical circumstances."[43] In a systemic review of implementation strategies spanning the last 30 years, Grimshaw et al. noted guideline dissemination efforts may lead to modest improvements in care.[44] However, guideline dissemination alone is not sufficient for implementation.[45]

For many clinical situations encountered today, thousands of evidence-based guidelines and practice recommendations have been published. Such sheer volume often precludes the individual practitioner from implementing all recommendations for every patient. As an example, Boyd et al. noted that if one were to treat a hypothetical 79 year old woman with diabetes, chronic obstructive pulmonary disease (COPD), hypertension, osteoporosis, and osteoarthritis, and follow all recommended guidelines for her multiple co-morbidities, the patient would require 12 medications at a cost of $406 per month.[46]

### Continuing Medical Education

Continuing medical education (CME), a requirement for ongoing medical licensure, has traditionally relied on text-based, didactic methods to promulgate clinical information. However, passive, text-based educational materials and formal CME conferences do not lead to measurable improvements in practice patterns.[47,48] Rather, CME using interactive techniques which actively engage physicians are more effective in improving practice patterns and patient outcomes.[49] Physicians who reflect on their own individual performance may identify areas for improvement and seek CME through multifaceted, self-directed learning opportunities. Modalities that promote active learning – such as case-based problem solving – have shown to produce modest improvements in clinical practice.[50]

With the advantages of being convenient, flexible, and inexpensive, the Internet has become a useful platform to reach a wider audience for interactive CME. Fordis et al. conducted a randomized controlled trial comparing live, small-group interactive CME workshops with Internet CME.[51] Both groups focused on cholesterol management. All physicians received didactic instruction, interactive cases with feedback, practice tools and resources, and access to expert advice. Knowledge scores for physicians in the Internet CME group increased more than scores for those in the live CME group. Additionally, the online CME group demonstrated a statistically significant improvement in appropriate drug treatment for high-risk patients. Success of the Internet CME may have been partially driven by the participants' ability to repeatedly return to the website for reinforcement and the ability to structure the learning experience to meet individual needs.

### Academic Detailing

Academic detailing relies on site visits to physicians' offices for intense relationship building and one-on-one information delivery. Important components for successful detailing include: (1) assessment of baseline knowledge and motivations for current behavior; (2) articulating clear objectives for education and behavior; (3) gaining credibility with ties to respected organizations through ongoing relationship building; (4) encouraging physicians to actively participate in educational interventions; (5) using graphic representations for educational materials; (6) focusing on a limited number of 'take-home' points; and, (7) supplying positive reinforcement for improved behaviors during follow up.[52] Representatives from pharmaceutical companies have effectively used academic detailing to boost product sales. In a systematic review, academic detailing yielded modest effects; however, significant resources were needed to sustain these projects.[44]

### Opinion Leaders

Several implementation programs have relied on influential colleagues. Opinion leader strategies may include using celebrities, employing people in leadership positions, and asking those doing front-line work to refer 'up the ladder.' Studies examining the effectiveness of opinion-leader strategies have produced both positive and negative findings, and the precise mechanism for how change is accomplished remains elusive.[53]

### Physician Audit and Feedback

The utility of audit and feedback hinges on developing credible data-driven summaries of how patient populations are being managed. In theory, such reports may prompt clinicians to reflect on their personal clinical practices and motivate

subsequent improvement. Performance feedback may focus on outcomes (such as percentage of patients with diabetes who have achieved glycemic control) or process (such as the percentage of patients with diabetes for whom the physician measured glycemic control). The credibility of performance feedback relies on the ability to capture the many clinical nuances that the physician must consider when delivering care to the individual patient. Because the difficulties in capturing these clinical nuances have not yet been completely surmounted, comparisons of performance to a data-driven, peer-based benchmark may be more appropriate than comparison to an arbitrary standard of perfect performance. Kiefe et al. found that feedback with peer-based benchmarks led to better quality of care, but other studies have reported mixed or modest results.[44,54]

## *Organization-Based Implementation Tools*

### Industrial-Style Quality Improvement

This type of improvement activity originated outside of health care and has acquired such labels as Total Quality Management (TQM) and Continuous Quality Improvement (CQI). These approaches make two fundamental assumptions: (a) that poor outcomes are attributable to system failures, rather than lack of individual effort or individual mistakes, and (b) achieving improvement and excellence, even in the absence of system failures, is possible through iterative cycles of planning, acting, and observing the results. In general, complex systems must have built-in redundancy to function well. If an individual makes a mistake at one point in the system, checks and balances built into other parts of the system may prevent an adverse event. However, as described in the example below, patient safety may be endangered by simultaneous failure of multiple system components, thus defeating built-in redundancy.

As a simple example, multiple mechanisms should be in place to ensure that incompatible blood products are not given to hospitalized patients. Delivery of the wrong blood type to a patient requires failure at multiple points, including preparation of the blood in the blood bank and administration of the blood by the nurse. Taking such a systems approach stands in stark contrast to blaming individuals, thereby avoiding low morale and reluctance to disclose mistakes.

Improvement activity usually proceeds through a series of ‘plan-do-study-act’ cycles. These cycles emphasize measuring the process of clinical care delivery at the level of the clinical microsystem, which has been previously described. Here, small amounts of data guide the initial improvement process. The process emphasizes small, continuous gains through repeated cycles and does not rely on the statistical significance of the measurements. Although many health care institutions have adopted such methodology based on compelling case studies, additional studies with high-quality experimental methods are still needed.[55]

### Systems Reengineering

Instead of incremental changes to clinical microsystems, major redesign of the entire system may be undertaken. For example, in the 1990s the Veterans' Health Administration (VHA) undertook a major reengineering of its health care system, focusing on the improved use of information technology (IT), the measurement and reporting of performance, and the integration of services.[56] By 2000, the VHA had made statistically significant improvements in nine areas, including preventive care, outpatient care (diabetes, hypertension, and depression), and inpatient care (acute myocardial infarction and congestive heart failure). Additionally, the VHA performed better than the fee-for-service Medicare system on 12 of 13 quality measures.[56] Because systems engineering requires changes on such a large scale, little evidence exists about its efficiency and effectiveness in yielding more improvements than smaller changes.[3]

### Computer-Based Systems

Computer-based systems target links in the process of care delivery that are most prone to human error. Such systems may provide clinical decision support by assisting the clinician with making a diagnosis, choosing among alternative treatments, or deciding upon a particular drug dosage. Other functions may include delivery of clinical reminders and computerized provider order entry (CPOE).[57] A systematic review documented improvements in time to therapeutic goals, decreases in toxic drug levels and adverse reactions, and shorter hospital stays.[58] However, adverse effects of computer-based systems have also been reported, including increased mortality rates, increased rates of adverse drug reactions, delays in medication administration, increased work load, and new types of errors.[59–62] These data illustrate that adverse drug reactions may be either increased or decreased after the introduction of computer-based systems. Therefore, computer-based systems should not be implemented without safeguards to prevent unintended consequences. We need more work to better understand how computer-based systems interact with human users and the complex health care environment and how these interactions affect quality, safety, and outcomes.

### Public Report Cards

Public reports on the quality of health care delivered by institutions are proliferating. For example, public reports may focus on risk-adjusted mortality after cardiac surgery or quality at long-term care facilities. In addition, such reports will probably be expanded to include physician groups and individual physicians. Public reports are often promoted under the assumption that the public will use them to

choose high-quality providers, thus better enabling a competitive 'medical marketplace.' However, this promise has yet to be realized. Although scant evidence links report cards to improved health care, report cards may have profound adverse effects: (1) physicians may avoid sicker patients to improve their ratings; (2) physicians may strive to meet the targeted rates for interventions even in situations where intervention is inappropriate; and, (3) physicians may ignore patient preferences and neglect clinical judgment.[63] Even worse, report cards may actually widen gaps in health disparities.[64]

### Pay-for-Performance (P4P)

Currently, there is mounting pressure to tie reimbursement for health care services to quality measurement. Although allowing market forces to freely operate through P4P reimbursement may seem logical, systematic reviews have not yielded conclusive results. Because not everything that is important is currently measured, linking reimbursement to measured quality may divert attention from important, but unmeasured aspects of care (i.e., 'spotlight' effect). As with public reporting, P4P may actually widen health disparities, although empiric data are lacking.

To date, evidence informing the effectiveness of P4P in improving the delivery of health care is limited. One study found that when implemented in physician practice groups, P4P produced improvements for those with higher baseline performance but had minimal effect on the lowest performers.[65] Glickman et al. found hospitals voluntarily participating in the P4P initiative for myocardial infarction did not show appreciable improvement.[66] A recent study found that hospitals participating in P4P and public reporting programs sponsored by the Centers for Medicare and Medicaid Services had slightly greater improvements in quality than those only participating in the public reporting program.[67] Several ongoing studies may soon deliver new insights about P4P.

### Advancing Implementation Science

Because the implementation science base is still emerging, researchers have at their disposal an array of tools which are variously effective depending upon the patient population and delivery setting. Moving beyond the tools described above, we need to develop innovative adaptations and approaches to bridge the gap between clinical knowledge and health care practice. We need to test the effectiveness of these new approaches with rigorous scientific methods to avoid adverse consequences from the wide-spread dissemination and adoption of unproven interventions.[22] Therefore, in the remainder of this chapter, we discuss the critical design elements for implementation randomized controlled trials, followed by an example of an implementation research study.

## Designing Implementation Research Studies

### *Overview of Implementation Research Study Design*

Multiple designs are available for implementation research projects. Somewhat analogous to the traditional clinical trial, randomized designs for implementation research allow causal inference and offer protection from measured and unmeasured confounding.[35] As described below in more detail, such designs include an active intervention, random allocation to a comparison or intervention group, and blinded assessment of objective endpoints.

Falling lower in the hierarchy of evidence, implementation studies may use designs that are neither randomized nor controlled. For example, a research team may observe a single group for changes in health care delivery or patient outcomes before and after intervention implementation. In this case, the observed changes may result from multiple factors not associated with the intervention. Secular trends, such as increasing use of specific medications, may produce broad, population-based changes, irrespective of the intervention under study. Without a comparison group, secular trends may be confused with intervention effects.[35] Interrupted time-series designs use advanced statistical methodology with data collected from multiple points in time before and after the intervention to better account for secular trends.

In addition to confounding from secular trends, uncontrolled study designs are susceptible to other 'non-interventional' aspects of the intervention. For example, an intervention may bestow more attention on patients or clinicians through data collection, leading to self-reported improvement through placebo-like effects. Comparison groups, even without randomization, offer important protection against secular trends and placebo-like effects. Non-randomized allocation to intervention and comparison groups does not assure that both groups are similar in all important characteristics. Matched study designs may balance study groups for a limited number of measured characteristics. In contrast, successfully implemented randomization equalizes recognized and unrecognized confounders across study groups and is, therefore, essential for cause-and-effect inference.

In summary, limitations of study designs without randomization or a comparison group include difficulty establishing causality, confounding, bias, and spurious associations from multiple comparisons.[23] Although such studies are generally considered to be lower within the evidence hierarchy, they may provide useful information when randomized controlled trials (RCTs) are not feasible or generate important hypotheses for subsequent testing with more rigorous study designs. In keeping with the theme of this book, we focus the remainder of this chapter on RCTs for implementation research. In contrast to the traditional clinical RCT, implementation studies frequently randomize groups (clusters) rather than individuals. Therefore, we place particular emphasis on the cluster RCT.[68] Because implementation studies typically involve a complex set of design issues, we strongly recommend that investigators obtain expert consultation with methodologists and statisticians during the planning stages, rather than postponing this activity until after the intervention has been completed and the data are ready to analyze.

## *Implementation Randomized Controlled Trials*

Many principles for the design of high-quality, traditional RCTs discussed elsewhere in this book also apply to implementation research. In contrast, the following discussion emphasizes particular facets of the implementation RCT that may diverge from the more traditional clinical trial. As a discussion guide, our approach is approximately parallel to the Consolidated Standards of Reporting Trials (CONSORT), which were designed to encourage high-quality clinical randomized trials and promote a uniform reporting style. The CONSORT criteria emphasize the ability to understand the flow of all actual and potential research participants through the experimental design. Although originally designed for the traditional or 'parallel' clinical trial,[69,70] the CONSORT criteria were subsequently modified for the cluster RCT.[71,72] Finally, an exhibit at the end of the discussion provides a specific example of an implementation randomized trial.

## *Participants and Recruitment*

In contrast to the randomized clinical trial where patients are the unit of intervention and analysis, implementation randomized trials have a broader reach. For example, key participants in implementation RCTs may be doctors, patients, clinics, or hospitals, or hospital wards. Because implementation research is conducted in the 'real world' and often seeks to engage busy clinicians who are otherwise overwhelmed with their usual activities, recruitment may be particularly difficult. Therefore, recruitment protocols for implementation research demand careful consideration and may require a dedicated recruitment and retention team. Often multiple options (e.g., word of mouth, e-mail, phone, fax, personal contacts, or lists from professional organizations) must be pursued, and still the desired number of participants may not be reached.

## *Human Subjects*

The need for approval of implementation studies by an institutional review board (IRB) has sometimes been questioned under the assumption that the work is being performed for local quality improvement and not for research. However, randomization is not generally used for local quality improvement projects. In addition, the intention to publish study findings in the peer-reviewed literature or present at national scientific conferences clearly places the work in the research domain. Although IRB review is always required for implementation research, the research protocol may pose minimal danger to participants, and the review may be conducted under an expedited protocol. We refer the reader to more detailed reviews on this topic.[73–75]

Investigators designing cluster RCTs must carefully consider the ethical issues that arise when consent occurs at the cluster level with subsequent enrollment of participants within the cluster. If the target of the research is clearly the clinician, informed consent may often be waived for the patient. For studies that focus on the clinician but collect outcomes from medical record review or administrative patient records, the researchers may consider applying for a waiver of informed patient consent. Such waivers are especially reasonable when a large volume of patient records would make patient informed consent impractical. Implementation research usually generates personally identifiable health information, which may be subject to the Health Insurance Portability and Accountability Act (HIPAA). Waiver of HIPAA consent by the patient may often be obtained based on requirements similar to waiver of informed patient consent. Finally, it may be necessary to obtain consent from both patients and providers if the intervention targets both populations.

Investigators should develop detailed plans to protect the security and confidentiality of study data. Data should be housed in physically secured locations with strong logical protection, such as password protected and encrypted files. Access to study data should be only on a 'need-to-know' basis. Participant identifiers should be maintained only as necessary for data quality control and linkage. Patients and clinicians should be assured that personal information will not be revealed in publications or presentations. Data integrity should also be protected with detailed protocols for verification and cleaning, which are beyond the scope of this chapter.[76]

We agree with the International Committee of Medical Journal Editors (ICMJE) that descriptions of all randomized clinical trials should be deposited in publically available registries before recruitment begins.[77] The ICJME includes interventions focusing on process-of-care within the rubric of clinical trials. Trial registries guard against the well-recognized bias that negative studies are less likely to be published than positive studies. Negative publication bias may significantly limit meta-analytic studies, leading to the false conclusion that ineffective interventions are actually effective. Registries also increase the likelihood that participation in clinical trials will promote the public good, even if the study is negative. Although the template is not customized for implementation research, one such registry may be found at http://clinicaltrials.gov.

## *Intervention Design*

The previously described tools may serve as useful starting points for an innovative intervention design, which is often achieved using a formative-evaluation process.[78,79] Formative evaluation incorporates input from end users to refine an intervention during the early stages of development. Following this approach, Glasgow et al. recommend key features to include in the content design: (1) barrier analysis; (2) integration of multiple types of evidence; (3) adoption of practical trials that address clinician concerns; (4) investigation of multiple outcomes, generalizability, and contextual factors; (5) design of multilevel programs using systems and social

networking models mindful of the integration of the study's components and levels; and (6) adaptation of program to local needs and ongoing issues.[29] It is critical that investigators carefully explore and understand the need of those who will be affected by the intervention. Therefore, implementation studies may use such techniques as focus groups or nominal group technique in the planning phase.[80–83]

In the exhibit at the end of the chapter, we provide an example of an Internet-based strategy for delivering continuing medical education and promoting practice improvement for rural physicians. Casebeer et al. have identified the most important features in Internet-based instruction for physicians: (1) needs assessment from office practice data; (2) multimodal strategies; (3) modular design with multiple parts; (4) clinical cases for contextual learning; (5) tailoring intervention based on individual responses; (6) interactivity with the learner; (7) audit and feedback; (8) evidence-based content; (9) established credibility of organization providing website and funding entity; (10) patient education resources; (11) high level of usability; and finally, (12) accessibility to the Internet site despite limited bandwidth.[84]

## *Comparison Group*

It is often appropriate to randomize participants in behavioral research to either an active intervention versus an attention control. In contrast to the traditional placebo, the attention control accounts for changes in behavior attributable to social exposure when participants receive services and attention from study personnel.[85] Positive social interactions may create expectations for positive outcomes, potentially confounding intervention effects collected through such methods as self report. Although attention controls are widely recommended, their precise implementation may be difficult.[86]

In our experience, clinicians and communities may be reluctant to enter a study with the possibility of being randomized to a group with no apparent benefit. This problem may be compounded by intensive procedures needed for data collection, regardless of the study group. To overcome such barriers, investigators may offer to open the intervention to the comparison group at the close of the study. Alternatively, study design might more formally incorporate a delayed intervention or test two variations of an active intervention.

## *Blinding*

As with traditional clinical randomized trials, 'blinding' is important to decrease bias in outcome ascertainment. Study personnel who perform outcome assessment should be unaware of whether an individual participant has been assigned to the intervention or comparison group. For example, it may be necessary to blind those doing patient examinations, those performing medical record abstraction, or

those administering patient, physician, or organizational surveys. When participants are blinded to the allocation arm, the study is single-blinded. If those delivering the intervention and collecting the outcomes are blinded as well, then the study is double-blinded. If the analysts are unaware of the assignments, then the study is triple-blinded. For implementation research, it is often not feasible to conceal study allocation from the research team, as illustrated by the RDOC exhibit.

## *Units of Intervention, Randomization, and Analysis*

Investigators planning an implementation randomized trial must carefully consider the units of study assignment for intervention, randomization, and analysis. Within any given study, the unit level may vary across components, meaning that the analysis plan must account for the clustered nature of the outcome data. For example, consider a study of a patient-based intervention that will be implemented through a group of affiliated multi-physician clinics. 'Contamination' could arise from physicians learning about the intervention and then exposing comparison patients to part of the intervention. Therefore, for this particular study, the investigators may chose to randomize at the physician level to avoid contamination. Thus, all patients assigned to a given physician will be allocated to the same condition: intervention or comparison.

In practice, the threat of contamination may be more perceived than real, depending upon the exact nature of the intervention and study setting. When present, contamination decreases the precision with which the intervention effect will be measured and increases the risk of a Type II error. As an alternative to cluster-based randomization to overcome contamination, the sample size could be increased.[87]

## *Approaches to Randomization*

The construct of randomization is described elsewhere in this book. In summary, randomization is a procedure to assure that study units (e.g., patients, physicians, clinics, hospitals, hospital wards) are allocated to the study conditions (e.g., intervention, comparison) according to chance alone. The specific approach to randomization is described as 'sequence generation' and may include matching or stratification as described below in more detail.[71] Allocation concealment is a 'technique used to prevent selection bias by concealing the allocation sequence from those assigning participants to intervention groups, until the moment of assignment.'[69,70] In other words, the purpose of this arrangement is to prevent researchers from influencing which participants are assigned to a given group. The concealment may be simply based on a coded list of randomly ordered study groups created by

a statistician who is not a member of the intervention team. After enrollment, each participant is assigned to a study group based on the sequence in the list.

For cluster randomized trials, the assignment of individuals to a study group is determined at the level of the cluster, which increases the opportunity for selection bias from failed concealment. For example, consider the cluster RCT described above where randomization occurs at the physician level with subsequent enrollment of patients from the physicians' practice. Depending upon the nature of the intervention, physicians may be able to determine their randomization group. If the randomized physician also recruits patients for the study, this knowledge of the randomization group may lead to biased patient selection.

Successful randomization ensures balanced characteristics at the unit of randomization, and larger numbers of randomized units increase the chance of successful randomization. Investigators should be aware that for cluster RCTs, successful randomization does not ensure balanced characteristics at units below the level of randomization.[88] Again, consider the illustration above where randomization occurs at the physician level. Although this design may produce intervention and comparison groups that are balanced based on physician characteristics, there may be important imbalances in patient characteristics, decreasing the power of randomization. To guard against imbalances of lower-level units in cluster randomized trials, investigators might consider stratifying or matching on a limited number of critical characteristics.[89] Alternatively, imbalances may require statistical adjustment at the point of analysis after the study has been completed. Decisions about matched study designs for cluster randomized trials are complex and beyond the scope of this chapter.

## *Intent-to-Treat and Loss to Follow Up*

As with the traditional clinical randomized trial, the primary analysis for an implementation randomized trial should test hypotheses specified *a priori* and should follow intent-to-treat principles.[90] With the intent-to-treat approach, all units are analyzed with the group to which they were originally randomized, regardless of whether the units are subsequently exposed to the intervention (i.e., cross over). For example, in a randomized trial of an Internet-based continuing medical education (CME) intervention for physicians, outcomes for all physicians randomized to the intervention group must be analyzed as part of the intervention group, regardless of whether the physician visited the Internet site. Intent-to-treat protocols preserve the power of randomization by protecting against bias resulting from differential participation or cross-over among intervention units with a greater or lesser propensity for success.

Unfortunately, participants lost to follow up may generate no data for analysis. As with violation of the intent-to-treat principle, loss to follow up may reduce the power of randomization. Although complete follow up is desirable, it is usually not

obtainable. Many scientists hold that for clinical trials, loss to follow up of greater than 20% introduces severe potential for bias.[91]

Therefore, many study designs include run-in phases before randomization. From the perspective of internal validity, it is better to exclude participants before randomization than have participants lost to follow up, cross between study groups, or become non-adherent to intervention protocols after randomization. For example, in the study of Internet-based CME described above, physicians might be required to demonstrate a willingness to engage in Internet learning and submit data for study evaluation before randomization. According to the CONSORT criteria for group randomized trials, investigators must carefully account for all individuals and clusters that were screened or randomized.[71]

## *Statistical Analysis*

Statistical analysis for cluster RCTs is a vast, technical topic which falls largely beyond the domain of the basic introduction provided in this book. However, an example will illustrate some important principles. More specifically, consider the previous illustration in which physicians are randomized to an intervention or comparison group, with patients being subsequently enrolled and assigned to the same study condition as their physician. To conduct the analysis at the physician level, the investigators might simply compare the mean post-intervention outcomes for the two study groups. However, this approach leads to loss of statistical power, because the number of physicians randomized will be less than the number of patients included in the study. Alternatively, the investigators could plan a patient-level analysis that appropriately considers the clustering of patients within physicians. The investigators could also collect outcomes for intervention and comparison patients before and after intervention implementation. Generalized estimation equations could then be used to compare the change in study endpoints over time for the intervention versus comparison group. Here, the main study effect will be reflected by a group-time interaction variable included in the multivariable model. This approach uses a marginal, population-averaged model to account for clustered observations and potentially adjust for observed imbalances in the study groups. Alternatively, the analyst may use a cluster-specific (or conditional) approach that directly incorporates random effects. Murray reviewed the evolving science and controversies surrounding the analysis of group-randomized trials.[89]

Although the main analysis should follow intent-to-treat principles as described above, most implementation randomized trials include a range of secondary analyses. Such secondary analyses may yield important findings, but they do not carry the power of cause-and-effect inference. ‘Per-protocol’ or ‘compliers only’ analyses may address the impact of the intervention among those who are sufficiently exposed or may examine dose-response relationships between intervention exposure and outcomes. Mediation analysis using a series of staged

regression models may investigate mechanisms through which an intervention leads to a positive study effect.[92,93]

## *Sample Size Calculations*

When designing an implementation trial, the investigator must determine the number of participants necessary to detect a meaningful difference in study endpoints between the intervention and comparison groups, i.e., the power of the study. Typically, a power of 80% is considered adequate to decrease the likelihood of a false negative result. If an intervention is sustained over an extended period of time, the investigators may wish to test specifically for effect decay, perhaps with a time-trend analysis. Such a hypothesis of no difference demands a special approach to power calculation. Sample size calculations for traditional randomized trials are discussed elsewhere in this book.

As described above, the analysis for an implementation randomized trial may be at a lower level than the unit of randomization. Under these circumstances, the power calculations must account for the clustering of participants within upper-level units, such as the clustering of patients within physicians from the example above. Failure to account for the hierarchical data structure may inflate the observed statistical significance and increase the likelihood of a false positive finding.[94]

Several approaches to accounting for the clustering of, say, patients within physicians from the above example, rely on the intra-class correlation coefficient (ICC). The ICC is the ratio of the between-cluster variance to the total sample variance (between clusters + within cluster). In this example, the ICC would be a measure of how 'alike' patient outcomes were within the physician clusters. If the ICC is 1, the outcomes for all patients clustered within a given physician are identical. If the ICC is 0, clustering within physicians is not related to patient outcomes.[95] In other words, with an ICC of 1, adding additional patients provides no additional information. Therefore, as the ICC increases, one must increase the sample size to retain the same power. For $0 < ICC < 1$, increasing the number of patients will increase study power less than increasing the number of physicians. Typical values for ICCs range from 0.01–0.50.[96]

Although the topic of power calculations for group randomized trials is vast and largely beyond the scope of this book, Donner provides a straight-forward framework for simple situations.[94] Taking this approach, the analyst first calculates an unadjusted sample size ($N_{un}$) using approaches identical to those described elsewhere in this book for the traditional randomized clinical trial. Next, the analyst calculates a sample inflation factor (IF) which is used to derive a cluster-adjusted sample size ($N_{adj}$). Then:

$$IF = [1+(m-1)\rho] \text{ and}$$
$$N_{adj} = (N_{un})*IF,$$

where m is the number of study units per cluster, and ρ is the ICC.

### *Situational Analysis and External Validity*

Because implementation randomized trials occur in a 'real-word' setting, we place special emphasis on understanding and reporting of context. In contrast to the traditional randomized clinical trial, the study setting for the implementation trial is an integral part of the study design. To address the importance of context in implementation research, Davidoff and Batalden promote the concept of situational analysis for quality improvement studies.[55] We believe that many of these principles are relevant to the implementation randomized trial. For example, published reports for implementation research should include specific details about the clinic setting, patient population, prior experience with system change, and how the context contributed to understanding the problem for which the study was designed.

Because implementation research often focuses on dissemination to large populations, external validity, or generalizability, acquires special importance. One must consider how study findings are applicable to other patients, doctors, clinics, or geographic locations. Fortunately, established criteria for external validity are available and are applicable to the implementation trial.[29] In summary, these criteria hinge upon: (1) the study's reach and sample representativeness, which includes the participants and setting; (2) the consistency of intervention implementation and the ability to adapt the intervention to other settings; (3) the magnitude of intervention effect, adverse outcomes, program intensity, and cost; and (4) the intervention's long-term effects, sustainability, and attrition rates. Finally, specialized approaches to economic evaluation provide additional important context for interpreting the results from implementation trials.[97]

## Summary

Implementation research bridges the gap between scientific knowledge and its application to daily practice with the overall purpose of improving the health of individuals and populations. To advance the science of implementation research, the Institute of Medicine published findings from the Forum on the Science of Health Care Quality Improvement and Implementation in 2007[98] and the Veterans' Health Administration sponsored a state-of-the-art (SOTA) conference in 2004.[3] Together, these documents summarized current knowledge, identified barriers to implementation research, and defined strategies to overcome these barriers. Given the well-documented quality and safety problems of our health care system despite the vast resources invested in the biomedical sciences, we need to promote interest in implementation research, an emerging scientific discipline focused on improving health care for all, regardless of geography, socioeconomic status, race, or ethnicity.

## Exhibit: Rural Diabetes Online Care (RDOC)

### *Background*

As the prevalence of type II diabetes in the United States continues to rise, rural physicians face important barriers to helping their patients achieve adequate disease control. In particular, the rural South has many disadvantaged and minority patients with limited health care access. Therefore, the goal of the Rural Diabetes Online Care (RDOC) project is to evaluate the effectiveness of a multifaceted, professional-development Internet intervention for rural primary care physicians. We hypothesize that patients of intervention physicians will achieve lower risk of cardiovascular and diabetes-related complications through improved control of diabetes, blood pressure, and lipids.

### *Objectives*

The objectives of RDOC are to: (1) assess barriers to implementation of diabetes guidelines and identify solutions through physician focus groups and case-based vignette surveys; (2) develop and implement an interactive Internet intervention including individualized physician performance feedback; (3) evaluate the intervention in a randomized controlled trial; and (4) examine the sustainability of improved guideline adherence after feedback.

### *Methods*

RDOC is a group-randomized implementation trial for health care providers in rural primary care offices. At the time of press, the intervention has been completed and recruitment and retention activities are ongoing. The study is open to physicians, nurses, and office personnel. Offices of primary care physicians located in rural areas were identified, and a recruitment plan was developed that included material distributed by mail, facsimile, presentations at professional meetings, physician-to-physician telephone conversations, and on-site office visits.

To enroll, a primary care physician must access the study Internet site and review the online consent material. Randomization to an intervention or comparison group occurs on-line immediately after consent. The first physician from an office to enroll is designated as the ‘lead physician.’ Subsequent physicians or office personnel participating in the study are assigned to the same study arm as the lead physician.

The intervention website, which was developed with input from rural primary physicians, contains: (1) practice timesavers; (2) practical goals and guidelines; (3) challenging cases; and, (4) patient education materials. Lead physicians receive feedback about areas for practice improvement based on medical record review. Those in the intervention group also receive feedback from interactive and challenging case vignettes. Based on data from medical record review and the case vignettes, intervention physicians will be able to compare their performance with that of their peers. The control website contains traditional text-based continuing medical education (CME) and links to nationwide diabetes resources. Participants are eligible to receive CME credits for completing sections from the website.

Outcomes will be ascertained before and after intervention implementation through medical record abstraction. For intervention physicians, medical record abstraction will also used to generate performance feedback that is delivered through the RDOC Internet site. Providers in the physician offices, chart abstractors, and statisticians are blinded to the study group assignments (intervention versus comparison), but the implementation team must be aware of study assignment for recruitment and retention activities.

The main analysis, conducted at the patient level based on intent-to-treat principles, will compare differential improvement in guideline adherence and intermediate physiologic outcomes between the study groups. More specifically, study outcomes will be linked to the lead physician and will focus on appropriate therapy and levels of control for blood sugar, hypertension, and lipids. Ancillary analyses will examine the effects of physician characteristics, other providers in the office, and patient characteristics (e.g., comorbidity ethnicity, gender, age, and socioeconomic status). Multivariable techniques will account for the clustering of patients within physicians and multiple providers within a single office. Based on the sample size calculations, we plan to equally randomize 200 physician offices to the intervention or comparison group. We will abstract 10–15 medical records for each lead physician.

## *Significance*

This study offers a technologically advanced, theory-grounded intervention to improve the care of a high-risk, underserved population. The implementation team has interdisciplinary expertise in translating research into practice, rural medicine, behavioral medicine, health informatics, and clinical diabetes. Our goal is to produce an evidence-based intervention that is sustainable in the ‘real world,’ and easily modified for other diseases.

ClinicalTrials.gov Identifier: NCT00403091 (Available at: http://clinicaltrials.gov).

## Resources

### *Selected Journals That Publish Implementation Research*

- Annals of Internal Medicine
- Implementation Science
- JAMA
- Medical Care
- Quality and Safety in Health Care

### *Selected Checklists and Reporting Guidelines*

- Enhancing the Quality and Transparency of health Research (EQUATOR)
  - EQUATOR is an initiative of the National Knowledge Service and the National Institute for Health Research that seeks to improve the quality of scientific reporting.
  - This initiative includes statements about reporting for a range of experimental and observational study types, including randomized trials, group randomized trials, behavioral trials, and quality interventions.
  - http://www.equator-network.org
- Consolidated Standards of Reporting Trials (CONSORT)
  - This initiative focuses on design and reporting standards for randomized controlled trials (RCTs) in health care.
  - Although originally designed for the traditional 'parallel' randomized clinical trial, the CONSORT criteria have been extended to include cluster RCTs and behavioral RCTs.
  - http://www.consort-statement.org/.
- Quality improvement evaluations
  - Davidoff F, Batalden P. Toward stronger evidence on quality improvement. Draft publication guidelines: the beginning of a consensus project. Qual Saf Health Care 2005; 14:319–325.

### *Selected Resources for Intervention Design*

- Evidence-based Practice Centers (EPC)
  - These centers are funded by the Agency for Healthcare Research and Quality to conduct systematic literature reviews and generate evidence reports.

- Several publically available reports focus on information technology and interventions to improve health care quality and safety.
- http://www.ahrq.gov/clinic/epc.

- Veterans' Administration Quality Enhancement Research Initiative (QUERI) Implementation Guides
  - The QUERI Implementation Guide is a four-part series focusing on practical issues for designing and conducing implementation research.
  - The guide includes material on conceptual models, diagnosing performance gaps, developing interventions, quasi-experimental study design.
  - http://hsrd.research.va.gov/queri/implementation.
- Finding Answers
  - This program is sponsored by the Robert Wood Johnson Foundation to develop interventions for eliminating racial/ethnic disparities in health care.
  - The Finding Answers Intervention Research (FAIR) database includes 206 manuscripts from a systematic review of interventions to decrease racial/ethnic disparities for breast cancer, cardiovascular disease, and diabetes. Interventions based on cultural leverage and performance-based reimbursement are also included.
  - http://www.solvingdisparities.org/toolsresources.
- National Center for Cultural Competence
  - This center is sponsored by Georgetown University and offers several implementation tools, manuscripts, and policy statements for organizations, clinicians, and consumers.
  - The Internet site has a section for 'promising practices' which may be particularly useful in designing new interventions.
  - http://www11.georgetown.edu/research/gucchd/nccc/.
- Clinical microsystems
  - The Dartmouth Institute for Health Policy and Clinical Practice maintains this Internet resource that offers tools for improving clinical microsystems.
  - Most tools are generally available to the public at no cost.
  - The Clinical Microsystems Action Guide may be particularly useful for designing new interventions.
  - http://clinicalmicrosystem.org.
- Institute for Healthcare Improvement
  - This not-for-profit organization maintains an Internet site that contains several tools for improving the quality, safety, and efficiency of health care. Many tools are publically available at no cost.
  - White papers describing the 'Breakthrough Series' may be particularly useful for those developing new interventions.
  - http://www.ihi.org.

**Acknowledgements** The authors thank Sei Lee, MD and Brook Watts, MD for their review and comments on a prior version of this chapter. The RDOC project is supported by NIDDK R18DK65001 grant to Dr. Allison.

## References

1. Berwick DM. Disseminating innovations in health care. JAMA 2003; 289:1969–75.
2. Lenfant C. Shattuck lecture–clinical research to clinical practice–lost in translation? N Engl J Med 2003; 349:868–74.
3. Kiefe CI, Sales A. A state-of-the-art conference on implementing evidence in health care. Reasons and recommendations. J Gen Intern Med 2006; 21 Suppl 2:S67–70.
4. McGlynn EA, Asch SM, Adams J, et al. The quality of health care delivered to adults in the United States. N Engl J Med 2003; 348:2635–45.
5. Rubenstein LV, Pugh J. Strategies for promoting organizational and practice change by advancing implementation research. J Gen Intern Med 2006; 21 Suppl 2:S58–64.
6. Kiefe CI, Safford M, Allison JJ. Forces influencing the care of complex patients: a framework. In: Academy Health Annual Meeting, 2007. Orlando, Fl; 2007.
7. Unequal treatment: confronting racial and ethnic disparities in health care. Washington, DC: National Academies Press; 2003.
8. Allison JJ. Health disparity: causes, consequences, and change. Med Care Res Rev 2007; 64:5S–6S.
9. Chin MH, Walters AE, Cook SC, Huang ES. Interventions to reduce racial and ethnic disparities in health care. Med Care Res Rev 2007; 64:7S–28S.
10. Kilbourne AM, Switzer G, Hyman K, Crowley-Matoka M, Fine MJ. Advancing health disparities research within the health care system: a conceptual framework. Am J Public Health 2006; 96:2113–21.
11. Smedley BD, Stith AY, Nelson AR. Unequal treatment: confronting racial and ethnic disparities in health care. Washington, DC: Institute of Medicine; 2003.
12. Flaskerud JH, Nyamathi AM. Attaining gender and ethnic diversity in health intervention research: cultural responsiveness versus resource provision. ANS Adv Nurs Sci 2000; 22:1–15.
13. Hayward RA, Asch SM, Hogan MM, Hofer TP, Kerr EA. Sins of omission: getting too little medical care may be the greatest threat to patient safety. J Gen Intern Med 2005; 20:686–91.
14. Kohn LT, Corrigan JM, Donaldson MS. To err is human: building a safer health system. Washington, DC: Institute of Medicine; 1999.
15. Public Health Action Plan to Combat Antimicrobial Resistance Centers for Disease Control and Prevention, 1999. (Accessed November 2007, at http://www.cdc.gov/drugresistance/actionplan/html/index.htm.)
16. Snow V, Mottur-Pilson C, Gonzales R. Principles of appropriate antibiotic use for treatment of acute bronchitis in adults. Ann Intern Med 2001; 134:518–20.
17. Wenzel RP, Fowler AA, 3rd. Clinical practice. Acute bronchitis. N Engl J Med 2006; 355:2125–30.
18. Roumie CL, Halasa NB, Grijalva CG, et al. Trends in antibiotic prescribing for adults in the United States–1995 to 2002. J Gen Intern Med 2005; 20:697–702.
19. Metlay JP, Camargo CA, Jr., MacKenzie T, et al. Cluster-randomized trial to improve antibiotic use for adults with acute respiratory infections treated in emergency departments. Ann Emerg Med 2007; 50:221–30.
20. Crossing the quality chasm: a new health system for the 21st century. Washington, DC: Institute of Medicine; 2001.
21. Berwick DM, Calkins DR, McCannon CJ, Hackbarth AD. The 100,000 lives campaign: setting a goal and a deadline for improving health care quality. JAMA 2006; 295:324–7.

22. Auerbach AD, Landefeld CS, Shojania KG. The tension between needing to improve care and knowing how to do it. N Engl J Med 2007; 357:608–13.
23. Smith GC, Pell JP. Parachute use to prevent death and major trauma related to gravitational challenge: systematic review of randomised controlled trials. BMJ 2003; 327:1459–61.
24. Brennan TA, Gawande A, Thomas E, Studdert D. Accidental deaths, saved lives, and improved quality. N Engl J Med 2005; 353:1405–9.
25. Berwick D. The stories beneath. Medical Care 2007; 45:1123–5.
26. Bhattacharyya O, Reeves S, Garfinkel S, Zwarenstein M. Designing theoretically-informed implementation interventions: fine in theory, but evidence of effectiveness in practice is needed. Implement Sci 2006;1 Feb 23:5.
27. Effective health care: getting evidence into practice. National Health Service Center for Reviews and Dissemination, Royal Society of Medicine Press, 1999; 5(1). (Accessed November 2007, at http://www.york.ac.uk/inst/crd/ehc51.pdf.)
28. Rogers EM. Diffusion of innovations (5th ed.). New York: Free Press; 2003.
29. Glasgow RE, Emmons KM. How can we increase translation of research into practice? Types of evidence needed. Annu Rev Public Health 2007; 28:413–33.
30. Sung NS, Crowley WF, Jr., Genel M, et al. Central challenges facing the national clinical research enterprise. JAMA 2003; 289:1278–87.
31. Nelson EC, Batalden PB, Huber TP, et al. Microsystems in health care: part 1. Learning from high-performing front-line clinical units. Joint Comm J Qual Im 2002; 28:472–93.
32. Zerhouni EA. Medicine. The NIH roadmap. Science 2003; 302:63–72.
33. Zerhouni EA. US biomedical research: basic, translational, and clinical sciences. JAMA 2005; 294:1352–8.
34. Chao SR. The state of quality improvement and implementation research: expert views–workshop summary. Washington, DC: National Academies Press; 2007.
35. Shojania KG, Grimshaw JM. Evidence-based quality improvement: the state of the science. Health Aff (Millwood) 2005; 24:138–50.
36. Shojania KG, McDonald KM, Wachter RM, Owens DK. Closing The Quality Gap: A Critical Analysis of Quality Improvement Strategies, Volume 1—Series Overview and Methodology. Technical Review 9. (Contract No. 290-02-0017 to the Stanford University—UCSF Evidence-based Practices Center). AHRQ Publication No. 04-0051-1. Rockville, MD: Agency for Healthcare Research and Quality. August 2004.
37. Williams GC, Deci EL. Activating patients for smoking cessation through physician autonomy support. Med Care 2001; 39:813–23.
38. Williams GC, McGregor H, Zeldman A, Freedman ZR, Deci EL, Elder D. Promoting glycemic control through diabetes self-management: evaluating a patient activation intervention. Patient Educ Couns 2005; 56:28–34.
39. Turner RC, Waivers LE, O'Brien K. The effect of patient-carried reminder cards on the performance of health maintenance measures. Arch Intern Med 1990; 150:645–7.
40. Bodenheimer T, Lorig K, Holman H, Grumbach K. Patient self-management of chronic disease in primary care. JAMA 2002; 288:2469–75.
41. Eng E, Parker E, Harlan C. Lay health advisor intervention strategies: a continuum from natural helping to paraprofessional helping. Health Educ Behav 1997; 24:413–7.
42. Swider SM. Outcome effectiveness of community health workers: an integrative literature review. Public Health Nurs 2002; 19:11–20.
43. Institute of Medicine. Clinical practice guidelines: directions for a new program. Washington, DC: National Academy Press; 1990.
44. Grimshaw J, Eccles M, Thomas R, et al. Toward evidence-based quality improvement. Evidence (and its limitations) of the effectiveness of guideline dissemination and implementation strategies 1966–1998. J Gen Intern Med 2006; 21 Suppl 2:S14–20.
45. Cabana MD, Rand CS, Powe NR, et al. Why don't physicians follow clinical practice guidelines? A framework for improvement. JAMA 1999; 282:1458–65.
46. Boyd CM, Darer J, Boult C, Fried LP, Boult L, Wu AW. Clinical practice guidelines and quality of care for older patients with multiple comorbid diseases: implications for pay for performance. JAMA 2005; 294:716–24.

47. Davis D, O'Brien MA, Freemantle N, Wolf FM, Mazmanian P, Taylor-Vaisey A. Impact of formal continuing medical education: do conferences, workshops, rounds, and other traditional continuing education activities change physician behavior or health care outcomes? JAMA 1999; 282:867–74.
48. Davis DA, Thomson MA, Oxman AD, Haynes RB. Changing physician performance. A systematic review of the effect of continuing medical education strategies. JAMA 1995; 274:700–5.
49. Mazmanian PE, Davis DA. Continuing medical education and the physician as a learner: guide to the evidence. JAMA 2002; 288:1057–60.
50. Centor R, Casebeer L, Klapow J. Using a combined CME course to improve physicians' skills in eliciting patient adherence. Acad Med 1998; 73:609–10.
51. Fordis M, King JE, Ballantyne CM, et al. Comparison of the instructional efficacy of Internet-based CME with live interactive CME workshops: a randomized controlled trial. JAMA 2005; 294:1043–51.
52. Soumerai SB, Avorn J. Principles of educational outreach ('academic detailing') to improve clinical decision making. JAMA 1990; 263:549–56.
53. Valente TW, Pumpuang P. Identifying opinion leaders to promote behavior change. Health Educ Behav 2007; 34(6):881–96.
54. Kiefe CI, Allison JJ, Williams OD, Person SD, Weaver MT, Weissman NW. Improving quality improvement using achievable benchmarks for physician feedback: a randomized controlled trial. JAMA 2001; 285:2871–9.
55. Davidoff F, Batalden P. Toward stronger evidence on quality improvement. Draft publication guidelines: the beginning of a consensus project. Qual Saf Health Care 2005; 14:319–25.
56. Jha AK, Perlin JB, Kizer KW, Dudley RA. Effect of the transformation of the Veterans Affairs Health Care System on the quality of care. N Engl J Med 2003; 348:2218–27.
57. Payne TH. Computer decision support systems. Chest 2000; 118:47S–52S.
58. Walton RT, Harvey E, Dovey S, Freemantle N. Computerised advice on drug dosage to improve prescribing practice. Cochrane Database Syst Rev 2001:CD002894.
59. Han YY, Carcillo JA, Venkataraman ST, et al. Unexpected increased mortality after implementation of a commercially sold computerized physician order entry system. Pediatrics 2005; 116:1506–12.
60. Nebeker JR, Hoffman JM, Weir CR, Bennett CL, Hurdle JF. High rates of adverse drug events in a highly computerized hospital. Arch Intern Med 2005; 165:1111–6.
61. Scalise D. Technology. CPOE: are you really ready? Hosp Health Netw 2006; 80:14, 6.
62. Ash JS, Sittig DF, Poon EG, Guappone K, Campbell E, Dykstra RH. The extent and importance of unintended consequences related to computerized provider order entry. J Am Med Inform Assoc 2007; 14:415–23.
63. Werner RM, Asch DA. The unintended consequences of publicly reporting quality information. JAMA 2005; 293:1239–44.
64. Werner RM, Asch DA, Polsky D. Racial profiling: the unintended consequences of coronary artery bypass graft report cards. Circulation 2005; 111:1257–63.
65. Rosenthal MB, Frank RG, Li Z, Epstein AM. Early experience with pay-for-performance: from concept to practice. JAMA 2005; 294:1788–93.
66. Glickman SW, Ou FS, DeLong ER, et al. Pay for performance, quality of care, and outcomes in acute myocardial infarction. JAMA 2007; 297:2373–80.
67. Lindenauer PK, Remus D, Roman S, et al. Public reporting and pay for performance in hospital quality improvement. N Engl J Med 2007; 356:486–96.
68. Murray DM. Design and analysis of group-randomized trials. New York: Oxford University Press; 1998.
69. Begg C, Cho M, Eastwood S, et al. Improving the quality of reporting of randomized controlled trials. The CONSORT statement. JAMA 1996; 276:637–9.
70. Moher D, Schulz KF, Altman D. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. JAMA 2001; 285:1987–91.
71. Campbell MK, Elbourne DR, Altman DG. CONSORT statement: extension to cluster randomised trials. BMJ 2004; 328:702–8.
72. Elbourne DR, Campbell MK. Extending the CONSORT statement to cluster randomized trials: for discussion. Stat Med 2001; 20:489–96.

73. Casarett D, Karlawish JH, Sugarman J. Determining when quality improvement initiatives should be considered research: proposed criteria and potential implications. JAMA 2000; 283:2275–80.
74. Emanuel EJ, Wendler D, Grady C. What makes clinical research ethical? JAMA 2000; 283:2701–11.
75. Lynn J, Baily MA, Bottrell M, et al. The ethics of using quality improvement methods in health care. Ann Intern Med 2007; 146:666–73.
76. Van den Broeck J, Cunningham SA, Eeckels R, Herbst K. Data cleaning: detecting, diagnosing, and editing data abnormalities. PLoS Med 2005; 2:e267.
77. Uniform Requirements for Manuscripts Submitted to Biomedical Journals: Writing and Editing for Biomedical Publication: International Committee of Medical Journal Editors. Available at www.ICMJE.org (last accessed November 2007).
78. Scriven M. Beyond formative and summative evaluation. In: McLaughlin MW, Phillips DC, eds. Evaluation and education: 90th yearbook of the National Society for the Study of Education. Chicago, IL: University of Chicago Press; 1991:18–64.
79. Weston CB, McAlpine L, Bordonaro T. A model for understanding formative evaluation in instructional design. Education Tech Research Dev 1995; 43:29–49.
80. Delbecq AL, Van de Ven AH, Gustafson DH. Group techniques for program planning: a guide to nominal group and Delphi processes. Glenview, IL: Scott Foresman; 1975.
81. Krueger RA, Casey MA. Focus groups: a practical guide for applied research (3rd ed.). Thousand Oaks, CA: Sage; 2000.
82. Nielsen J, Mack R. Usability inspection methods. New York: Wiley; 1994.
83. Strauss A, Corbin J. Basics of qualitative research: grounded theory, procedures, and techniques. Newbury Park, CA: Sage; 1990.
84. Casebeer LL, Strasser SM, Spettell CM, et al. Designing tailored Web-based instruction to improve practicing physicians' preventive practices. J Med Internet Res 2003; 5:e20.
85. Bootzin RR. The role of expectancy in behavior change. In: White L, Turskey B, Schwartz G, eds. Placebo: theory, research, and mechanisms. New York: Guilford Press; 1985:196–210.
86. Gross D. On the merits of attention-control groups. Res Nurs Health 2005; 28:93–4.
87. Torgerson DJ. Contamination in trials: is cluster randomisation the answer? BMJ 2001; 322:355–7.
88. Puffer S, Torgerson D, Watson J. Evidence for risk of bias in cluster randomised trials: review of recent trials published in three general medical journals. BMJ 2003; 327:785–9.
89. Murray DM, Varnell SP, Blitstein JL. Design and analysis of group-randomized trials: a review of recent methodological developments. Am J Public Health 2004; 94:423–32.
90. Lachin JM. Statistical considerations in the intent-to-treat principle. Control Clin Trials 2000; 21:167–89.
91. Schulz KF, Grimes DA. Sample size slippages in randomised trials: exclusions and the lost and wayward. Lancet 2002; 359:781–5.
92. Baron RM, Kenny DA. The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. J Pers Soc Psychol 1986; 51:1173–82.
93. Preacher KJ, Hayes AF. SPSS and SAS procedures for estimating indirect effects in simple mediation models. Behav Res Methods Instrum Comput 2004; 36:717–31.
94. Donner A, Klar N. Pitfalls of and controversies in cluster randomization trials. Am J Public Health 2004; 94:416–22.
95. Beach ML. Primer on group randomized trials. Eff Clin Pract 2001; 4:42–3.
96. Campbell MK, Fayers PM, Grimshaw JM. Determinants of the intracluster correlation coefficient in cluster randomized trials: the case of implementation research. Clin Trials 2005; 2:99–107.
97. Sculpher M. Evaluating the cost-effectiveness of interventions designed to increase the utilization of evidence-based guidelines. Fam Pract 2000;17 Suppl 1:S26–31.
98. Institute of Medicine. Advancing quality improvement research: challenges and opportunities–workshop summary. Washington, DC: National Academies Press; 2007. (Accessed November 2007, at www.nap.edu/catalog/11884.html.)

# Chapter 14
# Research Methodology for Studies of Diagnostic Tests

**Stephen P. Glasser**

**Abstract** Much of clinical research is aimed at assessing causality. However, clinical research can also address the value of new medical tests, which will ultimately be used for screening for risk factors, to diagnose a disease, or to assess prognosis. In order to be able to construct research questions and designs involving these concepts, one must have a working knowledge of this field. In other words, although traditional clinical research designs can be used to assess some of these questions, most of the studies assessing the value of diagnostic testing are more akin to descriptive observational designs, but with the twist that these designs are not aimed to assess causality, but are rather aimed at determining whether a diagnostic test will be useful in clinical practice. This chapter will introduce the various ways of assessing the accuracy of diagnostic tests, which will include discussions of sensitivity, specificity, predictive value, likelihood ratio, and receiver operator characteristic curves.

## Introduction

Up to this point in the book, we have been discussing clinical research predominantly from the standpoint of causality. Clinical research can also address the value of new medical tests, which will ultimately be used for screening for risk factors, to diagnose a disease, or to assess prognosis. The types of research questions one might formulate for this type of research include: "How does one know how good a test is in giving you the answers that you seek?" or "What are the rules of evidence against which new tests should be judged?" In order to be able to construct research questions and designs involving these concepts, one must have a working knowledge of this field. In other words, although traditional clinical research designs can be used to assess some of these questions, most of the studies assessing the value of diagnostic testing are more akin to descriptive observational designs, but with the twist that these designs are not aimed to assess causality, but are rather aimed at determining whether a diagnostic test will be useful in clinical practice.

S.P. Glasser (ed.), *Essentials of Clinical Research*,

## Bayes Theorem

Thomas Bayes was an English theologian and mathematician who lived from 1702–1761. In an essay published posthumously in 1863 (by Richard Price), Bayes' offers a solution to the problem "…to find the chance of probability of its happening (a disease in the current context) should be somewhere between any two named degrees of probability."[1] Bayes' Theorem provides a way to apply quantitative reasoning to the scientific method. That is, if a hypothesis predicts that something should occur and it does, it strengthens our belief in that hypothesis; and, conversely if it does not occur, it weakens our belief. Since most predictions involve probabilities i.e. a hypothesis predicts that an outcome has a certain percentage chance of occurring, this approach has also been referred to as probabilistic reasoning. Bayes' Theorem is a way of calculating the degree of belief one has about a hypothesis. Said in another way, the degree of belief in an uncertain event is conditional on a body of knowledge. Suppose we're screening people for a disease (D) with a test which gives either a positive or a negative result (A and B, or T+ and T− respectively). Suppose further that the test is quite accurate, in the sense that, for example, it will give a positive result 95% of the time when the disease is present (D+), i.e. p(T+ |D+) = 0.95 (this formula asks what is the probability of the disease being present GIVEN a positive test?), or said another way, what is the probability that a person who tests positive has disease? The naive answer is 95%; but this is wrong. What we really want to know is p(D+ |T+), that is, what is the probability of testing positive if one has the disease; and, Bayes's theorem (or predictive value) tells us that.

In modern medicine the first useful application of Bayes' theorem was reported in 1959.[2] Ledley and Lusted demonstrated a method to determine the likelihood that a patient had a given disease when various combinations of symptoms known to be associated with that disease were present.[2] Redwood et al. utilized Bayesian logic to reconcile seemingly discordant results of treadmill exercise testing and coronary angiography.[3] In 1977, Rifkin and Hood pioneered the routine application of Bayesian probability in the non-invasive detection of coronary artery disease (CAD).[4] This was followed by other investigative uses of Bayesian analysis, an approach which has now become one of the common ways of evaluating all diagnostic testing.

As noted above, diagnostic data can be sought for a number of reasons beside just the presence or absence of disease. For example, the interest may be the severity of the disease, the ability to predict the clinical course of a disease, or to predict a therapy response. For a test to be clinically meaningful one has to determine how the test results will affect clinical decisions, what are its cost, risks, and what is the acceptability of the test; in other words, how much more likely will one be about this patients problem after a test has been performed than one was before the test; and, is it worth the risk and the cost? Recall, that the goal of studies of diagnostic testing seeks to determine whether a test is useful in clinical practice. To derive the latter we need to determine whether the test is reproducible, how accurate it is, whether the test affects clinical decisions, etc. One way to statistically assess test reproducibility (i.e. inter and intra-variability of test interpretation), is with a kappa statistic.[5] Note that reproducibility does not require a gold standard, while accuracy

does. In order to talk intelligently about diagnostic testing, some basic definitions and understanding of some concepts is necessary.

## Kappa Statistic (k)

The kappa coefficient is a statistical measure of inter-rater reliability. It is generally thought to be a more robust measure than simple percent agreement calculation since κ takes into account the agreement occurring by chance. Cohen's kappa measures the agreement between two raters.[5]

The equation for κ is:

$$\Pr(a)–\Pr(e)/1–\Pr(e)$$

where Pr(a) is the relative observed agreement among raters, and Pr(e) is the probability that agreement is due to chance.

If the raters are in complete agreement then $\kappa = 1$. If there is no agreement among the raters (other than what would be expected by chance) then $\kappa \leq 0$ (see Fig. 14.1). Note that Cohen's kappa measures agreement between two raters only. For a similar measure of agreement when there are more than two raters Fleiss' kappa is used.[5] An example of the use of the kappa statistic is shown in Fig. 14.2

## Definitions

### *Pre-test Probability*

The pre-test probability (likelihood) that a disease of interest is present or not, is the index of suspicion for a diagnosis, ***before*** the test of interest is performed. This

| Kappa | Strength of agreement |
|---|---|
| 0.00 | Poor |
| 0.01-0.20 | Slight |
| 0.21-0.40 | Fair |
| 0.41-0.60 | Moderate |
| 0.61-0.80 | Substantial |
| 0.81-1.00 | Almost perfect |

**Fig. 14.1** Strength of agreement using the kappa statistic

| | | Doctor A | | |
|---|---|---|---|---|
| | | No | Yes | Total |
| Doctor B | No | 10 (34.5%) | 7 (24.1%) | 17 (58.6%) |
| | Yes | 0 (0.0%) | 12 (41.4%) | 12 (41.4%) |
| Total | | 10 (34.5%) | 19 (65.5%) | 29 |

Kappa = (Observed agreement - Chance agreement)/
(1-Chance agreement)
Observed agreement = (10 +12)/29 = 0.76
Chance agreement = 0.586 * 0.345 + 0.655 * 0.414 = 0.474
Kappa = (0.76 – 0.474)/(1 – 0.474) = 0.54

**Fig. 14.2** An example of the use of the kappa statistic

index of suspicion is influenced by the prevalence of the disease in the population of patients you are evaluating. Intuitively, one can reason that with a rare disease (low prevalence) that even with a high index of suspicion, you are more apt to be incorrect regarding the disease's presence, than if you had the same index of suspicion in a population with high disease prevalence.

## *Post-test Probability and Test Ascertainment*

The post-test probability is your index of suspicion ***after*** the test of interest has been performed. Let's further explore this issue as follows. If we construct a 2 × 2 table (Table 14.1) we can define the following variables: If disease is present and the test is positive, that test is called a true positive (TP) test (this forms the definition of test sensitivity – that is the percentage of TP tests in patients with the index disease). If the index disease is present and the test is negative, that is called a false negative (FN) test. Thus patients with the index disease can have a TP or FN result (but by definition cannot have a false positive – FP, or a true negative – TN result).

## *Sensitivity and Specificity*

The sensitivity of a test then can be written as TP/TP + FN. If the index disease is not present (i.e. it is absent) and the test is negative, this is called a true negative (TN) test (this forming the definition of specificity – that is the percentage of TN's in the absence of disease). The specificity of a test can then be written as TN/TN + FP. Finally, if disease is absent and the test is positive one has a false positive (FP)

**Table 14.1** The relationship between disease and test result

| | Abnormal test | Normal test |
|---|---|---|
| Disease present | TP | FN |
| Disease absent | FP | TN |

test. Note that the FP percentage is 1-specificity (that is, if the specificity is 90% – in 100 patients without the index disease, 90 will have a negative test, which means 10 will have a positive test – i.e. FP is 10%).

### *Predictive Value*

Another concept is that of the predictive value (PV+ and PV–) of a test. This is asking the question differently than what sensitivity and specificity address – that is rather than asking what the TP and TN rate of a test is, the PV+ of a test result is asking how likely is it that a positive test is a true positive (TP)? i.e. TP/TP + FP (for PV- it is TN/TN + FN).

## Ways of Determining Test Accuracy and/or Clinical Usefulness

There are at least six ways of determining test accuracy and they are all interrelated so the determination of which to use is based on the question being asked, and one's personal preference. They are:

Sensitivity and specificity
2 × 2 tables
Predictive value
Bayes formula of conditional probability
Likelihood ratio
Receiver Operator Characteristic curve (ROC)

### *Bayes Theorem*

We have already discussed sensitivity and specificity as well as the tests predictive value, and the use of 2 × 2 tables; and, examples will be provided at the end of this chapter. But, understanding Bayes Theorem of conditional probability will help provide the student interested in this area with greater understanding of the concepts involved. First let's discuss some definitions and probabilistic lingo along with some shorthand. The conditional probability that event A occurs given population B is written as P(A|B). If we continue this shorthand, sensitivity can be written as P(T+/D+) and PV+ as P(D+/T+). Bayes' Formula can be written then as follows: The post test probability of disease =

*(Sensitivity)(disease prevalence)*
(Sensitivity)(disease prevalence) + (1-specificity)(disease absence)
or
*P(D±/T±) = p(T±/D±)(prevalence D±)*
p(T+/D+)(prevalence D+)p(T+/D-)p(D-)

where p(D+/T+) is the probability of disease given a T+ (otherwise known as PV+), p(T+/D+) is the shorthand for sensitivity, pT+/D- is the FP rate or 1-specificity. Some axioms apply. For example, one can arbitrarily adjust the "cut-point" separating a positive from a negative test and thereby change the sensitivity and specificity. However, any adjustment that increases sensitivity (this then increases ones comfort that they will not "miss" any one with disease as the false negative rate necessarily falls) will decrease specificity (that is the FP rate will increase – recall 1-specificity is the FP rate). An example of this is using the degree of ST segment depression during an electrocardiographic exercise test that one has determined will identify whether the test will be called "positive" or "negative". The standard for calling the ST segment response as positive is 1 mm of depression from baseline, and in the example in Table 14.2 this yields a sensitivity of 62% and specificity of 89%. Note what happens when one changes the definition of what a positive test is, by using 0.5 mm ST depression as the cut-point for calling test positive or negative. Another important axiom is that the prevalence of disease in the population you are studying does not significantly influence the sensitivity or specificity of a test (to derive those variables the denominators are defined as subjects with or without the disease i.e. if you are studying a population with a 10% disease prevalence one is determining the sensitivity of a test – against a gold standard – only in those 10%). In contrast, PV is very dependent on disease prevalence because more individuals will have a FP test in populations with a disease prevalence of 10% than they would if the disease prevalence was 90%. Consider the example in Table 14.3.

## *Receiver Operator Characteristic Curves (ROC)*

The ROC is another way of expressing the relationship between sensitivity and specificity (actually 1-specificity). It plots the TP rate (sensitivity) against the FP rate over a range of "cut-point" values. It thus provides visual information on the

**Table 14.2** Pre vs post-test probability

| Prev = 10% of 100 patients, | | Se = 70%, Sp = 90% |
|---|---|---|
| | T+ | T− |
| D+ | 7/10 (TP) | 3/10 (FN) |
| D− | 9/90 (FP) | 81/90 (TN) |
| | | PV+7/16 = 44% (10%→ 44%) |
| | | PV−81/84 = 97% (90%→ 96%) |

"trade off" between sensitivity and specificity, and the area under the curve (AUC) of a ROC curve is a measure of overall test accuracy (Fig. 14.3). ROC analysis was born during WW II as a way of analyzing the accuracy of sonar detection of submarines and differentiating signals from noise.[6] In Fig. 14.4, a theoretic "hit" means a submarine was correctly identified, and a false alarm means that a noise was incorrectly identified as a submarine and so on. You should recognize this figure as the equivalent of the table above discussing false and true positives.

Another way to visualize the tradeoff of sensitivity and specificity and how ROC curves are constructed is to consider the distribution of test results in a population. In Fig. 14.5, the vertical line describes the threshold chosen for a test to be called positive or negative (in this example the right hand curve is the distribution of subjects within the population that have the disease, the left hand curve those who do

**Table 14.3** Pre vs post-test probability

| Prev = 50% in 100 patients, | | Se = 70%, Sp = 90% |
|---|---|---|
| | T+ | T− |
| D+ | 0.7 × 50 = 35 (TP) | 0.3 × 50 = 15 (FN) |
| D− | 0.1 × 50 = 5 (FP) | 0.9 × 50 = 45 (TN) |
| | | PV+ 35/40 = 87% |
| | | PV− 45/60 = 75% |

$$P(D+T+)=\frac{0.7(0.5)}{0.7(0.5)+1\text{-}0.9(0.5)}=\frac{0.35}{0.35+0.05}=0.87$$

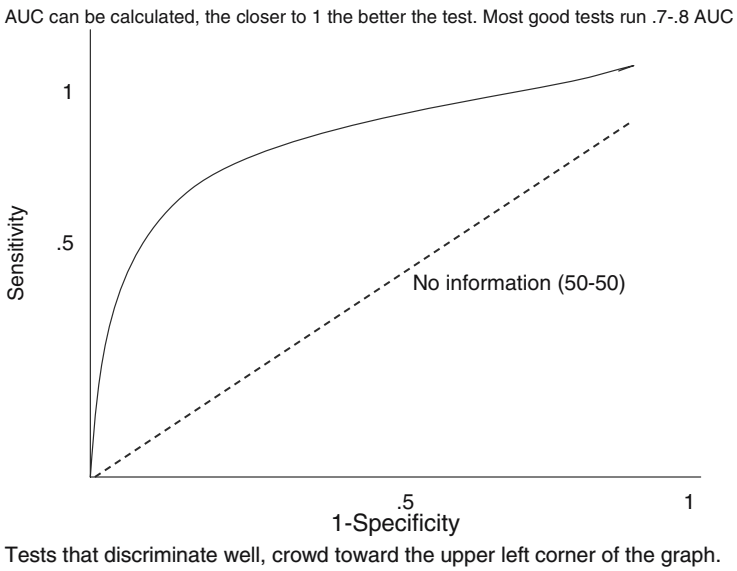



**Fig. 14.3** AUC can be calculated, the closer to 1 the better the test. Most good tests run 0.7–0.8 AUC

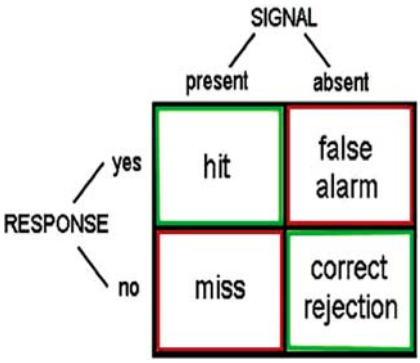



**Fig. 14.4** Depiction of true and false responses based upon the correct sonar signal for submarines

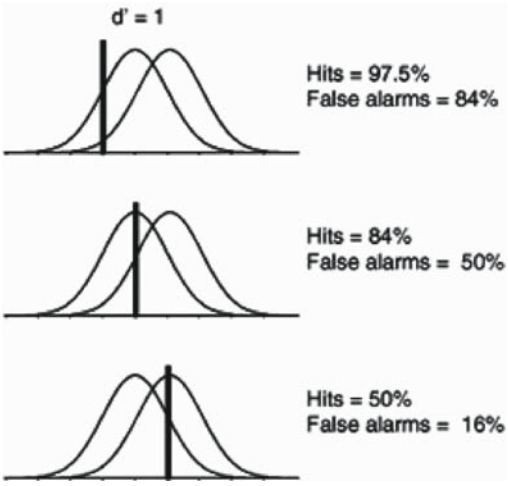



**Fig. 14.5** Demonstrates how changing the threshold for what divides true from false signals affects ones interpretation

not have the disease). The uppermost figure is an example of choosing a very low threshold value for separating positive from negative. By so doing, very few of the subjects with disease (recall the right hand curve) will be missed by this test (i.e. the sensitivity is high – 97.5%), but notice that 84% of the subjects without disease will also be classified as having a positive test (false alarm or false + rate is 84% and the specificity of the test for this threshold value is 16%). By moving the vertical line (threshold value) we can construct different sensitivity to false + rates and construct a ROC curve as demonstrated in Fig. 14.6.

As mentioned before, ROC curves also allow for an analysis of test accuracy (a combination of TP and TN), by calculating the area under the curve as shown in the figure above. Test accuracy can also be calculated by dividing the TP and TN by all possible test responses (i.e. TP, TN, FP, FN) as is shown in Fig. 14.4. The way ROC curves can be used during the research of a new test, is to compare the new test to existent tests as shown in Fig. 14.7.

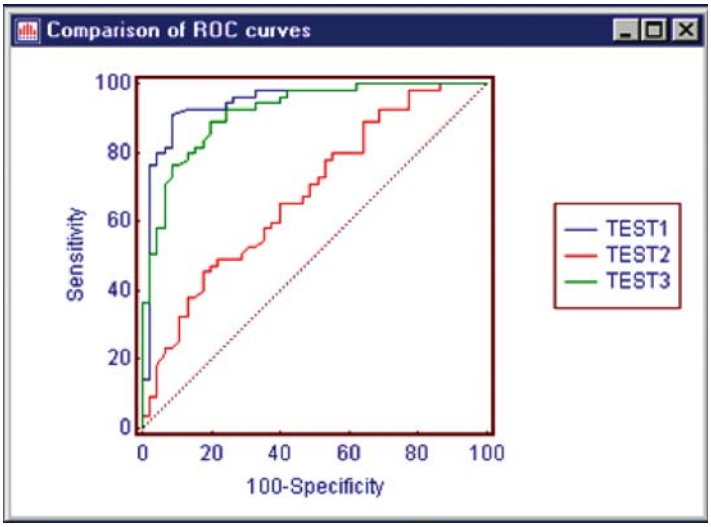
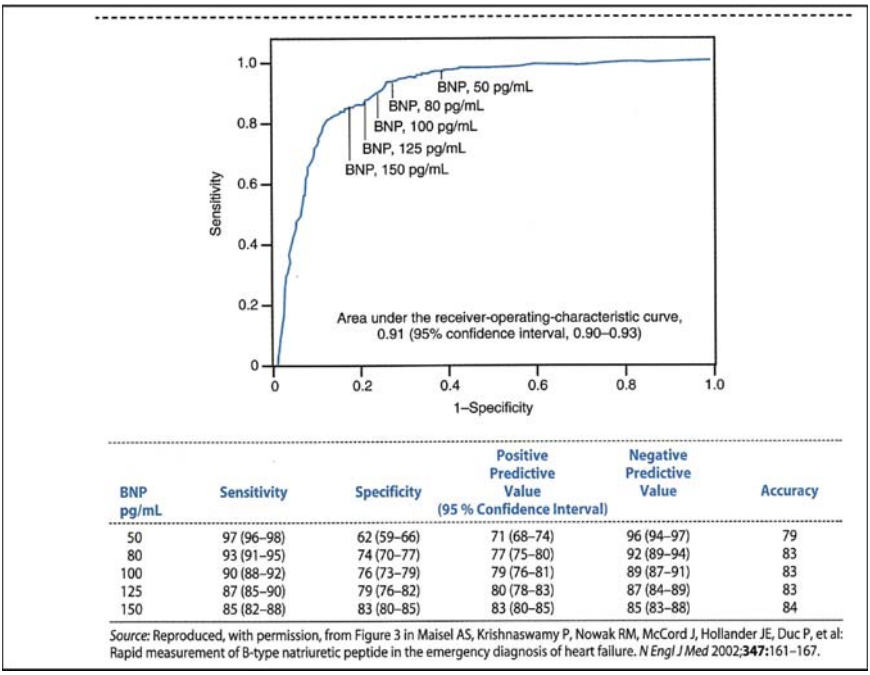



**Fig. 14.6** Comparison of ROC curves




| BNP pg/mL | Sensitivity | Specificity | Positive Predictive Value (95 % Confidence Interval) | Negative Predictive Value | Accuracy |
|---|---|---|---|---|---|
| 50 | 97 (96–98) | 62 (59–66) | 71 (68–74) | 96 (94–97) | 79 |
| 80 | 93 (91–95) | 74 (70–77) | 77 (75–80) | 92 (89–94) | 83 |
| 100 | 90 (88–92) | 76 (73–79) | 79 (76–81) | 89 (87–91) | 83 |
| 125 | 87 (85–90) | 79 (76–82) | 80 (78–83) | 87 (84–89) | 83 |
| 150 | 85 (82–88) | 83 (80–85) | 83 (80–85) | 85 (83–88) | 84 |

*Source:* Reproduced, with permission, from Figure 3 in Maisel AS, Krishnaswamy P, Nowak RM, McCord J, Hollander JE, Duc P, et al: Rapid measurement of B-type natriuretic peptide in the emergency diagnosis of heart failure. *N Engl J Med* 2002;**347:**161–167.

**Fig. 14.7** Box 12-1. Receiver operating characteristic curve for cutoff levels of B-type natriuretic peptide in differentiating between dyspnea due to congestive heart failure and dyspnea due to other causes

## *Likelihood Ratios*

Positive and Negative Likelihood Ratios (PLR and NLR) are another way of analyzing the results of diagnostic tests. Essentially, PLR is the odds that a person with a disease would have a particular test result, divided by the odds that a person without disease would have that result. In other words, how much more likely is a test result to occur in a person with disease than a person without disease. If one multiplies the pretest odds of having a disease by the PLR, one obtains the posttest odds of having that disease. The PLR for a test is calculated as the tests sensitivity/1-specificity (i.e. FP rate). So a test with a sensitivity of 70% and a specificity of 90% has a PLR of 7 (70/1-90). Unfortunately, it is made a bit more complicated by the fact that we generally want to convert odds to probabilities. That is, the PLR of 7 is really an odds of 7 to 1 and that is more difficult to interpret than a probability. Recall that odds of an event are calculated as the number of events occurring, divided by the number of events *not* occurring (i.e. non events, or p/p-1,). So if blood type O occurs in 42% of people, the odds of someone having a blood type of O are 0.42/1-0.42 i.e. the odds of a randomly chosen person having blood type O is 0.72:1. Probability is calculated as the odds/odds + 1, so in the example above 0.72/1.72 = 42% (or 0.42 – that is one can say the odds have having blood type O is 0.72 to 1 or the probability is 42% – the latter is easier to understand for most). Recall, that probability is the extent to which something is likely to happen. To review, take an event that has a 4 in 5 probability of occurring (i.e. 80% or 0.8). The odds of its occurring is 0.8/1-0.8 or 4:1. Odds then, are a ratio of probabilities. Note that an odds ratio (often used in the analysis of clinical trials) is also a ratio of odds.

To review:

The likelihood ratio of a positive test (LR+) is usually expressed as

Sensitivity/1-Specificity

and the LR- is ***usually*** expressed as

1-Sensitivity/Specificity

If one has estimated a pretest odds of disease, one can multiply that odds by the LR to obtain the post test odds, i.e.:

Post-test odds = pre-test odds × LR

To use an exercise test example consider the sensitivity for the presence of CAD (by coronary angiography) based on 1 mm ST segment depression. In this aforementioned example, the sensitivity of a "positive" test is 70% and the specificity is 90% (PLR = 7; NLR = 0.33). Let's assume that based upon our history and physical exam we feel the chance of a patient having CAD before the exercise test is 80% (0.8). If the exercise test demonstrated 1 mm ST segment depression, your post-test odds of CAD would be 0.8 × 7 or 5.6 (to 1). The probability of that patient having CAD is then 5.6/1 + 5.6 = 0.85 (85%). Conversely if the exercise test did not demonstrate 1 mm ST segment depression the odds that the patient did not have CAD is 0.33 × 7 = 2.3 (to 1) and the probability of his not having CAD is 70%. In other words *before* the exercise test there was an 80% chance of CAD, while *after* a positive test it was 85%. Likewise before the test, the chance of the patient not having CAD was 20%, and if the test was negative it was 70%.

To add a bit to the confusion of using LRs. there are two lesser used derivations of the LR as shown in Table 14.4. One can usually assume that if not otherwise designated, the descriptions for PLR and NLR above apply. But, if one wanted to

**Table 14.4** Pre vs post-test probabilities

| Clinical presentation | Pre test P (%) | Post test P (%) | T + Post test F (%) |
|---|---|---|---|
| Typical angina | 90 | 98 | 75 |
| Atypical angina | 50 | 88 | 25 |
| No symptoms | 10 | 44 | 4 |

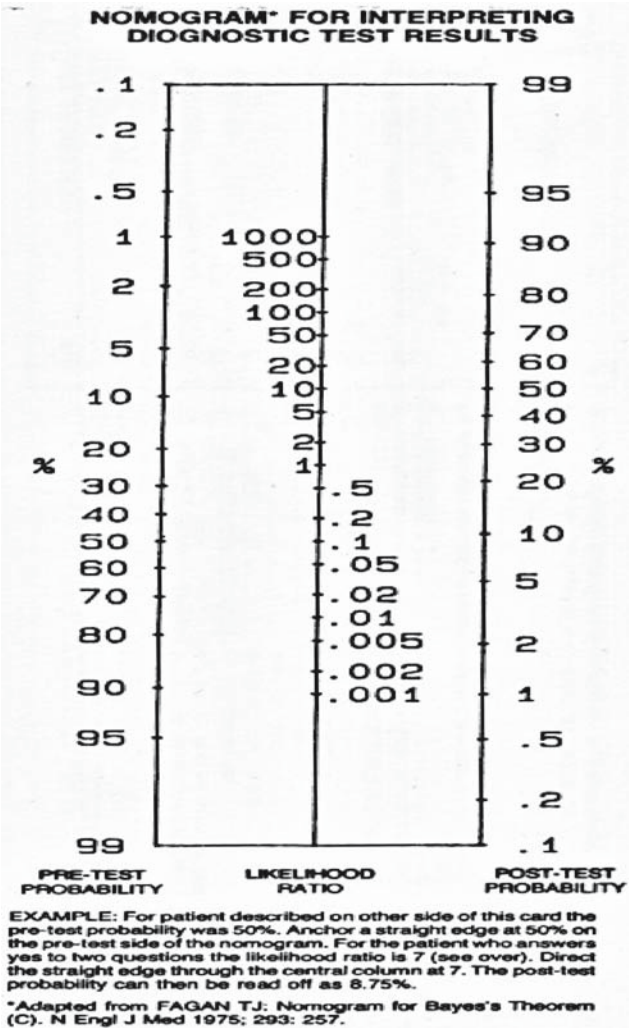



**Fig. 14.8** Nomogram for interpreting diagnostic test results (Adapted from Fagan[8])

**Table 14.5** Calcium Scores: Se, Sp, PPV and NPV

| CAC | Se % | Sp % | PPV % | NPV % |
|---|---|---|---|---|
| Age 40 to 49 | | | | |
| 1 | 88 | 61 | 17 | 98 |
| 100 | 47 | 94 | 42 | 95 |
| 300 | 18 | 97 | 60 | 93 |
| Age 60 to 69 | | | | |
| 1 | 100 | 26 | 41 | 100 |
| 300 | 74 | 81 | 67 | 86 |
| 700 | 49 | 91 | 74 | 78 |

CAC=Calcium artery Scores; Se=Senvitivity, Sp=Specificity, PPV=Positive predictive value, NPV=negative predictive value. Adapted from Miller DD

express the results of a negative test in terms of the chance that the patient **has** CAD (despite a negative test) rather than the chance that he **does not** have disease given a negative test; or wanted to match the NLR with NPV (i.e. the likelihood that the patient does NOT have the disease given a negative test result) an alternative definition of NLR can be used (of course one could just as easily subtract 70% form 100% to get that answer as well). To make things easier, a nomogram can be used in stead of having to do the calculations Fig. 14.8

In summary, the usefulness of diagnostic data depends on making an accurate diagnosis based upon the use of diagnostic tests, whether the tests are radiologic, laboratory based, or physiologic. The questions to be considered by this approach include: “How does one know how good a test is in giving you the answers that you seek?”, and “What are the rules of evidence against which new tests should be judged?” Diagnostic data can be sought for a number of reasons including: diagnosis, disease severity, to predict the clinical course of a disease, to predict therapy response. That is, what is the probability my patient has disease x, what do my history and PE tell me, what is my threshold for action, and how much will the available tests help me in patient management. An example of the use of diagnostic research is provided by Miller and Shaw.[7] From Table 14.5, one can see how the coronary artery calcium (CAC) score can be stratified by age and the use of the various definitions described above.

## References

1. Bayes T. An essay toward solving a problem in the doctrine of chances. *Phil Trans Roy Soc London*. 1764; 53:370–418.
2. Ledley RS, Lusted LB. Reasoning foundations of medical diagnosis; symbolic logic, probability, and value theory aid our understanding of how physicians reason. *Science*. July 3 1959; 130(3366):9–21.
3. Redwood DR, Borer JS, Epstein SE. Whither the ST segment during exercise. *Circulation*. Nov 1976; 54(5):703–706.

4. Rifkin RD, Hood WB, Jr. Bayesian analysis of electrocardiographic exercise stress testing. *N Engl J Med*. Sept 29, 1977; 297(13):681–686.
5. McGinn T, Wyer PC, Newman TB, Keitz S, Leipzig R, For GG. Tips for learners of evidence-based medicine: 3. Measures of observer variability (kappa statistic). *CMAJ*. Nov 23, 2004; 171(11):1369–1373.
6. Green DM, Swets JM. *Signal Detection Theory and Psychophysics*. New York: Wiley; 1966.
7. Miller DD, Shaw LJ. Coronary artery disease: diagnostic and prognostic models for reducing patient risk. *J Cardiovasc Nurs*. Nov–Dec 2006; 21(6 Suppl 1):S2–16; quiz S17–19.
8. Fagan TJ. Nomogram for Bayes's theorem (C). *N Engl J Med*. 1975; 293:257.

# Part III

This Part addresses statistical concepts important for the clinical researcher. It is not a Part that is for statisticians, but rather approaches statistics from a basic foundation standpoint.

*Statistician:* *Oh, so you already have calculated the p-value?*
*Surgeon:* *Yes, I used multinomial logistic regression.*
*Statistician:* *Really? How did you come up with that?*
*Surgeon:* *Well, I tried each analysis on the SPSS drop-down menu, and that was the one that gave the smallest p-value.*

Vickers A, Shoot first and ask questions later. Medscape Bus Med. 2006; 7(2), posted 07/26/2006

# Chapter 15
# Statistical Power and Sample Size: Some Fundamentals for Clinician Researchers

**J. Michael Oakes**

*Surgeon: Say, I've done this study but my results are disappointing.*
*Statistician: How so?*
*Surgeon: The p-value for my main effect was 0.06.*
*Statistician: And?*
*Surgeon: I need something less than 0.05 to get tenure.*

**Abstract** This chapter aims to arm clinical researchers with the necessary conceptual and practical tools (1) to understand what sample size or power analysis is, (2) to conduct such analyses for basic low-risk studies, and (3) to recognize when it is necessary to seek expert advice and input. I hope it is obvious that this chapter aims to serve as a general guide to the issues; specific details and mathematical presentations may be found in the cited literature. Additionally, it should be obvious that this discussion of statistical power is focused, appropriately, on quantitative investigations into real or hypothetical effects of treatments or interventions. It does not address *qualitative* study designs. The ultimate goal here is to help practicing clinical researcher *get started* with power analyses.

## Introduction

My experience as both and educator and collaborator is that clinical researchers are frequently perplexed if not unnerved by questions of statistical power, detectable effect, number-needed-to-treat, sample size calculations, and related concepts. Those who have taken a masters-level biostatistics course may even become paralyzed by authoritative cautions, supporting the quip that a little knowledge can be a dangerous thing. Unfortunately, anxiety and misunderstanding seem to push some to ignore the issues while others appear rigid in their interpretations, rejecting all 'under-powered' studies as useless. Neither approach is helpful to researchers or medical science.

I do not believe clinician researchers, especially, are to blame for the trouble. My take is that when it comes to statistical power and related issues, instructors, usually

S.P. Glasser (ed.), *Essentials of Clinical Research*,

biostatisticians, are too quick to present equations and related algebra instead of the underlying concepts of uncertainty and inference. Such presentations are understandable since the statistically-minded often *think* in terms of equations and are obviously equipped with sufficient background information and practice to make sense of them. But the same is not usually true of clinicians or perhaps even some epidemiologists. Blackboards filled with Greek letters and algebraic expressions, to say nothing of terms like 'sampling distribution,' only seem to intimidate if not turn-off students eager to understand and implement the ideas. What is more, I have come across strikingly few texts or articles aimed at helping clinician-researchers understand key issues. Most seem to address only experimental (e.g., drug trial) research, offer frightening cautions, or consider only painfully simple studies. Little attention is paid to less glorious but common clinical studies such as sample-survey research or perhaps the effects of practice/cultural changes to an entire clinic. Too little has written about the conceptual foundations of statistical power, and even less of this is tailored for clinician-researchers.

I find that clinical researchers gain a more useful understanding of, and appreciation for, the concepts of statistical power when the ideas are first presented with some utilitarian end in mind, and when the ideas are located in the landscape of inference and research design. Details and special-cases are important, but an emphasis must be placed on simple and concrete examples relevant to the audience. Mathematical nuance and deep philosophical issues are best reserved for the few who express interest. Still, I agree with Baussel and Li[1] who write,

> … a priori consideration of power is so integral to the entire design process that its consideration should not be delegated to individuals not integrally involved in the conduct of an investigation…

Importantly, emphasis on concepts and understanding may also be sufficient for clinical researchers since I believe the following three points are critical to a successful power-analysis:

1. *The More, the Merrier* – Except for exceptional cases when study subjects are exposed to more than minimal risk, there is hardly any *pragmatic* argument for not enrolling as many subjects as the budget permits. Over-powered studies are not much of a threat, especially when authors and readers appreciate the abundant limitations of p-values and other summary measures of 'significance.' While perhaps alarming, I have found analytic interest in subgroup comparisons or other 'secondary' aims to be universal; few researchers are satisfied when 'real' analyses are limited to main study hypotheses. It follows that more subjects are always needed. But let me be clear: when risk is elevated, clinical researchers must seek expert advice.
2. *Use Existing Software* – Novice study designers should rely on one or more of the high-quality and user-friendly software packages available for calculating statistical power. Novice researchers should not attempt to derive new equations nor should they attempt to implement any such equation into a spreadsheet package. The possibility of error is too great and efforts to 're-invent the wheel' will likely lead to mistakes. Existing software packages have been tested and will

give the correct answer, provided researchers input the correct information. This means, of course, that the researcher must understand the function of each input parameter and the reasonableness of the values entered.

3. *If No Software, Seek Expert* – If existing sample-size software cannot accommodate a particular study design or an analysis plan, novice researchers should seek expert guidance from biostatistical colleagues or like-minded scholars. Since existing software accommodates many (sophisticated) analyses, exceptions mean something unusual must be considered. Expert training, experience, and perhaps an ability to simulate data are necessary in such circumstances. Expert advice is also necessary when risks of research extend beyond the minimal threshold.

The upshot is that clinical researchers need to minimally know what sort of sample size calculation they need and, at most, what related information should be entered into existing software. Legitimate and accurate *interpretation* of output is then paramount, as it should be. Concepts matter most here, and are what seem to be retained anyway.[2]

Accordingly, this chapter aims to arm clinical researchers with the necessary conceptual and practical tools (1) to understand what sample size or power analysis is, (2) to conduct such analyses for basic low-risk studies, and (3) to recognize when it is necessary to seek expert advice and input. I hope it is obvious that this chapter aims to serve as a general guide to the issues; specific details and mathematical presentations may be found in the cited literature. Additionally, it should be obvious that this discussion of statistical power is focused, appropriately, on quantitative investigations into real or hypothetical effects of treatments or interventions. I do not address *qualitative* study designs. The ultimate goal here is to help practicing clinical researcher *get started* with power analyses. Alternative approaches to inference and 'statistical power' continue to evolve and merit careful consideration if not adoption, but such a discussion is far beyond the simple goals here; see.[3,4]

## Fundamental Concepts

### *Inference*

Confusion about statistical power often begins with a misunderstanding about the point of conducting research. In order to appreciate the issues involved in a power calculation, one must appreciate that the goal of research is to draw credible inferences about a phenomena under study. Of course, drawing credible inferences is difficult because of the many errors and complications that can cloud or confuse our understanding. Note that, ultimately, power calculations aim to clarify and quantify some of these potential errors.

To make issues concrete, consider patient A with systolic pressure of 140 mm Hg and, patient B, with a reading of 120 mm Hg. Obviously, the difference between

these two readings is 20 mm Hg. Let us refer to this difference as '*d*'. To sum up, we have

$$140-120 = d$$

Now, as sure as one plus one equals two, the measured difference between the two patient's BPs is 20. Make no mistake about it, the difference is 20, not more, not less.

So, what is the issue? Well, as any clinician knows either or both the blood-pressure measures could (probably do!) incorporate error. Perhaps the cuff was incorrectly applied or the clinician misread the sphygmomanometer. Or perhaps the patient suffers white-coat hypertension making the office-visit measure different from the patient's 'true' measure. Any number of measurement errors can be at work making the calculation of the observed difference between patients an error-prone measure of the true difference, symbolized by Δ, the uppercase Greek-letter 'D', for True or philosophically perfect difference.

It follows that what we actually measure is a mere *estimate* of the thing we are trying to measure, the True or parameter value. We measure blood-pressures in both patients and calculate a difference, 20, but no clinician will believe that the true or real difference in pressures between these two individuals is precisely 20 now or for all time. Instead, most would agree that the quantity 20 is an estimate of the true difference, which we may believe is 20, plus or minus 5 mm Hg, or whatever. And that this difference changes over time if not place.

This point about the observed difference of 20 being an estimate for the true difference is key. One takes measures, but appreciates that imprecision is the rule. How can we gauge the degree of measurement error in our estimate of $d = 20 \rightarrow \Delta$?

One way is to take each patient's blood-pressures (BP) multiple times and, say, average them. It may turn out that patient A's BP was measured as 140, 132, 151, 141 mm Hg, and patient B might have measures 120, 121, 123, 119, 117. The average of patient A's four measurements is, obviously, 141 mm Hg, while patient B's five measurements yield an average of 120 mm Hg. If we use these presumably more accurate average BPs, we now have this

$$141-120 = 21 = d^*$$

where d* is used to show that this 'd' is based on a different calculation (e.g., averages) than the previously discussed 'd'.

How many estimates of the true difference do we need to be comfortable making claims about it? Note that the p-value from the appropriate t-test is less than 0.001. What does this mean? Should we take more measures? How accurate do we need the difference in blood pressure to be before we are satisfied that patient A's BP is higher than patient B's? Should we worry that patient A's BP was much more variable (standard deviation = 7.8) than patient B's (standard deviation = 2.2)? If patient A is male and patient B female, can we generalize and say that, on average, males have high BP than females? If we are a little wrong about the differences in

blood pressures, which is more important: claiming there is no difference when in fact there is one, or claiming there is a difference when in fact there is not one? It is questions like these that motivate our discussion of statistical power.

The basic goal of a 'power analysis' is to appreciate *approximately* how many subjects are needed to detect a *meaningful* difference between two or more experimental groups. In other words, the goal of power analysis is to consider natural occurring variance of the outcome variable, errors in measurement, and the impact of making certain kinds of inferential errors (e.g., claiming a difference when in truth the two persons or groups are identical). Statistical power calculations are about inference, or making (scientific) leaps of faith from real-world observations to statements about the underlying truth.

Notice above, that I wrote 'approximately.' This is neither a mistake nor a subtle nuance. Power calculations are useful to determine if a study needs 50 or 100 subjects; the calculations are not useful in determining whether a study needs 50 or 52 subjects. The reason is that power calculations are loaded with assumptions, too often hidden, about distributions, measurement error, statistical relationships and perfectly executed study designs. As mentioned above, it is rare for such perfection to exist in the real world. Believing a given power analysis is capable of differentiating the utility of a proposed study within a degree of handful of study subjects is an exercise in denial and is sure to inhibit scientific progress.

I also wrote that power was concerned with differences between 'two groups.' Of course study designs with more groups are possible and perhaps even desirable. But power calculations are best done by keeping comparisons simple, as when only two groups are involved. Furthermore, this discussion centers on elementary principles and so simplicity is paramount.

The other important word is 'meaningful'. It must be understood that power calculations offer nothing by way of *meaning*; manipulation of arbitrary quantities through some algebraic exercise is a meaningless activity. The meaningfulness of a given power calculation can only come from scientific/clinical expertise. To be concrete, while some may believe a difference of, say, 3 mm Hg of systolic blood pressure between groups is important enough to act on, others may say such a difference is not meaningful *even if* it is an accurate measure of difference. The proper attribution of meaningfulness, or perhaps importance or utility, requires extra-statistical knowledge. Clinical expertise is paramount.

## *Standard Errors*

A fundamental component of statistical inference is the idea of 'standard error.' As an *idea*, a standard error can be thought of as the standard deviation of a test statistic in the sampling distribution. You may be asking, what does this mean?

Essentially, our simplified approach to inference is one of replicating a given study over and over again. This replication is not actually done, but is instead a though experiment, or theory that motivates inference. The key is to appreciate that

for each hypothetical and otherwise identical study we observe a treatment effect or some other outcome measure. Because of natural variation and such, for some studies the test statistic is small/low, for others, large/high. Hypothetically, the test statistic is distributed in a bell-shaped curve, with one point/measure for each hypothetical study. This distribution is called the *sampling distribution*. The standard deviation (or spread) of this sampling distribution is the standard error of the test statistic. The smaller the standard deviation, the smaller the standard error.

We *calculate* standard errors in several ways depending on the study design and the chosen test statistics. Standard error formulas for common analytic estimators (i.e., tests) are shown in Fig. 15.1. Notice the key elements of each standard error formula are the variance of the outcome measure, $\sigma^2$, and sample size, $n$. Researchers must have a sound estimate of the outcome measure variance at planning. Reliance on existing literature and expertise is a must. Alternative approaches are discussed by Browne.[5]

Since smaller standard errors are usually preferred (as they imply a more precise test statistic), one is encouraged to use quality measurement tools and/or larger sample sizes.

## *Hypotheses*

A fundamental idea is that of the 'hypothesis' or 'testable conjecture.' The term 'hypothesis' may be used synonymously with 'theory'. A necessary idea here is that the researcher has a reasoned and *a priori* guess or conjecture about the outcome

| Estimator | Standard Error |
|---|---|
| Sample mean | $\sqrt{\sigma^2/n}$ |
| Difference between independent sample means | $\sqrt{\sigma^2\left(1/n_1 + 1/n_2\right)}$ |
| Binomial proportion | $\sqrt{p(1-p)/n}$ |
| Log Odds-ratio | $\sqrt{1/a + 1/b + 1/c + 1/d}$ |
| Difference between two means in a Group-randomized trial | $\sqrt{\frac{2\left[\sigma^2/m + \tau^2\right]}{g}}$ |

**Fig. 15.1** Common standard error formulas

of their analysis or experiment. The *a priori* (or in advance) aspect is critical since power is done in the planning stage of a study.

For purposes here, hypotheses may be of just two types: the null and the alternative. The null hypothesis is, oddly, what is *not expected* from the study. The alternative hypothesis is what is expected given one's theory. This odd reversal of terms or logic may be a little tricky at first but everyone gets used to it. Regardless, the key idea is that researchers marshal information and evidence from their study to either confirm or disconfirm (essentially reject) their *a priori* null hypothesis. For us, a study is planned to test a theory by setting forth a null and alternative hypothesis and evaluating data/results accordingly. Researchers will generally be glad to observe outcomes that refute null hypotheses.

Several broad kinds of hypotheses are important for clinical researchers but two merit special attention:

1. Equality of groups – The null hypothesis is that the, say, mean in the treatment group is strictly equal to the mean in the control group; symbolically $\mu_T = \mu_C$, where $\mu_I$ represents the mean of the treatment group and $\mu_C$ represents the mean of the control group. The analysis conducted aims to see if the treatment is strictly different from control; symbolically $\mu_T \neq \mu_C$. As can be imagined, this strict equality or difference hypothesis is not much use in the real world.
2. Equivalence of groups – In contrast to the equality designs, equivalence designs do not consider just any difference to be important, even if statistically significant! Instead, equivalence studies require that the identified difference be clinically meaningful, above some pre-defined value, *d*. The null hypothesis in equivalence studies is that the (absolute value of) the difference between treatment and control groups be larger than some meaningful value; symbolically, $|\mu_T - \mu_C| \geq d$. The alternative hypothesis is then that the observed difference is smaller than the predefined threshold value *d*, or in symbols $|\mu_T - \mu_C| < d$. If the observed is less than *d*, then two 'treatments' are viewed as equivalent, though this does not mean strictly equal.

Finally, it is worth pointing out that authors typically abbreviate the term null hypothesis with $H_0$ and the alternative hypothesis with $H_A$.

## *Type I and Type II Error*

When it comes to elementary inference, it is useful to define two kinds of errors. Using loose terms, we may call them errors of commission and omission, with respect to stated hypotheses.

Errors of commission are those of inferring a relationship between study variables when in fact there is not one. In other words, errors of commission are rejecting a null hypothesis (no relationship) when in fact it should have been accepted it. In other words, you have done something you should not have.

Errors of omission are those of not inferring a relationship between study variables when in fact there is a relationship. In other words, not rejecting a null in favor

of the alternative, when in fact the alternative (a relationship) was correct. That is, you have failed to do something you should have.

The former – errors of commission – are called Type I errors. The latter, Type II errors. A simple figure is useful for understanding their inter-relationship, as shown in Fig. 15.2. Statistical researchers label Type I error α, the Greek letter 'a' or alpha. Type II errors are labeled β, the Greek letter 'b' or beta (the first and second letters of the Greek alphabet).

Both Type I and Type II errors are quantified as probabilities. The probability of incorrectly rejecting a true null hypothesis – or accepting that there is a relationship when in fact there is not – is $\alpha$ (ranging from 0 to 1). So, Type I error may be 0.01, 0.05 or any other such value. The same goes for Type II error.

For better or worse, by convention researchers typically plan studies with an Type I error rate of 0.05, or 5%, and a Type II error rate of 0.20 (20%) or less. Notice this implies that making an error of commission (5% alpha or Type I error) is four times *more* worrisome than making an error of omission (20% beta or Type II error). By convention, we tolerate less Type I error than Type II error. Essentially, this relationship reflects the conservative stance of science: scientists should accept the null (no relationship) unless there is strong evidence to reject it and accept the alternative hypothesis. That is the scientific method.

## *Statistical Power*

We can now define statistical power. Technically, power is the complement of the Type II error (i.e., the difference between 1 and the amount of Type II error in the study). A simple definitional equation is,

$$\text{Power} = 1 - \beta$$

| | **Mother Nature or True State of Null Hypothesis** | |
|---|---|---|
| **Researcher's Inference** | $H_0$ is True | $H_0$ is False |
| Reject $H_0$ | ***Type I error*** *probability =* $\alpha$ | Correct Inference *probability =* $1-\beta$ *Power* ($H_A$) |
| Accept $H_0$ | Correct Inference *probability =* $1-\alpha$ | ***Type II error*** *probability =* $\beta$ |

**Fig. 15.2** Type I and Type II errors

Statistical power is, therefore, about the probability of correctly rejecting a null hypothesis when in fact one should do so. It is a population parameter, loosely explained as a study's ability or strength to reject the null when doing so is appropriate. In other words, power is about a study's ability to find a relationship between study variables (e.g., treatment effect on mortality) when in fact there is such a relationship. Note that power is a function of the alternative hypothesis; which essentially means that the larger the (treatment) effect, the more power to detect it. It follows that having more power is usually preferred since researchers want to discover new relationships between study variables. Insufficient power means some existing relationships go undetected. This is why underpowered studies are so controversial; one cannot tell if there is in fact no relationship between two study variables or whether the study was not sufficiently powered to detect the relationship; inconclusive studies are obviously less than desirable.

Given the conventional error rates mentioned above (5% Type I and 20% Type II) we can now see where and why the conventional threshold of 80% power for a study obtains: it is simply

$$\text{Power} = 1-\beta = 1-0.20 = 0.80$$

To be clear, 80% statistical power means that if everything in the study goes as planned and the alternative hypothesis in fact is true, there is an 80% chance of observing a statistically significant result and a 20% chance of erroneously missing it. All else equal, lower Type II error rates mean more statistical power.

## *Power and Sample Size Formula*

There are a large number of formulae and approaches to calculating statistical power and related concepts, and many of these are quite sophisticated. It seems useful however to write down a/the very basic formula and comment on it. Such foundational ideas serve as building blocks for more advanced work. The basic power formula may be written as,

$$Z_{1-\alpha/2} + Z_{power} = \frac{\Delta}{SE(\Delta)}$$

where $Z_{\alpha/2}$ is the value of Z for a given $\alpha/2$ Type I error rate, $Z_{Power}$ is the value of Z for a given power value (i.e., 1 – Type II error rate), $\Delta$ is the minimal detectable effect for some outcome variable (discussed below), and $SE(\Delta)$ is the standard error for the same outcome variable.

Let us now explore each of the four (just four!) basic elements in more detail. In short, the equation states that the (transformed) probability of making the correct inference equals the effect of some intervention divided by the appropriate standard error.

The term $Z_{\alpha/2}$ is the value of a Z statistic (often found in the back of basic statistics textbooks) for the Type I error rate divided by two, for a two-sided hypothesis test. If Type I error rate is 0.05, the value of this element is 0.975. Looking up the value of Z shows that the Z at 0.975 is 1.96.

The term $Z_{Power}$ is the value of the Z statistic, a specified level of power. Type II error is often set a 20% (or 0.20), which yields a $Z_{Power}$ of 0.84.

We may now rewrite the equation for use when two-sided Type I error is 5% and power is set at 80% (Type II error is 20%),

$$1.96 + 0.84 = \frac{\Delta}{SE(\Delta)}$$

The other two elements in the equation above depend on the data and/or theory. The critical part is the standard error of the outcome measure, symbolized as $SE(\Delta)$. This quantity depends on the study design and the variability of the outcome measure under investigation. If may be helpful to regard this quantity as the noise that is recorded in the outcome measure. Less noise means a more precise outcome measure; and, the more precision the better.

It should now be easy to see that the key part of the formula is the standard error, and thus two elements really drive statistical power calculations: variance of the outcome measure, $\sigma^2$, and sample size, $n$. The rest is more or less given, although the importance of the study design and statistical test cannot be over emphasized. It follows that for any given design researchers should aim to decrease variance and increase sample size. Doing either or both reduces the minimal detectable effect, $\Delta$, which is generally a good thing.

## *Minimal Detectable Effect*

As mentioned above, applied or collaborating statisticians rarely directly calculate the statistical power of a given study design. Instead, we typically ask clinician researchers how many subjects can be recruited given budget constraints and then using the conventional thresholds of 80% power and 5% Type I error rates calculate the study's minimum detectable difference.[6] In other words, given that (1) most find 80% power and 5% Type I error satisfactory and (2) that budgets are always tight, there is no point in calculating power or how many subjects are needed. Instead the values of 80%, 5%, and number of subject's affordable, along with the variance and other information are taken as given or immutable. The formula is algebraically manipulated to yield the smallest or minimal study effect (on the scale of the outcome measure) that is to be expected.

$$\Delta = SE(\Delta)\left[Z_{1-\alpha/2} + Z_{power}\right].$$

For the conventional Type I and II error rates, the formula is simply

$$\Delta = SE(\Delta) * 2.8$$

If this value is clinically meaningful – that is, not as large as to be useless – then the study is well-designed. Notice, one essentially substitutes any appropriate standard error. Again, standard errors are a function of study design (cross-sectional, cohort, or experiment study, etc.) It is worth noting that there are some subtle but important aspects to this approach; advanced learners may begin with the insights of Greenland.[7]

## *P-Values and Confidence Interval*

P-values and confidence intervals are practically related and convey a sense of uncertainty about an effect estimate. There remains a substantial degree of controversy about the utility or misuse of p-values as a measure of meaning,[8–10] but the key idea is that some test statistic, perhaps Z or t, which is often the ratio of some effect estimate divided by its standard error, is assessed against a threshold value in a Z-table, say Z of 0.05 which is 1.96. If the ratio of the effect estimate divided by its standard error is greater than 1.96 (which is 1.96 standard deviations away from mean of the sample distribution) then we say the estimated effect is unlikely to arise by chance if the null hypothesis were in fact true… that is, the estimated effect is statistically significant.

Confidence intervals, often called 95% confidence intervals, are another measure of uncertainty about estimated effects.[11] Confidence intervals are often written as the estimated mean or other statistic of the effect plus or minus some amount, such as 24±11, which is to say the lower 95% confidence interval is 24 − 11 = 13 and the upper 95% confidence interval is 24 + 11 = 35. In other words, in 95 out of 100 replications of the study being conducted, the confidence interval will include (or cover) the true mean (i.e., parameter). Confidence intervals are too often erroneously interpreted as saying that there is a 95% probability of the true mean being within the limit bounds.

## Two Worked Examples

The benefits of working through a few common examples seem enormous. In what follows I offer two different ‘power analyses’ for common study designs: the first is a t-test for a difference between two group means, the second example considers an odds-ratio from a case-control study. I rely on the PASS software package for each analysis.[12] There are other programs that yield similar results and I do not mean to suggest PASS is the best. But I do rely on it personally and find it user-friendly.

Two points must be emphasized before proceeding: (1) power analyses are always tailored to a particular study design and null hypothesis and (2) use of existing software is beneficial, but if study risks are high then expert guidance is necessary.

## *(Example 1) T-Test with Experimental Data*

Imagine a simple randomized experiment where 50 subjects are given some treatment (the treatment group) and 50 subjects are not (the control or comparison group). Researchers might be interested in the difference in the mean outcome of some variable between groups. Perhaps we are interested in the difference in body mass index (BMI) between some diet regime and some control condition. Presume that it is known from pilot work and the existing literature that the mean BMI for the study population is 28.12 with a standard deviation of 7.14.

Since subjects were randomized to groups there is no great concern with confounding. A simple t-test between means will suffice for the analysis. Our null hypothesis is that the difference between means is nil; our alternative hypothesis is that the treatment group mean will be different (presumably but not necessarily less) than the control group mean.

Since we could only afford a total of N = 100 subjects, there is no reason to consider altering this. Additionally, we presume that in order to publish the results in a leading research journal we need 5% Type I error and 20% Type II error (or what is the same, 80% Power). The question is, given the design and other constraints, how small an effect of the treatment can we detect? Inputting the necessary information into a software program is easy. The PASS screen for this analysis is shown in Fig. 15.3.

Notice that we are solving for 'Mean 2 (Search < Mean 1)' which implies that we are looking for the difference between our two sample means, where the second mean is less than the first or visa versa. Again, the alternative hypothesis is that our treatment group BMI mean will be different from the control groups, which is a non-directional or two-sided test. The specification here merely adds a sign (+ or −) to the estimated treatment effect. The question at hand is how small an effect can we minimally detect?

- We have given error rates for 'Power' to be 0.80 and our 'Alpha (Significance)' to be 0.05.
- The sample size we have is 50 for 'N1 (sample size Group 1)' and the same for 'N2 (sample size Group 2)'. Again, we presume these are given due to budget constraints.
- The mean of group 1 'Mean1 (Mean of Group 1)' is specific at 28.12, a value we estimated from our expertise and the existing literature. We are solving for the mean of group two 'Mean2 (Mean of Group 2)'.
- The standard deviation of BMI also comes from the literature and is thought to be 7.14 for our target population (in the control or non-treatment arm). We assume

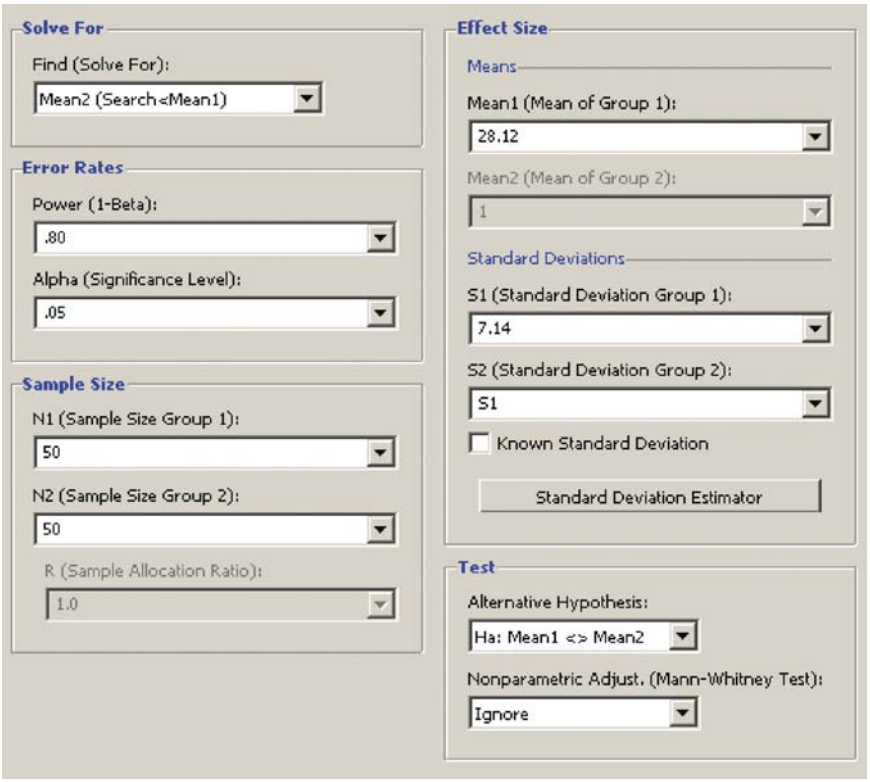



**Fig. 15.3** PASS input screen for t-test analysis

that the standard deviation for the treatment arm will be identical to S1 or 7.14. Again, these are hypothetical values for this discussion only.

- The alternative hypothesis under investigation is that the means are unequal. This framework yields a two-sided significance test, which is almost always indicated.

Clicking the 'run' button (top left) yields this PASS screen seen in Fig. 15.4, which is remarkably self-explanatory and detailed. The output shows that for 80% Power, 5% alpha or Type I error, two-sides significance test, 50 subjects per group, and a mean control-group BMI of 28.1 with a standard deviation of 7.1, we can expect to minimally detect a difference of 4.1 BMI units (28.1 − 24.1 = 4.0). To be clear, we have solved for Δ and it is 4.0. Given this design, we have an 80% chance to detect a 4.0 unit difference in BMI if in fact that difference exists. If our treatment actually has a larger impact on BMI, we will have more power to detect it.

If this change of 4.0 BMI units between treatment groups is thought to be possible and is clinically meaningful, then we have a well-designed study. If we can only hope for a 2.1 unit decrease in BMI from the intervention, then we are under-powered and should alter the study design. Possible changes include more subjects and or

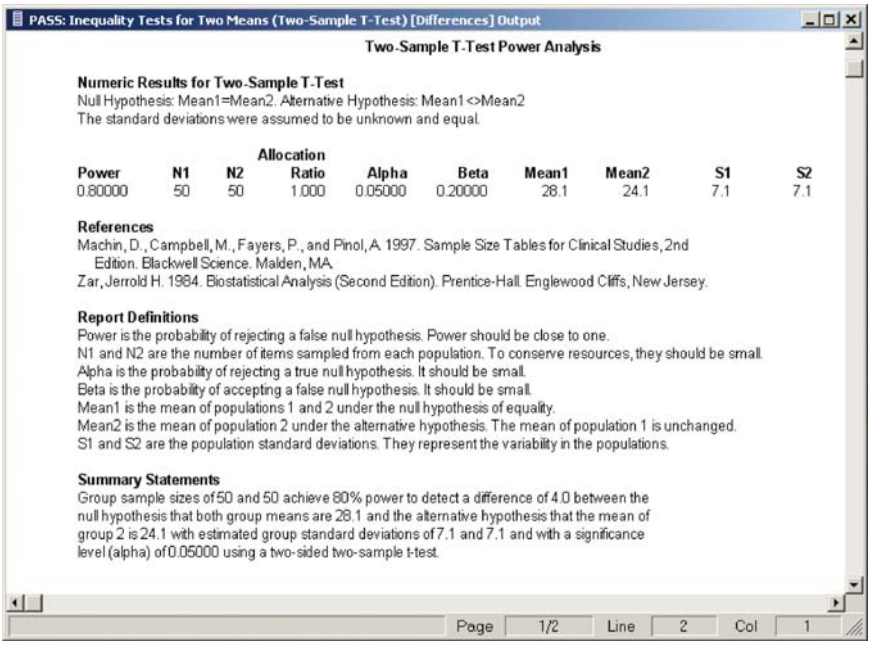



**Fig. 15.4** PASS output for t-test power analysis

reducing the standard deviation of the outcome measure BMI, presumably by using a more precise instrument, or perhaps stratifying the analysis.

It is worth noting that more experienced users may examine the range of minimal detectable differences possible over a range of sample sizes or a range of possible standard deviations. Such 'sensitivity' analyses are very useful for both investigators and peer-reviewers.

## *(Example 2) Logistic Regression with Case-Control Data*

The second example is for a (hypothetical) case-control study analyzed with a logistic regression model. Here again we navigate to the correct PASS input screen (Fig. 15.5) and input our desired parameters:

- Solve for an odds-ratios, expecting the exposure to have a positive impact on the outcome measure; in other words OR > 1.0
- Power = 80% and Type I or alpha error = 5%
- Let sample size vary from N = 100 to N = 300 by 25 person increments
- Two sided hypothesis test
- Baseline probability of exposure (recall this is case-control) of 20%

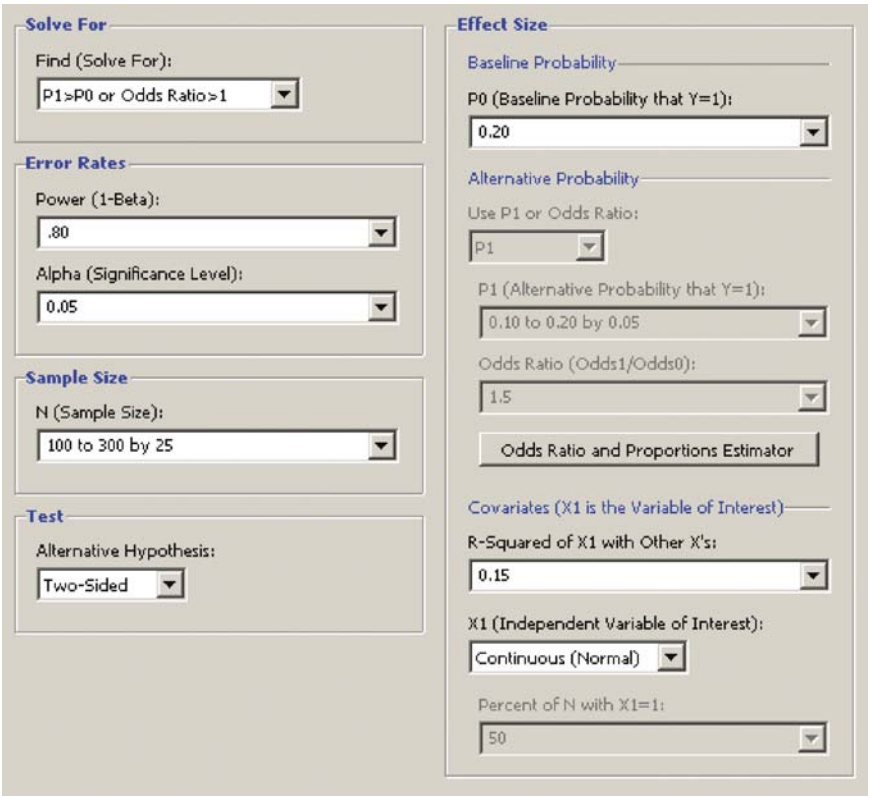


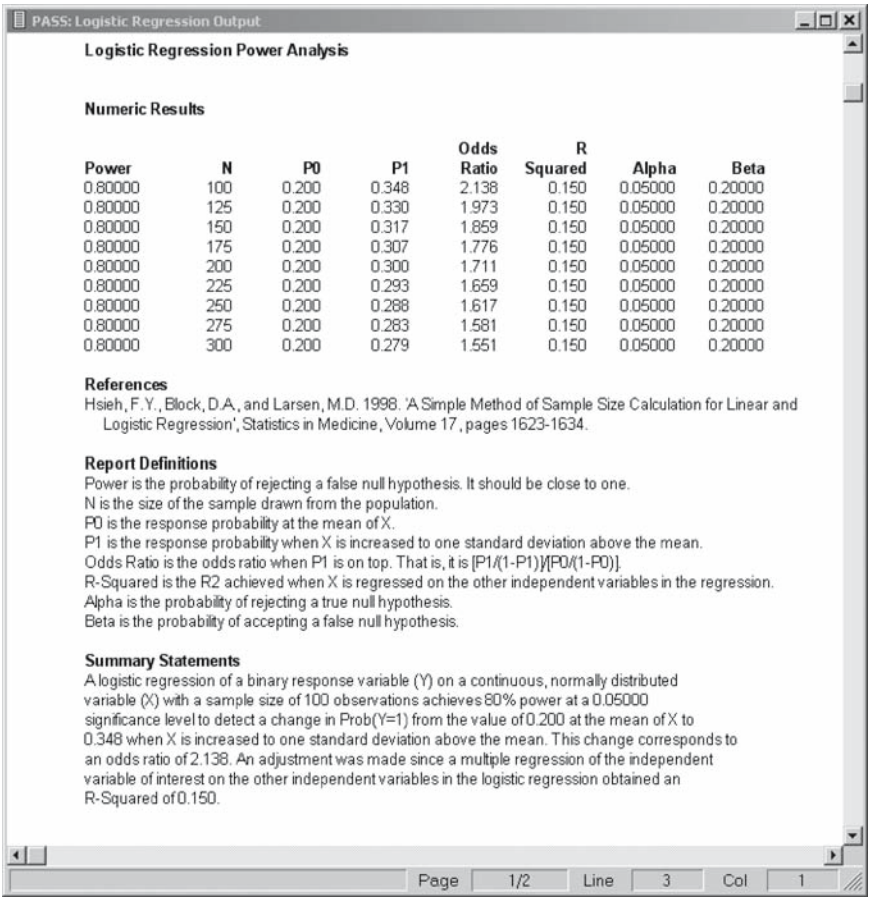

**Fig. 15.5** PASS input screen

And the explanatory influence of confounders included in the model is 15%.

But given the *range* of sample size values we specified, the output screen is shown in Fig. 15.6.

Given the null hypothesis of no effect (OR = 1.0), it is easy to see that the minimum detectable difference of exposure in this case-control study with N = 100 subject is 0.348 − 0.200 = 0.148, which is best understood as an OR = 2.138. With 300 subjects the same parameter falls to 1.551. As expected, increasing sample size (three fold) decreases the smallest effect one can expect to detect. Again, practically speaking, the smaller the better.

One can copy the actual values presented into a spreadsheet program (e.g., Microsoft Excel) and graph the difference in odds-ratios (that is, Δ) as a function of sample size. Reviewers tend to prefer such 'sensitivity' analyses. When it comes to such simple designs, this is about all there is to it, save for proper interpretation of course.


PASS: Logistic Regression Output

**Logistic Regression Power Analysis**

**Numeric Results**

| Power | N | P0 | P1 | Odds Ratio | R Squared | Alpha | Beta |
|---|---|---|---|---|---|---|---|
| 0.80000 | 100 | 0.200 | 0.348 | 2.138 | 0.150 | 0.05000 | 0.20000 |
| 0.80000 | 125 | 0.200 | 0.330 | 1.973 | 0.150 | 0.05000 | 0.20000 |
| 0.80000 | 150 | 0.200 | 0.317 | 1.859 | 0.150 | 0.05000 | 0.20000 |
| 0.80000 | 175 | 0.200 | 0.307 | 1.776 | 0.150 | 0.05000 | 0.20000 |
| 0.80000 | 200 | 0.200 | 0.300 | 1.711 | 0.150 | 0.05000 | 0.20000 |
| 0.80000 | 225 | 0.200 | 0.293 | 1.659 | 0.150 | 0.05000 | 0.20000 |
| 0.80000 | 250 | 0.200 | 0.288 | 1.617 | 0.150 | 0.05000 | 0.20000 |
| 0.80000 | 275 | 0.200 | 0.283 | 1.581 | 0.150 | 0.05000 | 0.20000 |
| 0.80000 | 300 | 0.200 | 0.279 | 1.551 | 0.150 | 0.05000 | 0.20000 |

**References**

Hsieh, F.Y., Block, D.A., and Larsen, M.D. 1998. 'A Simple Method of Sample Size Calculation for Linear and Logistic Regression', Statistics in Medicine, Volume 17, pages 1623-1634.

**Report Definitions**

Power is the probability of rejecting a false null hypothesis. It should be close to one.
N is the size of the sample drawn from the population.
P0 is the response probability at the mean of X.
P1 is the response probability when X is increased to one standard deviation above the mean.
Odds Ratio is the odds ratio when P1 is on top. That is, it is [P1/(1-P1)]/[P0/(1-P0)].
R-Squared is the R2 achieved when X is regressed on the other independent variables in the regression.
Alpha is the probability of rejecting a true null hypothesis.
Beta is the probability of accepting a false null hypothesis.

**Summary Statements**

A logistic regression of a binary response variable (Y) on a continuous, normally distributed variable (X) with a sample size of 100 observations achieves 80% power at a 0.05000 significance level to detect a change in Prob(Y=1) from the value of 0.200 at the mean of X to 0.348 when X is increased to one standard deviation above the mean. This change corresponds to an odds ratio of 2.138. An adjustment was made since a multiple regression of the independent variable of interest on the other independent variables in the logistic regression obtained an R-Squared of 0.150.

Page 1/2 Line 3 Col 1

**Fig. 15.6** PASS output screen

## Conclusions

Sample size and statistical power are important issues for clinical research and it seems clinical researchers continue to struggle with the basic ideas. Accordingly, this chapter has aimed to introduce some fundamental concepts too often ignored in the more technical (i.e., precise) literature. Abundant citations are offered for those seeking more information or insight.

In closing, five points merit emphasis. First, sound inference comes from well-designed and executed studies. Planning is the key. Second, power analyses are always directly linked to a particular design and analysis (i.e., null hypothesis). General power calculations are simply not helpful, correct, and may even lead to disaster. Third, while used throughout this discussion, I emphasize that I do not advocate for the conventional 80% power and 5% Type I error. I simply use these

above as common examples. Error rates should be carefully considered. Power analyses are properly done in the planning stage of a study. Retrospective power analyses are to be avoided.[13] Fourth, assumptions of planned analyses are key. Multiple comparisons and multiple hypothesis tests undermine power calculations and assumptions. Further, interactive model specification (i.e., data mining) invalidates assumptions. Finally, cautions of when to consult a statistical expert are important, especially when research places subjects at risk.

For greater technical precision and in-depth discussion, interested readers are encouraged to examine the following texts, ordered from simplest to more demanding discussions [1,14,15]: A solid and more technical recent but general discussion is by Maxwell, Kelley and Rausch.[16] Papers more tailored to particular designs include Oakes and Feldman,[17] Feldman and McKinlay,[18] Armstrong[19]; Greenland[20]; Self and Mauritsen.[21] Of note is that Bayesian approaches to inference continue to evolve and merit careful study if not adoption by practicing statisticians.[3] Because the approach incorporates *a priori* beliefs and is focused on decision-making under uncertainty, the Bayesian approach to inference is actually a more natural approach to inference in epidemiology and clinical medicine.

**Acknowledgements** This chapter would not have been possible without early training by Henry Feldman and the outstanding comments and corrections of Peter Hannan.

## References

1. Baussell RB, Li Y-F. *Power Analysis for Experimental Research: A Practical Guide for the Biological, Medical and Social Sciences*. New York: Cambridge; 2002 (p ix).
2. Herman A, Notzer N, Libman Z, Braunstein R, Steinberg DM. Statistical education for medical students–concepts are what remain when the details are forgotten. *Stat Med*. Oct 15, 2007; 26(23):4344–4351.
3. Berry DA. Introduction to Bayesian methods III: use and interpretation of Bayesian tools in design and analysis. *Clin Trials*. 2005; 2(4):295–300; discussion 301–294, 364–278.
4. Berry DA. Bayesian statistics. *Med Decis Making*. Sep–Oct 2006; 26(5):429–430.
5. Browne RH. Using the sample range as a basis for calculating sample size in power calculations. *Am Statistician*. 2001; 55:293–298.
6. Bloom HS. Minimum detectable effects: a simple way to report the statistical power of experimental designs. *Evaluat Rev*. Oct 1995; 10(5):547–556.
7. Greenland S. Power, sample size and smallest detectable effect determination for multivariate studies. *Stat Med*. Apr–June 1985; 4(2):117–127.
8. Poole C. Low P-values or narrow confidence intervals: which are more durable? *Epidemiology*. May 2001; 12(3):291–294.
9. Savitz DA, Tolo KA, Poole C. Statistical significance testing in the American Journal of Epidemiology, 1970–1990. *Am J Epidemiol*. May 15, 1994; 139(10):1047–1052.
10. Sterne JA. Teaching hypothesis tests–time for significant change? *Stat Med*. Apr 15, 2002; 21(7):985–994; discussion 995–999, 1001.
11. Greenland S. On sample-size and power calculations for studies using confidence intervals. *Am J Epidemiol*. July 1988; 128(1):231–237.
12. Hintz J. PASS 2008, NCSSLLC. www.ncss.com.
13. Hoenig JM, Heisey D. The abuse of power: the pervasive fallacy of power calculations for data analysis. *Am Stat*. 2001; 55:19–24.

14. Chow S-C, Shao J, Wang H. *Sample Size Calculations in Clinical Research*. New York: Marcel Dekker; 2003.
15. Lipsey M. *Design Sensitivity: Statistical Power for Experimental Research*. Newbury Park, CA: Sage; 1990.
16. Maxwell SE, Kelly K, Rausch JR. Sample size planning for statistical power and accuracy in parameter estimation. *Ann Rev Psychol*. 2008; 59:537–563.
17. Oakes JM, Feldman HA. Statistical power for nonequivalent pretest-posttest designs. The impact of change-score versus ANCOVA models. *Eval Rev*. Feb 2001; 25(1):3–28.
18. Feldman HA, McKinlay SM. Cohort versus cross-sectional design in large field trials: precision, sample size, and a unifying model. *Stat Med*. Jan 15, 1994; 13(1):61–78.
19. Armstrong B. A simple estimator of minimum detectable relative risk, sample size, or power in cohort studies. *Am J Epidemiol*. Aug 1987; 126(2):356–358.
20. Greenland S. Tests for interaction in epidemiologic studies: a review and a study of power. *Stat Med*. Apr–June 1983; 2(2):243–251.
21. Self SG, Mauritsen RH. Power/sample size calculations for generalized linear models. *Biometrics*. 1988; 44:79–86.

# Chapter 16
# Association, Cause, and Correlation

**Stephen P. Glasser and Gary Cutter**

*The star of the play is the effect size i.e. what you found*
*The co-star is the effect size7's confidence interval i.e. the precision that you found*
*If needed, supporting cast is the adjusted analyses i.e. the exploration of alternative explanations*
*With a cameo appearance of the p value, which, although its career is fading, insisted upon being included*
*Do not let the p value or an F statistic or a correlation coefficient steal the show, the effect size must take center stage!*
*But remember it takes an entire cast to put on a play!*

**Abstract** Anything one measures can become data, but only those data that have meaning can become information. Information is almost always useful, data may or may not be. This chapter will address the various ways one can measure the degree of association between an exposure and an outcome and will include a discussion of relative and absolute risk, odds ratios, number needed to treat, and related measures.

## Introduction

Types of data include dichotomous, categorical, and continuous. For finding associations in clinical research data, there are several tools available. For categorical analyses one can compare relative frequencies or proportions; cross classifications (grouping according to more than one attribute at the same time) offer three different kinds of percentages (row, column and joint probabilities) and to assess whether they are different from what one might expect by chance: chi square tests. When comparing continuous measures one can use correlation, regression, analysis of variance (ANOVA), and survival analyses. The techniques for continuous variables can also accommodate categorical data into their assessments.

## Relative Frequencies and Probability

Let's address relative frequencies first or how often something appears relative to all results. The simplest relative frequency can be a probability such as a rate (a numerator divided by what is in the numerator plus what is not in the numerator) i.e. A/A + B

S.P. Glasser (ed.), *Essentials of Clinical Research*,

(Influenza fatality rate: those who are infected with influenza and die denoted by A divided by those infected who die (A) plus those infected who recover (B). In contrast, a ratio is of the form A/B, where the numerator is not part of the denominator. Examples of rates are the probability of a 6 on the throw of a die (there is one 6 and 5 other 'points' on the die thus, the probability of a 6 is one out of six), or the probability of a winning number in roulette. Three key concepts in probability and associations are: joint probability, marginal probability, and conditional probability (i.e. probability of A occurs given B has already occurred). Figure 16.1 diagrams these three types of probabilities. These concepts are key to cross classifications of variables.

Dependence, is another way of saying association, and two events are dependent if the probability of A and B (A & B)_occurring is not equal to the probability of A times the probability of B. If the probability of A & B is equal to the product of the probability of A times the probability of B the two events are said to be independent. For example, there are four suits in a deck of cards, thus, the probability of drawing a card and it is a heart is ¼. There are four queens in a deck of cards, thus the probability of drawing a queen is 4/52. The probability of drawing the queen of hearts is 1 out of 52 cards which equals is ¼ times 4/52 = 1/52. Thus we can say that the suit of the card is independent of the face on the card. How does this apply to epidemiology and medical research? To illustrate, consider the 2 × 2 table shown in Table 16.1.

**Fig. 16.1** Some concepts of probability

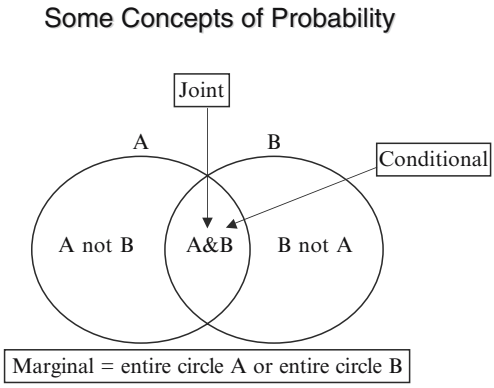



**Table 16.1** Cross classification: How to summarize or compare this data

| | Dx | No Dx | Total | |
|---|---|---|---|---|
| Exp | A | B | A + B | Conditional probability prob disease given exposure A/(A + B) |
| Not E | C | D | C + D | Conditional probability prob disease given no exposure C/(C + D) |
| Total | A + C↑ | B + D | N | |
| Marginal probability prob disease P(Dx) = (A + C)/N | | | Marginal probability prob exposure P(Exp) = (A + B)/N | |

By applying the above to an exploration of the association of hormone replacement therapy (HRT) and deep venous thrombosis (DVT) from the following theoretic data we can calculate the joint, marginal, and conditional probability as seen in Table 16.2.

To test the hypothesis of independence, we can use the above probability rules to determine how well the observed compares to the expected occurrence under the assumption that the HRT therapy is independent of the DVT. One way to do this is to use the chi square statistic (basically the observed value in any of the cells of our cross-classification minus the expected value squared, divided by the expected for each cell of our cross-classification table; and, add them up these squared deviations to achieve a test statistical value). If the statistical value that is calculated occurs by chance (found by comparing to the appropriate chi-square distribution table) is less than say 5% of the time we will reject the hypothesis that the row and column variables are independent, thereby implying that they are *not* independent i.e. an association exists. Any appropriately performed test of statistical significance lets you know the degree of confidence you can have in accepting or rejecting a hypothesis. Typically, the hypothesis tested with chi square is whether or not two different samples (of people, texts, whatever) are different enough in some characteristic or aspect of their behavior that we can say from our sample that the populations from which our samples are drawn appear to different from the expected behavior.

A non-parametric test (specific distributions of values are not specified *a priori* some assumptions are made such as independent and identically distributed values) makes fewer assumptions about form of the data occurring, but compared to parametric tests (like t-tests and analysis of variance, for example) is less powerful or likely to identify an association and, therefore, has less status in the list of statistical tests. Nonetheless, its limitations are also its strengths; thus, because chi- square is more 'forgiving' in the data it will accept, it can be used in a wide variety of research contexts.

**Table 16.2** The two by two table HRT and DVT

| | DVT | No DVT | Total |
|---|---|---|---|
| On HRT | 33 | 1,667 | 1,700 |
| No HRT | 27 | 2,273 | 2,300 |
| Total | 60 | 3,940 | 4,000 |
| Marginal probability of DVT | | = 60/4,000<br>= 0.0150 or 1.50% | On HRT = 1,700/4,000<br>= 0.4250 or 42.50% |
| Conditional probability of DVT given you are on HRT | | = 33/1,700 = 0.0194 or 1.94% | |
| Conditional probability of DVT given you're not on HRT | | = 27/2,300 = 0.0117 or 1.17% | |
| Joint probability of HRT AND DVT | | = 33/4,000 = 0.0083 or 0.83% | |

## Generalizing from Samples to Populations

Converting raw observed values or frequencies into percentages does allow us to more easily see patterns in the data, but that is all we can see, unless we make some additional assumptions about the data. Knowing with great certainty how often a particular drink is preferred in a particular group of 100 students is of limited use; we usually want to measure a sample in order to infer something about the larger populations from which our samples were drawn. On the basis of raw observed frequencies (or percentages) of a sample's behavior or characteristics, we can make claims about the sample itself, but we cannot generalize to make claims about the population from which we drew our sample, unless we make some assumptions on how that sample was obtained and submit our results to quantification, so called inferential statistics and often to make inferences, a test of statistical significance. A test of statistical significance tells us how confidently we can generalize to a larger (unmeasured) population from a (measured) sample of that population (see Chapter 18).

How does the chi square distribution and test statistic allow us to draw inferences about the population from observations on a sample? The chi-square statistic is what statisticians call an enumeration statistic. Rather than measuring the value of each of a set of items, a calculated value of chi-square compares the frequencies of various kinds (or categories) of items assuming a random sample, to the frequencies that are expected if the population frequencies are as hypothesized by the investigator. Chi square is often called a 'goodness of fit' statistic. That is it compares the observed values to how well they fit what is expected in a random sample and what is expected under a given statistical hypothesis. For example, chi-square can be used to determine if there is a reason to reject the statistical hypothesis (the chance that it arose from the underlying model given that the expected frequency is so unlikely that we choose to assume the underlying model is incorrect). For example we might want to know that the frequencies in a random sample that are collected are consistent with items that come from a normal distribution. We can divide up the Normal distribution into areas, calculate how many items would fall within those areas assuming the normal model is correct and compare to how many fall in those areas from the observed values.

Basically then, the chi square test of statistical significance is a series of mathematical formula which compares the actual observed frequencies of some phenomenon (in a sample) with the frequencies we would expect. In terms of determining associations, we are testing the fit of the observed data to that expected if there were no relationships at all between the two variables in the larger (sampled) population, that in the card example above. The chi square tests our actual results against the null hypothesis that the items were the result of an independent process and assesses whether the actual results are different enough from what might occur just by sampling error.

## *Chi Square Requirements*

As mentioned before, chi square is a nonparametric test. It does not require the sample data to be more or less normally distributed (like parametric tests such as the

t-tests do), although it relies on the assumption that the variable is sampled randomly from an appropriate population.

But chi square, while forgiving, does have some requirements as noted-below:

1. The sample must be assumed to be randomly drawn from the population. As with any test of statistical significance, your data is assumed to be from a random sample of the population to which you wish to generalize your claims. While nearly never technically true, we make this assumption and must consider the implications of violating this assumption (i.e. a biased sample).
2. Data must be reported in raw frequencies (not percentages). One should only use the chi square when your data are in the form of raw frequency counts of things in two or more mutually exclusive and exhaustive categories. As discussed above, converting raw frequencies into percentages standardizes cell frequencies as if there were 100 subjects/observations in each category of the independent variable for comparability, but this is not to be used in calculation of the chi square statistic. Part of the chi square mathematical procedure accomplishes this standardizing, so computing the chi square on percentages would amount to standardizing an already standardized measurement and would always assume that there were 100 observations irrespective of the true number, thus in general would give the wrong answer except when there are exactly 100 observations.
3. Measured variables must be measured independently between people; That is, if we are measuring disease prevalence using sisters in the group, this may not be an independent assessment, since there may be strong familial risk of the disease. Similarly, using two mammograms from the same woman taken two years apart would not be an independent assessment.
4. Values/categories on independent and dependent variables must be mutually exclusive and exhaustive (each person or observation can only go into one place).
5. Expected frequencies cannot be too small. The computation of the chi-square test involves dividing the difference between the observed and expected value squared by the expected value. If the expected value were too small, this calculation could wildly distort the statistic. A general rule of thumb is that the expected must be greater than 1 and not more than 20% of the expected values should be less than 5.

We will discuss expected frequencies in greater detail later, but for now remember that expected frequencies are derived from observed frequencies under an independence model.

## Relative Risk and Attributable Risk

One of the more common measures of association is relative risk (RR) (Fig. 16.2). Relative Risk is the incidence of disease in one group compared to the other. As such, it is used as a measure of association in cohort studies and RCTs. Said in other ways, RR is the risk of an event (or of developing a disease) in one group relative to another; is a ratio of the probability of the event occurring in the exposed group versus the probability of the event occurring in the control (non-exposed) group.

$$RR = \frac{P \text{ exposed}}{P \text{ control}}$$

For example, if the probability of developing lung cancer among smokers was 20% and among non-smokers 1%, then the relative risk of cancer associated with smoking would be 20. Smokers would be 20 times as likely as non-smokers to develop lung cancer. Relative risk is used frequently in the statistical analysis of binary outcomes where the outcome of interest has relatively low probability. It is thus often an important outcome of clinical trials, where it is used to compare the risk of developing a disease say in people not receiving a new medical treatment (or receiving a placebo) versus people who are receiving a new treatment. Alternatively, it is used to compare the risk of developing a side effect in people receiving a drug as compared to the people who are not receiving the treatment (or receiving a placebo). A relative risk of 1 means there is no difference in risk between the two groups (since the null hypothesis is operative a RR implies no association between exposure and outcome) and the study then seeks to disprove that there is no association, the alternative hypothesis).

- A RR of <1 means the event is less likely to occur in the experimental group than in the control group.
- A RR of >1 means the event is more likely to occur in the experimental group than in the control group.

In the standard or classical hypothesis testing framework, the null hypothesis is that RR = 1 (the putative risk factor has no effect). The null hypothesis can be rejected in favor of the alternative hypothesis that the factor in question does affect risk (if the confidence interval for RR excludes 1, a so-called two sided test, since the RR can be

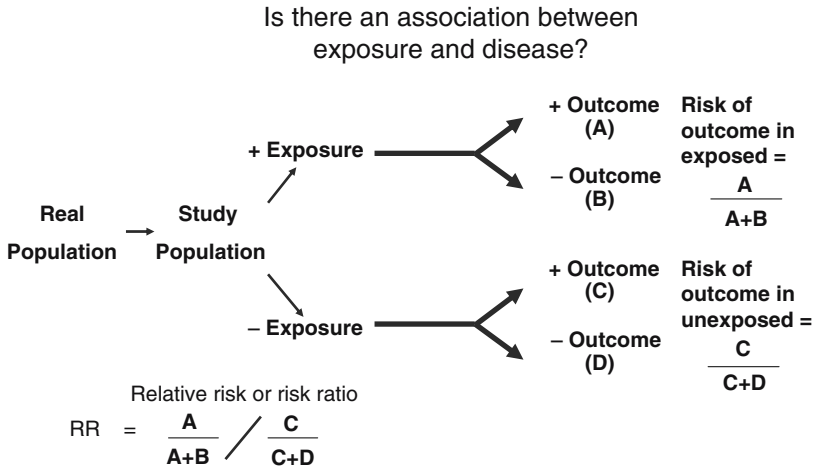



**Fig. 16.2** Is there an association between exposure and disease?

less than one or greater than 1). A RR of >2 suggests that the intervention is 'more likely than not' (also a legal term) responsible for the outcome. Since RR is a measure of incident cases, RR cannot be used in case control studies because case control studies begin with the identification of existent cases, and matches controls only to cases. With the RR one needs to know the incidence in the unexposed group, and because in case control studies the number of nonexposed cases is under the control of the investigator, there isn't an accurate denominator from which to compute incidence. This latter issue also prevents the use of RR even in prospective case-control studies (see discussion of case control study designs). In such situations, we use an approximation to the RR, called the Odds Ratio (OR discussed below), which for incident rates under 10% works very well.

Attributable risk (AR) is a measure of the excess risk that can be attributed to an intervention, above and beyond that which is due to other causes. When the AR exceeds 50%, it is about equivalent to a RR > 2. AR = incidence in the exposed group minus incidence in the unexposed divided by the incidence in the exposed. Thus, if the incidence of disease in the exposed group is 40% and in the unexposed is 10%, the proportion of disease that is attributable to the exposure is 75% (30/40). That is, 75% of the cases are due to the exposure. By the way, 'attributable' does not mean causal.

## Odds Ratio

Another common measure of association is the odds ratio (OR). As noted above, it is used in case control studies as an alternative to the RR. The OR is a way of comparing whether the probability of a certain event is the same for two groups, with an OR of 1 implying that the event is equally likely in both groups as with the RR. The odds of an event is a ratio; the occurrence of the event divided by the lack of its occurrence (Table 16.3). Commonly one hears in horse racing that the horse has to 1 odds of winning. This means that if the race were run five times, this horse is expected to win three times and lose one time. Another horse may have 2 to 1 odds. The odds ratio between the two horses would be 3/1 divided by 2/1 or 1.5. Thus, the odds ratio of the first horse winning to the second is 1.5.

$$\text{Odds ratio} = (Pi/(1-Pi))/(Pc/(1-Pc)$$

**Table 16.3** The odds ratio

| | Cancer | No Cancer |
|---|---|---|
| Exp | A | B |
| Not E | C | D |

The odds of cancer given exposure is A:B or A/B
The odds of cancer given no exposure is C:D or C/D
The odds ratio of cancer is: A/B divided by C/D
*O.R. = AD/BC*

The odds ratio approximates well the relative risk only when the probability of end-points (event rate or incidence) is lower than 10%. Above this threshold, the odds ratio will overestimate the relative risk. It is easy to verify the 'lower than 10%' rule. The relative risk from the odds ratio is:

$$\text{Relative risk} = \text{Odds ratio}/(1+Pc*(\text{Odds ratio} - 1)$$

Thus, for ORs larger than 1, the RR is less than or equal to the OR. The odds ratio has much wider use in statistics because of the approximation of the RR and the common use of logistic regression in epidemiology. Because the log of the odds ratio is estimated as a linear function of the explanatory variables, statistical models of the odds ratio often reflect the underlying mechanisms more effectively. When the outcome under study is relatively rare, the OR and RR are very similar in terms of their measures of association, but as the incidence of the outcome under study increases, the OR will underestimate the RR as shown in Fig. 16.3 taken from Zhang and Yu.[2]

Since relative risk is a more intuitive measure of effectiveness, the distinction above is important, especially in cases of medium to high event rates or probabilities. If action A carries a risk of 99.9% and action B a risk of 99.0% then the relative risk is just slightly over 1, while the odds associated with action A are almost 10

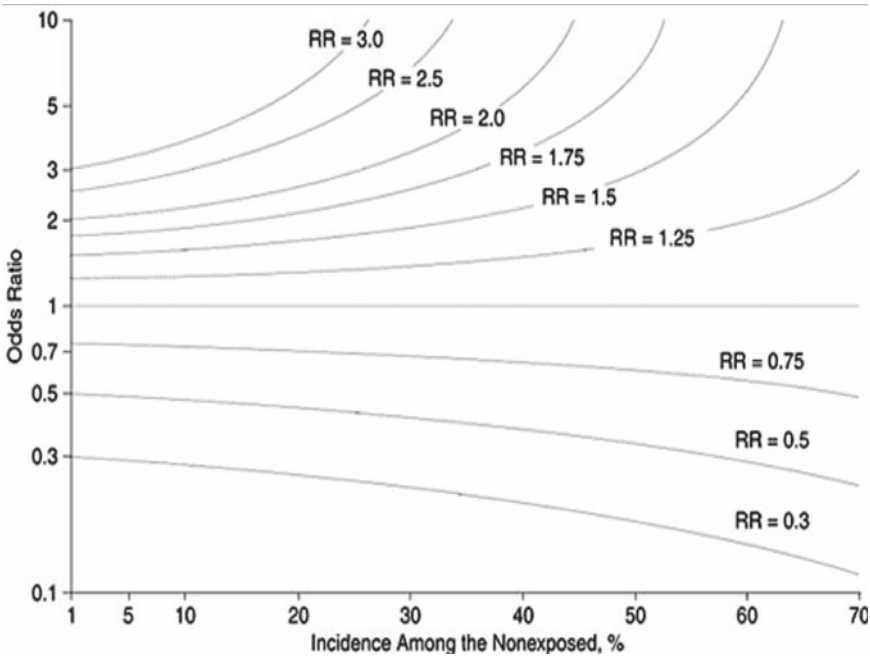



**Fig. 16.3** The relationship of RR to OR dependent upon the rarity of the outcome

times higher than the odds with B. In medical research, the odds ratio is used frequently for case-control studies and retrospective studies because it can be obtained easier and with less cost than studies which must estimate incidence rates in various risk groups. Relative risk is used in randomized controlled trials and cohort studies, but requires longitudinal follow-up and thus is more costly and difficult to obtain.[2]

## Relative Risk Reduction (RRR) and Absolute Risk Reduction (ARR) and Number Needed to Treat (NNT)

The RRR is simply 1 – RR times 100 and is the difference in event rates between two groups (e.g. a treatment and control group). Let's say you have done a trial where the event rate in the intervention group was 30/100 and the event rate in the control group was 40/100. The RRR is 25% (i.e. 10% absolute reduction divided by the events in the control group of 10/40). The absolute risk reduction ARR is just the difference in the incidence rates. So the ARR above is 0.40 minus 0.30 or 0.10, a difference of 10 cases. But what if in another trial we see 20% events in the control group of size N vs. 15 in the intervention group of size N? The RRR is 5/20 or 25% while the ARR is only 5%.

Absolute risk reduction (ARR) is another possible measure of association that is becoming more common in reporting clinical trial results of a drug intervention. Its inverse is called the number needed to treat or NNT. The ARR is computed by subtracting the proportion of events in the control group from the proportion of events in the intervention group. NNT is 1/ARR and is a relative measure of how many patients need to be treated to prevent one outcome event (in a specified time period). If there are 5/100 outcomes in the intervention group (say you are measuring strokes with BP lowering in the experimental group over 12 months of follow-up) and 30/100 in the control group, the ARR is 0.30 - 0.05 = 0.25, and the NNT is 4(1/0.25), that is for every four patients treated for a year (in the period of time of the study usually amortized per year) one stroke would be prevented (this, by the way, would be a highly effective intervention). The table below (16.4) summarizes the formulas for commonly used measures of therapeutic effect and Table 16.5 summarizes the various measures of association.

The main issue in terms of choosing any statistic, but specifically a measure of association, is to not use a measure of association that could potentially mislead the reader. An example of how this can happen is shown in Table 16.6.

**Table 16.4** Formulas for commonly used measures of therapeutic effect

| Measure of effect | Formula |
|---|---|
| Relative risk | (Event rate in intervention group) – (event rate in control group) |
| Relative risk reduction | 1 – relative risk or (Absolute risk reduction) – (event rate in control group) |
| Absolute risk reduction | (Event rate in intervention group) – (event rate in control group) |
| Number needed to treat | 1/(absolute risk reduction) |

**Table 16.5** Measures of Association

| Parameter | Treatment drug M | Control treatment |
|---|---|---|
| Recur/N = Rate | 5/100 = 0.05 | 30/100 = 0.30 |
| Relative risk | 0.05/0.30 = 0.17 | 0.30/0.05 = 6 |
| Odds ratio | 5 × 70/30 × 95 = 0.12 | 30 × 95/5 × 70 = 8.1 |
| Absolute risk reduction | 0.30 − 0.05 = 0.25 | |
| Number needed to treat | 1/(0.30 − 0.05) = 4 | |

**Table 16.6** Comparison of RR and AR

| | Annual mortality rate per 100,000 | |
|---|---|---|
| | Lung cancer | Coronary heart disease |
| Smokers | 140 | 669 |
| Non smokers | 10 | 413 |
| Relative risk | 14.0 | 1.6 |
| Attribute risk | $130/10^5$/year | $256/10^5$/year |

**Table 16.7** Number Needed to Treat (NNT) to avoid one death with converting enzyme inhibitor captopril after myocardial infarction

| | Control Number of deaths/Pts | Intervention Number of deaths/Pts | RR | NNT |
|---|---|---|---|---|
| SAVE trial (42 months) | 275/1,115 (24.7%) | 228/1,116 (20.4%) | 0.828 | 1/(0.247 – 0.204) 23.5 (24) |
| ISIS 4 (5 weeks) | 2,231/29,022 (7.69%) | 2,088/29,028 (7.19%) | 0.936 | 1/(0.0769 – 0.0719) 201.1 (202) |

In the above example the RR of 14 for annual lung cancer mortality rates is compared to the RR of 1.6 for the annual mortality rate of CAD. However, at a population level, the mortality rate for CAD per 100,000 is almost twice that of lung cancer. Thus, while the RR is enormously higher, the impact of smoking on CAD in terms of disease burden (ARR) is nearly double. A further example from the literature is shown in Table 16.7.

One can also compute the NNH (number needed to harm), an important concept to carefully present the down sides of treating along with the upsides. The NNH is computed by subtracting the proportion of adverse events in the control and intervention group per Table 16.8.

## Correlations and Regression

Other methods of finding associations are based on the concepts above, but use methods that afford the incorporation of other variables and include such tools as correlations and regression (e.g. logistic, linear, non-linear, least squares regression

**Table 16.8** Number needed to harm

• Similar to NNT
– 1/difference in side effects or adverse events:
• For example 1998 a study of Finasteride showed: NNT for various side effects:

| | Finast(%) | Control(%) | Number needed to harm |
|---|---|---|---|
| Impotence | 13.2 | 8.8 | 1/(0.132 − 0.088) = 22.7 or 23 |
| Decreased libido | 9.0 | 6.0 | 1/(0.09 − 0.6) = 33 |

line, multivariate or multivariable regression, etc.). We use the term regression to imply a co-relationship, and the term correlation to show relatedness of two or more variables. Linear regression investigates the linear association between two continuous variables. Linear regression gives the equation of the straight line that best describes an association in terms of two variables, and enables the prediction of one variable from the other. This can be expanded to handle multiple variables. In general, regression analysis examines the dependence of a random variable, called the dependent or response variable, on other random or deterministic variables, called independent variables or predictors. The mathematical model of their relationship is known as the regression equation. This is an extensive area statistics and in its fullest forms are beyond the scope of this chapter. Well known types of regression equations are linear regression for continuous responses, the logistic regression for discrete responses, and nonlinear regression. Besides dependent and independent variables, the regression equations usually contain one or more unknown regression parameters, which are to be estimated from the given data in order to maximize the quality of the model. Applications of regression include curve fitting, forecasting of time series, modeling of causal relationships and testing scientific hypotheses about relationships between variables. A graphical depiction of regression analysis is shown in Fig. 16.4. Correlation is the tendency for one variable to change as the other variable changes (it is measured by rho -ρ).

Correlation, also called correlation coefficient, indicates the strength and direction of a linear relationship between two random variables. In general statistical usage, correlation or co-relation refers to the departure of two variables from independence, that is, knowledge of one variable better informs an investigator of the expected results of the dependent variable than not considering this covariate. Correlation does not imply causation, but merely that additional information is provided about the dependent variable when the covariate, independent variable, is known. In this broad sense there are several coefficients, measuring the degree of correlation, adapted to the nature of data. The rate of change of one variable tied to the rate of change of another is known as a slope. The correlation coefficient and the slope of the regression line are functions of one another, and a significant correlation is the same as a significant regression. You may have heard of a concept called the r-squared. We talk of r-squared as the percent of the variation in one variable explained by the other. This means that if we compute the variation in the dependent variable by taking each observation, subtracting the overall mean and summing the squared deviations and dividing by the sample size to get our estimated variance.

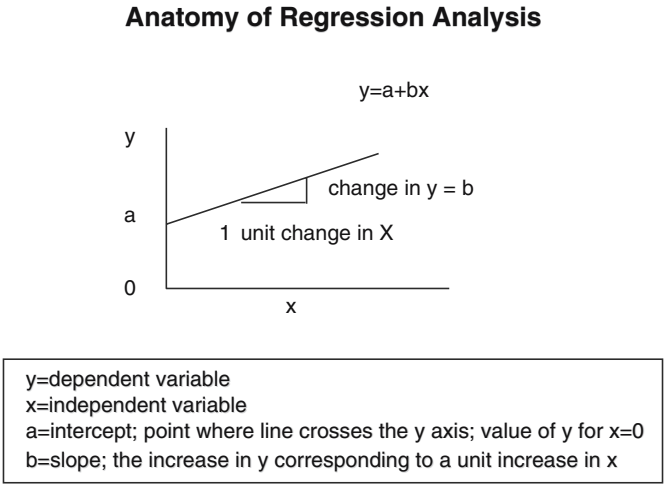



**Fig. 16.4** Anatomy of regression analysis

To assess the importance of the covariate, we compute a 'regression' model using the covariate and assess how well our model explains the outcome variable. We compute an expected value based on the regression model for each outcome. Then we assess how well our observed outcomes fit our expected. We compute the observed minus the expected, called the residual or unexplained portion and find the variance of these residuals. The ratio of the variance of residuals to the variation in the outcome variable overall is the proportion of unexplained variance and 1 minus this ratio is the R-squared or proportion of variance explained. A number of different coefficients are used for different situations. The best known is the Pearson product-moment correlation coefficient, which is easily obtained by standard formulae. Geometrically, if one thinks of a regression line, it is a function of the angle that the regression line makes with a horizontal line parallel to the x-axis, the closer the angle is to a 45 degree angle the better the correlation. Importantly, it should be realized that correlation can measure precision and/or reproducibility, but does not accuracy or validity.

## Causal Inference

An association (or a correlation) does not imply causation. In an earlier chapter, various clinical research study designs were discussed, and the differing 'levels of scientific evidence' that are associated with each were addressed. A comparison of study designs is complex, with the metric being that the study design providing the highest level of scientific evidence (usually experimental studies) is the one that yields the greatest likelihood of cause and effect relationship between the exposure and the outcome. The basic tenet of science is that it is almost impossible to prove an association or cause, but it is easier to disprove it. Causal effect focuses on outcomes among

exposed individuals, but what would have happened had they not been exposed? The outcome among exposed individuals is called the factual outcome. To draw inferences, exposed and non-exposed individuals are compared. Ideally, one would use the same population, expose them, observe the result, and then go back in time and repeat the same experiment among the same individuals but without the exposure in order to observe the counterfactual outcome. Randomized clinical trials attempt to approximate this ideal by using randomly assigned individuals to groups (to avoid any bias in assignment) and observe the outcomes. Because the true ideal experiment is impossible, replication of results with multiple studies is the norm. Another basic tenet is that even when the association is statistically significant, association does not denote causation. Causes are often distinguished into two types: Necessary and Sufficient.

## *Necessary Causes*

If x is a necessary cause of y; then the presence of y necessarily implies the presence of x. The presence of x, however, does not imply that y will occur. For example, poison ivy oils cause a purulent rash, but not everyone exposed will develop the rash, but all who develop the rash will be exposed to poison ivy oils.

## *Sufficient Causes*

If x is a sufficient cause of y, then the presence of x necessarily implies the presence of y. However, another cause z, may alternatively cause y. Thus the presence of y does not imply the presence of x.

The majority of these tenets and related ones (Koch's postulates, Bradford Hills tenets of causation) were developed with infectious diseases in mind. There are more tenuous conclusions that emanate from chronic diseases Giovannoni [1].

Consider the finding of an association between coffee drinking and myocardial infarction (MI) (Table 16.9). Coffee drinking might be a 'cause' of the MI, as the finding of that association from a study might imply. However, some persons who have had an MI may begin to drink more coffee, in which case (instead of a cause-effect relationship) the association would be an 'effect-cause' relationship (sometimes referred to as reverse causation).

**Table 16.9** Five explanations of association

| Association | Basis | Type | Explanation |
|---|---|---|---|
| 1. C → MI | Cause-effect | Real | Cause-effect |
| 2. MI → C | Cart before horse | Real | Effect-cause |
| 3. C ← × → MI | Confounding | Real | Effect-cause |
| 4. C ≠ MI | Random error | Spurious | Chance |
| 5. C ≠ MI | Systematic error | Spurious | Bias |

The association between coffee drinking and MI might be mediated by some confounder (e.g., persons who drink more coffee may smoke more cigarettes, and it is the smoking that precipitates the MI) (Table 16.3). Finally, observed associations may be spurious as a result of chance (random error) or due to some systematic error (bias) in the study design. To repeat, in the first conceptual association in Table 16.9, coffee drinking leads to MI, so it could be casual. The second association represents a scenario in which MI leads to coffee drinking (effect-cause or reverse causation). An association exists, but coffee drinking is not causal of MI. In the third association, the variable x results in coffee drinking and MI, so it confounds the association between coffee drinking and MI. In the fourth and fifth associations, the results are spurious because of chance or some bias in the way in which the trial was conducted or the subjects were selected. Thus, establishing cause and effect, is notoriously difficult and within chronic diseases has become even more of a challenge. In terms of an infectious disease – think about a specific flu – many flu-like symptoms occur without a specific viral agent, but for the specific flu, we need the viral agent present to produce the flu. What about Guillian-Barre Syndrome – it is caused by the Epstein Barr Virus (EBV), but the viral infection and symptoms have often occurred previously. It is only thru the antibodies to the EBV that this cause was identified. Further, consider the observation that smokers have a dramatically increased lung cancer rate. This does not establish that smoking must be a cause of that increased cancer rate: maybe there exists a certain genetic defect which both causes cancer and a yearning for nicotine; or even perhaps nicotine craving is a symptom of very early-stage lung cancer which is not otherwise detectable. In statistics, it is generally accepted that observational studies (like counting cancer cases among smokers and among non-smokers and then comparing the two) can give hints, but can never establish cause and effect. The gold standard for causation is the randomized experiment: take a large number of people, randomly divide them into two groups, force one group to smoke and prohibit the other group from smoking, then determine whether one group develops a significantly higher lung cancer rate. Random assignment plays a crucial role in the inference to causation because, in the long run, it renders the two groups equivalent in terms of all other possible effects on the outcome (cancer) so that any changes in the outcome will reflect only the manipulation (smoking). Obviously, for ethical reasons this experiment cannot be performed, but the method is widely applicable for less damaging experiments. And our search for causation must try to inform us with data as similar to possible as the RCT.

Because causation cannot be proven, how does one approach the concept of ‘proof’? The Bradford Hill criteria for judging causality remain the guiding principles as follows. The replication of studies in which the magnitude of effect is large, biologic plausibility for the cause-effect relationship is provided, temporality and a dose response exist, similar suspected causality is associated with similar exposure outcomes, and systematic bias is avoided, go a long way in suggesting that an association is truly causal.

## *Deductive vs Inductive Reasoning*

Drawing inferences about associations can be approached with deductive and inductive reasoning. An overly simplistic approach is to consider deductive reasoning as truths of logic and mathematics. Deductive reasoning is the kind of reasoning in which the conclusion is necessitated by, or reached from, previously known facts (the premises). If the premises are true, the conclusion must be true. This is distinguished from inductive reasoning, where the premises may predict a high probability of the conclusion, but do not ensure that the conclusion is true. That is, induction or inductive reasoning, sometimes called inductive logic, is the process of reasoning in which the premises of an argument are believed to support the conclusion but do not ensure it.

For example, beginning with the premises 'All ice is cold' and 'This is ice', you may conclude that 'This is cold'. An example where the premise being correct but the reasoning incorrect is 'this French person is rude so all French must be rude' (although some still argue that this is true). That is, deductive reasoning is dependent on its premises-a false premise can possibly lead to a false result, and inconclusive premises will also yield an inconclusive conclusion. We induce truths based on the interpretation of empirical evidence; but, we learn that these 'truths' are simply our best interpretation of the data at the moment and that we may need to change as new evidence is presented.

When using empirical observations to make inductive inferences, we have a greater ability to falsify a principle than to affirm it. This was pointed out by Karl Popper[3] in the late 1950s with his now classic example: if we observe swan after swan, and each is white, we may infer that all swans are white. We may observe 10,000 white swans and feel more confident about our inference. However, it takes but a single observation of a non-white swan to disprove the assertion. It is this Popperian view from which statistical inferences using the null hypothesis is born. That is we set our hypothesis that our theory is not correct, and then set out to disprove it. The p value is the probability (thus 'p'), that is the mathematical probability, that we would find a difference if the null hypothesis was true. Thus, the lower the probability of the finding, the more certain we can be in stating that we have falsified the null hypothesis.

Errors in making inferences about associations can also occur due to chance, bias, and confounding (see Chapter 17). Bias refers to anything that results in error i.e. compromises validity in a study. It is not (in a scientific sense) an intentional behavior, but rather it is an unintended consequence of a flaw in study design or conduct that affects an association. The two most common examples are selection bias (the inappropriate selection of study participants) and information bias (a flaw in measuring either the exposure group or disease group). These biases are the 'achilles heel' of observational studies which are essentially corrected for in randomized trials. However, randomized trials may restrict the populations to a degree that also leads to selection biases. When an association exists, it must be determined whether the exposure caused the outcome, or the association is caused by

some other factor (i.e. is confounded by another factor). A confounding factor is both a risk factor for the disease and a factor associated with the exposure. Some classify confounding as a form of bias. However, confounding is a reality that actually influences the association, although confounding can introduce bias (i.e. error) into the findings of a study. Confused with confounding is effect modification. Confounding and effect modification are very different in both the information each provides as well as what is done with that information. For confounding to exist, a factor must be unevenly distributed in the study groups, and as a result has influenced the observed association. Confounding is a nuisance effect, and the researchers main goal is to control for confounding and eliminate its effect (by stratification or multivariate analysis). In a statistical sense confounding is inextricably tied to the variable of interest, but in epidemiology we consider confounding a covariate. Effect modification is a characteristic that exists irrespective of study design or study patients. It is to be reported, not controlled.

Stratification is used to control for confounding, and to describe effect modification. If, for example, an association observed is stratified for age and the effect is uniform across age groups, this suggests confounding by age. In contrast, if the observed association is not uniform effect modification is present. For example, amongst premature infants, stratified by birth weight; 500–749 g, 750–999 g and 1,000–1,250 g, the incidence of intracranial hemorrhage (ICH) is vastly different across these strata, thus birth weight is an effect modifier of ICH.

## References

1. Giovannoni G, et al. Infectious causes of multiple sclerosis. *Lancet Neurol*. 2006 Oct; 5(10): 887-94.
2. Zhang J, Yu KF. What's the relative risk? A method of correcting the odds ratio in cohort studies of common outcomes. *JAMA*. Nov 18, 1998; 280(19):1690–1691.
3. Relative risk. *Wikipedia*.
4. http://en.wikiquote.org/wiki/Karl_Popper.

# Chapter 17
# Bias, Confounding, and Effect Modification

**Stephen P. Glasser**

*You're like the Tower of Pisa-always leaning in one direction*[1]

**Abstract** Bias, jaconfounding, and random variation/chance are the reasons for a non-causal association between an exposure and outcome. This chapter will define and discuss these concepts so that they may be appropriately considered whenever one is interpreting the data from a study.

## Introduction

Bias, confounding, and random variation/chance are alternate explanations for an observed association between an exposure and outcome. They represent a major threat to the internal validity of a study, and should always be considered when interpreting data. Whereas statistical bias is usually an unintended mistake made by the researcher; confounding is not a mistake; rather, it is an additional variable that can impact the outcome (negatively or positively; all or in part) separately from the exposure. Sometimes, confounding is considered to be a third major class of bias.[2]

As will be further discussed, when a confounding factor is *known* or suspected, it can be controlled for in the design phase (randomisation, restriction and matching) or in the analysis phase (stratification, multivariable analysis and matching). The best that can be done about *unknown* confounders is to use a randomised design (see Chapter 3). Bias and confounding are not affected by sample size, but chance effect (random variation) diminishes as the sample size gets larger. A small p-value and a narrow odds ratio or relative risk are reassuring signs against chance effect but the same cannot be said for bias and confounding.[3]

S.P. Glasser (ed.), *Essentials of Clinical Research*,

## Bias

Bias is a systematic error that results in an incorrect (invalid) estimate of a measure of association. That is, the term bias 'describes the systematic tendency of any factors associated with the design, conduct, analysis, and interpretation of the results of clinical research to make an estimate of a treatment effect deviate from its true value'.[3] Bias can either create or mask an association; that is, bias can give the appearance of an association when there really is none, or can mask an association when there really is one. Bias can occur with all study designs, be it experimental, cohort, or case-control; and, can occur in either the design phase of a study, or during the conduct of a study. For example, bias may occur from an error in the measurement of a variable; confounding involves an incorrect interpretation of an association even when there has been accurate measurement. Also, whereas adjustments can be made in the analysis phase of a study for confounding variables, bias can not be controlled, at best; one can only suspect that it has occurred. The most important design techniques for avoiding bias are blinding and randomization.

An example of systematic bias would be a thermometer that always reads three degrees colder than the actual temperature because of an incorrect initial calibration or labeling, whereas one that gave random values within five degrees either side of the actual temperature would be considered a random error.[4] If one discovers that the thermometer always reads three degrees below the correct value one correct for the bias by simply making a systematic correction by adding three degrees to all readings. In other cases, while a systematic bias is suspected or even detected, no simple correction may be possible because it is impossible to quantify the error. The existence and causes of systematic bias may be difficult to detect without an independent source of information; the phenomenon of scattered readings resulting from random error calls more attention to itself from repeated estimates of the same quantity than the mutually consistent incorrect results of a biased system.

There are two major types of bias; selection and observation bias.[5]

### *Selection Bias*

Selection bias is the result of the approach used for subject selection. That is, when the sample in the study ends up being different from the target population, selection bias is a cause. Selection bias is more likely to be present in case-control or retrospective cohort study designs, because the exposure and the outcome have already occurred at time of subject selection. For a case-control study, selection bias occurs when controls or cases are more (or less) likely to be included in study if they have been exposed – that is, inclusion in the study is not independent of the exposure. The result of this is that the relationship between exposure and disease observed among study participants is different from relationship between exposure

and disease in individuals who would have been eligible but were not included, thus the odds ratio from a study that suffers from selection bias will incorrectly represent the relationship between exposure and disease in the overall study population.[6]

A biased sample is a statistical sample of a population in which some members of the population are less likely to be included than others. If the bias makes estimation of population parameters impossible, the sample is a non-probability sample. An extreme form of biased sampling occurs when certain members of the population are totally excluded from the sample (that is, they have zero probability of being selected). For example, a survey of high school students to measure teenage use of illegal drugs will be a biased sample because it does not include home schooled students or dropouts. A sample is also biased if certain members are underrepresented or overrepresented relative to others in the population. For example, a "man on the street" interview which selects people who walk by a certain location is going to have an over-representation of healthy individuals who are more likely to be out of the home than individuals with a chronic illness. A biased sample causes problems because any statistic computed from that sample has the potential to be consistently erroneous.[7] Bias can lead to an over- or under-representation of the corresponding parameter in the population. Almost every sample in practice is biased because it is practically impossible to ensure a perfectly random sample. If the degree of under-representation is small, the sample can be treated as a reasonable approximation to a random sample. Also, if the group that is underrepresented does not differ markedly from the other groups in the quantity being measured, then a random sample can still be a reasonable approximation.

The word bias in common usage has a strong negative connotation, and implies a deliberate intent to mislead. In statistical usage, bias represents a mathematical property. While some individuals might deliberately use a biased sample to produce misleading results, more often, a biased sample is just a reflection of the difficulty in obtaining a truly representative sample.[7]

Let's take as an example the data shown in Fig. 17.1, which addresses the question of whether otitis media differs in bottle feeding, as opposed to breast feeding. 100 infants with ear infection are identified among members of one HMO, and the controls are 100 infants in that same HMO without otitis. The potential bias is whether being included in the study as a control is not independent of the exposure, that is, they were not representative of the whole study population that produced the cases. In other words, one could ask the reason(s) that infants were being seen in an HMO in the first place and how many might have had undiagnosed otitis.

So, what are the solutions for selection bias? Little or nothing can be done to fix selection bias once it has occurred. Rather one needs to avoid it during the design and conduct the study by, for example, using the same criteria for selecting cases and controls, obtaining all relevant subject records, obtaining high participation rates, and taking into account diagnostic and referral patterns of disease. But, almost always (perhaps always) one can not totally remove selection bias form any study.

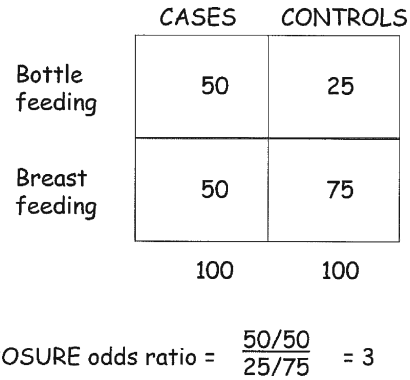



**Fig. 17.1** Odds ratio of developing otitis media dependent upon bottle vs. breast feeding

## *Observation Bias*

While selection bias occurs as subjects enter the study, observation bias occurs after the subjects have entered the study. Observation bias is the result of incorrectly classifying the study participant's exposure or outcome status. There are several types of observation bias: recall bias, interviewer bias, loss to follow up, and differential and non-differential misclassification.

Recall bias occurs because participants with and without the outcome of interest do not report their exposure accurately (because they do not remember it accurately) and more importantly report the exposure differently (this can result in an over- or under-estimate of the measure of association). It is not that unlikely that subject's with an outcome might remember the exposure more accurately than subjects without an outcome, particularly if the outcome is a disease. Solutions for recall bias include using controls, who are themselves sick; and/or, using standardized questionnaires that obtain complete information and that mask subjects to the study hypothesis.[8]

Whenever exposure information is sought, information is recorded and interpreted. If there is a systematic difference in the way the information is solicited, recorded, or interpreted, interviewer bias can occur. One solution to reduce interviewer bias is to mask interviewers, so that they are unaware of the study hypothesis and disease or exposure status of subjects, and to use standardized questionnaires or standardized methods of outcome (or exposure) ascertainment.[9]

Loss to follow up is a concern in cohort and experimental studies if people who are lost to follow up differ from those that remain in the study. Bias results if subjects lost, differ from those that remain, with respect to both the outcome and exposure. The main solution for lost to follow up is to minimize its occurrence. Excessive numbers of subjects lost to follow up can seriously damage the validity of the study. (See also discussion of lost to follow up in Chapter 3.)

Misclassification bias occurs when a subject's exposure or disease status is erroneously classified. Two types of misclassification are non-differential (random) and differential (non random). Non-differential misclassification results in inaccuracies with respect to disease classification that is independent of the exposure; or, with inaccuracies with respect to the exposure that are independent of disease. Non-differential misclassification makes the exposure and non exposure groups more similar. The probability of misclassification may be the same in all study groups (non-differential misclassification) or may vary between groups (differential misclassification).

### *Measurement Bias*

Let's consider that a true value does in fact exist. Both random and biological variation modifies that true value by the time the measurement is made. Performance of the instrument and observer bias, and recording and computation of the results further modifies the 'true value' and this now becomes the value used in the study. Reliability has to do with the ability of an instrument to measure consistently, repeatedly, and with precision and reproducibility. But, the fact is, that every instrument has some inherent imprecision and/or unreliability. This latter fact negatively impacts one of the main objectives of clinical research, to isolate subject variability between subjects, from measurement variability. Measurement error is, therefore, intrinsic to research.

In summary, in order to reduce bias, ask yourself these questions:' given the conditions of the study, could bias have occurred? Is bias actually present? Are consequences of the bias large enough to distort the measure of association in an important way? Which direction is the distortion, that is, is it towards the null or away from the null?[9]

## Confounding

A confounding variable (confounding factor or confounder) is a variable that correlates (positively or negatively) with both the exposure and outcome. One, therefore, needs to control for these factors in order to avoid what is known as a type 1 error, which is a 'false positive' conclusion that the exposure is in a causal relationship with the outcome. Such a relation between two observed variables is termed a spurious relationship. Thus, confounding is a major threat to the validity of inferences made about cause and effect, i.e. internal validity, as the observed effects should be attributed all or in part to the confounder rather than the outcome. For example, assume that a child's weight and a country's gross domestic product (GDP) rise with time. A person carrying out an experiment could measure weight and GDP, and conclude that a higher GDP causes children to gain weight. However, the confounding variable, time, was not accounted for, and is the real cause of both rises.[10] By definition, a confounding variable is associated with both the probable

cause and the outcome, and the confounder should not lie in the causal pathway between the cause and the outcome. Though criteria for causality in statistical studies have been researched intensely, Pearl has shown that confounding variables cannot be defined in terms of statistical notions alone; some causal assumptions are necessary.[11] In a 1965 paper, Austin Bradford Hill proposed a set of causal criteria.[12] Many working epidemiologists take these as a good place to start when considering confounding and causation.

There are various ways to modify a study design to actively exclude or control confounding variables[13]:

- Case-control studies assign confounders to both groups, cases and controls, equally. For example if somebody wanted to study the cause of myocardial infarct and thinks that the age is a probable confounding variable, each 67 years old infarct patient will be matched with a healthy 67 year old "control" person. In case-control studies, matched variables most often are the age and sex.
- Cohort studies: A degree of matching is also possible and it is often done by only admitting certain age groups or a certain sex into the study population, and thus all cohorts are comparable in regard to the possible confounding variable. For example, if age and sex are thought to be confounders, only 40 to 50 years old males would be involved in a cohort study that would assess the myocardial infarct risk in cohorts that either are physically active or inactive.
- Stratification: As in the example above, physical activity is thought to be a behavior that protects from myocardial infarct; and age is assumed to be a possible confounder. The data sampled is then stratified by age group – this means, the association between activity and infarct would be analyzed per each age group. If the different age groups (or age strata) yield much different risk ratios, age must be viewed as a confounding variable. There are statistical tools like Mantel-Haenszel methods that deal with stratified data.

All these methods have their drawbacks. This can be clearly seen in the following example: a 45 year old Afro-American from Alaska, avid football player and vegetarian, working in education, suffers from a disease and is enrolled into a case-control study. Proper matching would call for a person with the same characteristics, with the sole difference of being healthy – but finding such individuals would be an enormous task. Additionally, there is always the risk of over- and under-matching of the study population. In cohort studies, too many people can be excluded; and in stratification, single strata can get too thin and thus contain only a small, non-significant number of samples.[4]

An additional major problem is that confounding variables are not always known or measurable. This leads to 'residual confounding' – epidemiological jargon for incompletely controlled confounding. Hence, randomization is often the best solution since, if performed successfully on sufficiently large numbers, all confounding variables (known and unknown) will be equally distributed across all study groups.

In summary, confounding is an alternative explanation for an observed association between the exposure and outcome. Confounding is basically a mixing of effects such that the association between exposure and outcome is distorted

because it is mixed with the effect of another factor that is associated with the disease. The result of confounding is to distort the true association toward the null (negative confounding) or away from the null (positive confounding). It should be re-emphasized, that a variable cannot be a confounder if it is in the causal chain or pathway. For example, moderate alcohol consumption increases serum HDL-C levels which, in turn, decrease the risk of heart disease. Thus, HDL-C levels are a step in the causal chain, not a confounder that needs to be controlled.[9] This latter example is rather something interesting that helps us understand the disease mechanism. In contrast, smoking is confounder of effect of occupational exposures (to dyes) on bladder cancer and does need to be controlled for because, confounding factors are nuisance variables. They get in the way of the relation you want to study; as a result, one wants to remove their effect. Recall that here are three ways of eliminating a confounder. The first is with the use of a case-control design, in which the confounder is matched between the cases and the controls. The second way of eliminating a confounder is mathematically, by the use of multivariate analysis. And, the third and best way for reducing the effect of confounding is to use a randomized design; but, remember "likely to control" means just that. It's not a guarantee.

## Confounding vs. Effect Modification

As discussed above, confounding is another explanation for apparent associations that are not due to the exposure. Also recall, that confounding is defined as an extraneous variable in a statistical or research model that affects the outcome measure, but has either not been considered or has not been controlled for during the study. The confounding variable can then lead to a false conclusion that the outcome has a causal relationship with the exposure. Consider the example where coffee drinking is found to be associated with myocardial infarction (MI). If there is really no effect of coffee intake on MI but more coffee drinkers smoke cigarettes than non coffee drinkers, then cigarette smoking is a confounder in the apparent association of coffee drinking and MI. If one corrects for smoking, the true absence of the association of coffee drinking and MI will become apparent.

Effect modification is sometimes confused with confounding. In the example above, let us say that both coffee drinking and smoking impact on the outcome (MI). If one corrects for smoking, and there is still some impact of coffee drinking on MI, some association is imparted by cigarette smoking. In the hypothetical example above, let's say we find a RR of 5 for the association of coffee drinking and MI. When cigarette smokers are eliminated from the analysis and smoking is a confounder, the RR will be 1. In the case of effect modification where both coffee drinking and smoking equally contribute to the outcome (i.e. both smoking and coffee drinking have an equal impact on the association) the RR for each will be 2.5.

## References

1. Cited in “Quote Me” How to add wit and wisdom to your conversation. Compiled by J Edward Breslin. Hounslow Press, Ontario, Canada, 1990, p 44
2. http://dorakmt.tripod.com/epi/bc/html
3. http://www.dorak.info/epi/bc.html
4. http://en.wikipedia.org/wiki/Systematic_bias
5. Davey Smith G, Ebrahim S. Data dredging, bias, or confounding. BMJ 2002; 325: 1437–1438
6. http://dorakmt.tripod.com/epi/bc/html)
7. http://en.wikipedia.org/wiki/Biased_sample
8. Sackett DL. Bias in analytic research. J Chronic Dis 1979; 32:51–63
9. http://publichealth.jbpub.com/aschengrau/ppts/confounding.ppt10.
10. http://en.wikipedia.org/wiki/Confounding_variable
11. Pearl, Judea (2000). Causality: Models, Reasoning, and Inference. Cambridge University Press, Cambridge. ISBN 0-521-77362-8.
12. Bradford Hill, Austin. “The environment or disease: association Or causation?”. Proc Roy Soc Med May 1965; 58:295–300. PMID 14283879
13. Mayrent, Sherry L. Epidemiology in Medicine. Williams & Wilkins, Baltimore, MD; 1987. ISBN 0-316-35636-0.

# Chapter 18
# It's All About Uncertainty

**Stephen P. Glasser and George Howard**

*Not everything that counts can be counted; and, not everything that can be counted counts.*

Albert Einstein

**Abstract** This chapter is aimed at providing the foundation for common sense issues that underlie why and what statistics is, so it is not a math chapter, relax! We will start with the concepts of "the universe" and a "sample", discuss the conceptual issues of estimation and hypothesis testing and put into context the question of how certain are we that a research result in the sample studied reflects what is true in the universe.

## Introduction

It is surprising that as a society we accept poor math skills. Even if one is not an active researcher, one has to understand statistics to read the literature. Fortunately, most of statistics are common sense. This chapter is aimed at providing the foundation for common sense issues that underlie why and what statistics is, so it is not a math chapter, relax!

Let us start with the concepts of "the universe" and of a "sample". The "universe" is that group of people (or things) that we really want to know about … it is what we are really trying to describe. For example, for a study we might be interested in the blood pressure for British white males – then the "universe" is every white man in the Great Britain. The trouble is that the "universe" is simply too big for research purposes, we cannot begin to measure everybody in Great Britain, so we select a representative part of the universe – which is our study sample. Since the sample is much smaller than the universe we have the ability to measure things on everybody in the sample and analyze relationships between factors in the sample. If the sample is really representative of the universe, and we understand what is going on in that sample, we gain an *inferential* understanding of what is happening in the universe. The critical concept is that we perform our analysis on the sample (which is not what we really want to describe) and infer that we understand what is going on in the universe (which is our real goal) (as an aside, when the entire universe is measured it is called performing a *census* and we all know even that has its problems). There are, however, advantages of measuring everyone if we could For example, if we could measure everyone, we will get the correct answer – there is almost no uncertainty when everyone is measured; and, one will not need a statistician – because the main

S.P. Glasser (ed.), *Essentials of Clinical Research*,

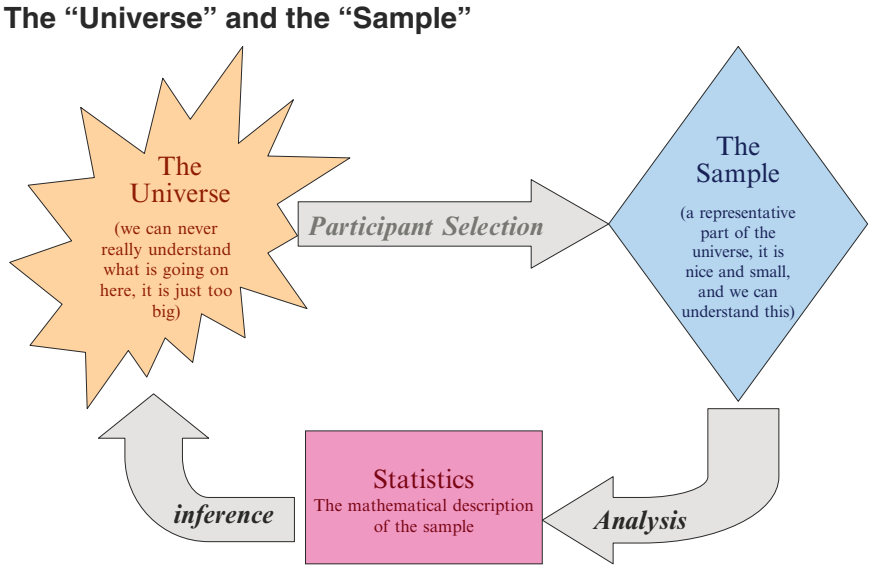



**Fig. 18.1** The "Universe" and the "Sample"

job of a statistician is to deal with the uncertainty involved in making inferences from a sample. However, since measuring everyone is impractical (impossible?), and very expensive, for practical reasons one is forced to use an inferential approach, which if done correctly, one can *almost* be certain to get *nearly* the correct answer. The entire field of statistics deals with this uncertainty, specifically to help define or quantify "almost" and "nearly" when making an inference (Fig. 18.1). The characteristic that defines any statistical approach is how it deals with uncertainty. The traditional approach to dealing with uncertainty is the frequentist approach, which deals with fixed sample sizes based upon prior data; but the information present from prior studies is not incorporated into the study being now implemented. That is, with the frequentist approach "the difference between treatment groups is assumed to be an unknown and fixed parameter"; and, estimates the minimal sample size in advance of the trial and analyzing the results of the trial using p values. A Bayesian approach uses previous data to develop a prior distribution of potential differences between treatment groups and updates this data with that collected during the trial being performed to develop a posterior distribution (this is akin to the discussion in Chapter 14 that addresses pre and post test probability). We will discuss the Bayesian approach later in this chapter.

There are two kinds of inferential activities statisticians perform – estimation and hypothesis testing, each described below.

## Conceptual Issues in Estimation

Estimation is simply the process of producing a very educated guess for the value of some parameter ("truth") in the universe. In statistics, as in guessing in other fields, the key is to understand how close the estimate is to the true value. Conceptually, *parameters* (such as an average BP of men in the US) exist in the *universe* and do not change, but we cannot know them without measuring everyone. The natural question would then be "*how good is our guess*;" and, for this we need to have some measure of the reliability of our estimate or guess.

If we select two people out of the universe, one would not expect them to have the same exact measurement (i.e. for example, we would not expect them to have the same blood pressure). People in a population have a dispersion of outcomes that is characterized by the standard deviation. We might recall from standardized testing for college and graduate programs that about 95% of the people are within about two standard deviations of the average value. That is, getting people who are more than two standard deviations away from the mean will not happen very often (in fact, less than 5% of the time).

Returning to the example mentioned above, suppose we are interested in estimating (guessing) the mean blood pressure of white men in Great Britain. How much variation (uncertainty) can we reasonably expect between two estimates of the mean blood pressure? To answer this, consider that the correct answer exists in the universe, but the estimate from a sample will likely be somewhat different from that true value. In addition, a different sample would likely give a result that is both different from the "true" value and different from the first estimate. If one repeats the experiment in a large number of samples, the different estimates that would be produced from the repeated experiments would have a standard deviation. This standard deviation of estimates from repeated estimates has a special name – the *standard error of the estimate*. The standard error is nothing more than the standard deviation of the estimate, if the same experiment was repeated a large number of times. That is, if one repeats an experiment 100 times (i.e. obtain100 different samples of white men, and each time calculate a mean blood pressure), just as we would not expect individual people to have the same blood pressure, we would not expect these samples to have the same mean blood pressure. The standard deviation of means is called the standard error of the mean. The real trick of the field of statistics is to provide an estimate of the standard error of a parameter when the experiment is only performed a single time. That is, if a single sample is drawn from the universe, on the basis of that single sample is it possible to say how much you would expect the mean of future samples to differ from that obtained in this first sample (if one thinks about it … this is quite a trick).

As mentioned above, we are all familiar with the fact that 95% of people are within two standard deviations of the mean (again, think about the standardized tests we have all taken). It turns out that 95% of the estimates are also within two standard deviations (except we call it two standard errors) of the true mean. This observation is the basis for "confidence intervals" and this can be used to characterize the uncertainty of the estimation. The calculation of a confidence interval is

nothing more than a reflection of the same concept that 95% of the people (estimates) are within about two standard deviations (standard errors) of the mean. The use of confidence intervals permits a more refined assessment of the uncertainty of the guess, and is a range of values calculated from the results of a study, within which the true value lies; the width of the interval reflecting random error. The width of the confidence limit differs slightly from the two standard errors (due to adjustment for the uncertainty from sampling), and the width is also a function of sample size (a larger sample size reduces the uncertainty). Also, the most common interpretation of a confidence interval is that "I am 95% sure that the real parameter is within this range" is technically incorrect, albeit not that incorrect. The correct interpretation is much less intuitive (and therefore is not as frequently used) – that if an experiment were repeated a large number of times, and 95% confidence limits were calculated each time using similar approaches, then 95% of the time these confidence limits would include the true parameter. We are all accustomed to hearing about confidence limits, since confidence intervals are what pollsters mean when they talk about the "margin of error" of their poll.

To review, estimation is an educated guess of a parameter, and every estimate (not only estimated means, but also estimated proportions, slopes, and measures of risk) has a standard error. The 95% confidence limits depict the range that we can "reasonably" expect the true parameter to be within (approximately ±2 SE). For example, if the mean SBP is estimated to be 117 and the standard error is 1.4, then we are "pretty sure" the true mean SBP is between 114.2 and 119.8 (the slightly incorrect interpretation of the 95% confidence limit is "I am 95% sure that the real parameter is between these numbers").

Studies frequently focus on the association between an "exposure" (treatment) and an "outcome". In that case, parameter(s) that describe the strength of the association between the exposure and the outcome are of particular interest. Some examples are:

- The difference in cancer recurrence at a given time, between those receiving a new versus a standard treatment
- The reduction in average SBP associated with increased dosage of an antihypertensive drug
- The differences in the likelihood of being a full professor before age 40 in those who read this book versus those who do not

Let's say we have a sample of 51 University of Alabama at Birmingham students some of whom have read an early draft of this book years ago. We followed each of these students to establish their academic success, as measured by whether they made the rank of full professor by age 40. The, resulting data is portrayed in Table 18.1. From a review of Table 18.1 what types of estimates of the measure of association can we make from this sample? We can:

1. Calculate the *absolute difference* in those achieving the goal:
   (a) Calculating the proportion that achieved the goal among those reading the book (20/31 = 0.65 or 65%).

(b) Calculating the proportion that achieved the goal among those not reading the book (8/20 = 0.40 or 40%).
(c) By calculating the difference in these two proportions (0.65 – 0.40 = 0.25), we can demonstrate a 25% increase in the likelihood of academic success by this measure.

Or

2. We can calculate the *relative risk (RR)* of achieving the goal:

(a) By, calculating the proportion that achieved the goal among those reading the book (20/31 = 0.65 or 65%)
(b) By, calculating the proportion that achieved the goal among those not reading the book (8/20 = 0.40 or 40%)
(c) And then calculating the ratio of these two proportions (RR is 0.65/0.40 = 1.61) – or there is a 61% increase in the likelihood of making full professor among those reading the book

Or

3. We can calculate the *odds ratio (OR)* of achieving this goal:

(a) By calculating the odds (the "odds" is the chance of something happening divided by the chance of it not happening) of achieving the goal among those reading the book (20/11 = 1.81)
(b) By calculating the odds of achieving the goal among those not reading the book (8/12 = 0.67)
(c) And then, calculating the ratio of these two odds (OR is 1.81/0.67 = 2.73) – or there is a 2.73 times greater odds of making full professor among those reading the book

The point of this example is to demonstrate that there are different estimates that can reasonably be produced from the very same data. Each of these approaches is correct, but they give extremely different impressions of what is occurring in the study (that is, is there a 25% increase, a 65% increase or a 173% increase?). In estimation, therefore, great care should be taken to make sure that there is a deep understanding of what is being estimated.

To review the major points about estimation:

- Estimates from samples are only educated guesses of the truth (of the parameter).
- Every estimate has a standard error, which is a measure of the variation in the estimates. When standard errors are not provided, care should be taken in the inter-

**Table 18.1** A 2 × 2 table from which varying estimates can be derived

| | | Full Professor by 40 | | |
|---|---|---|---|---|
| | | Yes | No | Total |
| Attend course | Yes | 20 | 11 | 31 |
| | No | 8 | 12 | 20 |
| | Total | 28 | 23 | |

pretation of the estimates – they are guesses without an assessment of the quality of the guess (by the way, note that standard errors were not provided for the guesses made from Table 18.1 of the difference, the relative risk, or the odds ratio of the chance of making full professor).

- If you were to repeat a study, one should not expect to get the same answer (just like if one sampled people from a population, one should not expect them to have the same blood pressure amongst individuals in that sample).
- When you have two estimates, you can conclude:
  - It is almost certain that neither is correct.
  - However, in a well-designed experiment
    - The guesses should be "close" to "correct".
    - Statistics can help us understand how far our guesses are likely to be from the truth, and how far they would be from other guesses (were they made).

## Conceptual Issues in Hypothesis Testing

The other activity performed by statisticians is hypothesis testing, which is simply making a yes/no decision regarding some parameter in the universe. In statistics, as in other decision making areas, the key to decision making is to understand what kind of errors can be made; and, what the chances are of making an incorrect decision. The basis of hypothesis testing is to assume that whatever you are trying to prove is not true – i.e. that there is no relationship (or technically, that the null hypothesis $H_o$ is supported). To test the hypothesis of no difference, one collects data (on a sample), and calculates some "test statistic" that is a function of that data. In general, if the null hypothesis is true, then the test statistic will tend to be "small;" however, if the null hypothesis is incorrect the test statistic is likely to be "big." One would then calculate the chance that a test statistic as big (or bigger) as we observed would occur under the assumption of no relationship (this is termed the *p-value*!). That is, if the observed data is unlikely under the null, then we either have a strange sample, or the null hypothesis of no difference is wrong and should be rejected. To return to Table 18.1, let's ask the question "how can one calculate the chance of getting data this different for those who did versus those who did not read a draft of this book, under the assumption that reading the book has no impact?" The test statistic is then calculated to assess whether there is evidence to reject the hypothesis that the book is of no value. Specifically, the test statistic used is the Chi-square ($\chi^2$), the details of which are unimportant in this conceptual discussion – but the test statistic value for this particular table is 2.95. Now the question becomes is 2.95 "large" (providing evidence that the null hypothesis of no difference is not likely) or "small" (failing to provide such evidence). It can be shown that in cases like the one considered here, that if there is really no association between reading the book and the outcome, that only 5% of the time is the value of the

test statistic larger than 3.84 (this, therefore, becomes the definition of "large"). Since 2.95 is less than 3.84, this is not a "large" test statistic; and, therefore, there is not evidence to support that the null hypothesis is wrong (i.e. that reading the book has no impact is wrong - however, one cannot use these hypothetical data to prove that you are currently otherwise spending your time wisely). We acknowledge and regret that this double-negative statement must be made, i.e. "there is not evidence that the null hypothesis is wrong". This is because, one does not "accept" the null hypothesis of no effect, one just does not reject it. This is a small, but critically important concept in hypothesis testing – that a "negative" test (as was true in the above example) does not prove the null hypothesis, it only fails to support the alternative. On the other hand, if the test statistic had been bigger than 3.84, then we would have rejected the null hypothesis of no difference and accepted the alternative hypothesis of an effect (i.e. that reading this book does improve ones chances of early academic advancement – obviously the correct answer).

## *P Value*

The "*p-value*" is the chance that the test statistic from the sample could have happened under the null hypothesis. What constitutes a situation where it is "unlikely" for the data to have come from the null, that is, how much evidence are we going to require before one "rejects" the null? The standard is that if the data has less than a 5% chance ($p < 0.05$) of happening by chance alone, then the observation is considered "unlikely". One should realize that this p value (0.05) is an arbitrary number, and many argue that too much weight is given to the p-value. None-the-less, the p-value being less than or greater than 0.05 is inculcated in most scientific work. However, consider the example of different investigators performing an identical experiment and one gets $p = 0.053$, whereas the other gets $p = 0.049$. Should one really come to different conclusions? In one case there is a 5.3% chance of getting data as observed under the null hypothesis, and in the other there is a 4.9% chance. If one accepts the 0.05 threshold as "gospel," then these two very similar results appear to be discordant. Many people do, in fact, adhere to the position that they are "different" and are discordant, while others feel that they are confirmatory. To make things even more complex, one could argue that the interpretation of the p value may depend on the context of the problem (that is, should one always require the same level of evidence?).

Aside from the arguments above, there are a number of ways to "mess up" the p value. One certain way is to not follow the steps in hypothesis testing, one surprising, but not uncommon way to mess things up. Consider the following steps one researcher took: after looking at the data the investigator created a hypothesis, tested that hypothesis, and obtained a p-value; that is, the hypothesis was created from the data (see discussion of subgroup and post-hoc analysis). Forming a hypothesis from data already collected is frequently referred to as "data dredging" (a polite term for the same activity is "exploratory data analysis"). Another way of messing up the p value is to look at the

data multiple times during the course of an experiment. If one looks at the data once, the chance of a spurious finding is 0.05; but with multiple "peeks", the chance of spurious findings increase significantly (Fig. 18.2). For example, if one "peeks" at the data five times during the course of one's experiment, the chance of a spurious finding increases to almost 20% (i.e. we went from 1 chance in 20 to about a 4 in 20 chance of a spurious finding). What do we mean by peeking at the data? This frequently occurs from: interim examinations of study results; looking at multiple outcome measures; analyzing multiple predictor variables; or, performing subgroup analyses. Of course, all of these can be legitimate, it just requires planning (that is pre-planning).

Regarding subgroup analysis, It is not uncommon that after trial completion, and while reviewing the data one discovers a previously unsuspected relationship (i.e. a post-hoc observation). Because this relationship was not an *a priori* hypothesis, the interpretation of the p value is no longer reliable. Does that mean that one should ignore the relationship and not report it in one's manuscript? Of course not, it is just that one should be honest about the conditions of the discovery of the observation. What should be said in the paper is something similar to:

> *In exploratory analysis, we noted an association between X and Y. While the nominal p-value of assessing the strength of this association is 0.001, because of the exploratory nature of the analysis we encourage caution in the interpretation of this p-value and encourage replication of the finding.*

This is a "proper" and honest statement that might have been translated from:

> *We were poking around in our data we found something that is really neat. We want to be on record as the first to report this. We sure do hope that you other guys see this in your data too.*

## *Type I Error, Type II Error, and Power*

To this point, we have been focusing on a specific type of error – one where there really is no difference (null hypothesis is true) between the groups, but we are con-

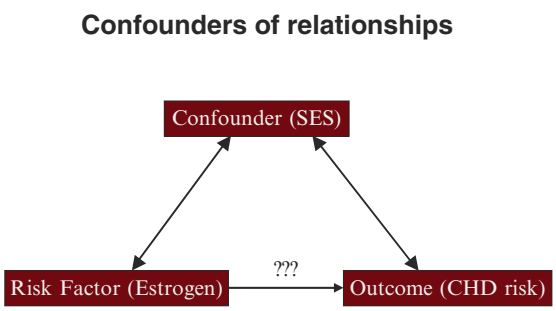



**Fig. 18.2** Depicts an example of trying to prove an association of estrogen and CHD (indicated by the question marks) but that socioeconomic status (SES) is a factor that influences the use of estrogen and also affects CHD risk separate from estrogen. As such, SES is a confounder for the relationship between estrogen and CHD risk

cerned about falsely saying there is a difference. This would be akin to a false positive result and this is termed a "Type I Error." Type II errors occur if one says there is not evidence of a difference when a difference does indeed exist; and this is akin to a false negative result (Table 18.2). To recap, recall that one initially approaches hypothesis testing with the statement that there was no difference (the null hypothesis is true), one then calculated the chance that a difference as big as the one you observed in the data was due to chance alone, and if you reject that hypothesis ($P < 0.05$), you say there really is a difference, then the p value gives you the chance that you are wrong (i.e. $p < 0.05$ means there is less than 1 chance in 20 that you are wrong and 19 chances out of 20 that you are right – i.e., that there really is a difference). Table 18.1 portrays all the possibilities in a $2 \times 2$ table.

## *Statistical Power*

Statistical power (also see Chapter 15), is the probability that given that the null hypothesis is false (i.e. that there really is a difference) that we will see that difference in our experiment. Power is influenced by:

- The significance level ($\alpha$): if we require more evidence to declare a difference (i.e. a lower p value – say $p < 0.01$), it will be harder to get, and the sample size will have to be larger, as this determination will allow one to provide for greater (or less) precision (i.e. see smaller differences).
- The true difference: this is from the null hypothesis (i.e. big differences are easier to see than small differences).
- The other parameter values related to "noise" in the experiment. For example, if the standard deviation ($\delta$) of measurements within the groups is larger (i.e., there is more "noise" in the study) then it will be harder to see the differences that exist between groups.
- The sample size (n). It is not wrong to think of sample size as "buying" power. The only reason that a study is done with 200 rather than 100 people is to buy the additional power.

To review, some major conceptual points about hypothesis testing are:

- Hypothesis testing is making a yes/no decision.
- The order of steps in statistical testing is important (the most important thing is to state the hypothesis before seeing the data).

**Table 18.2** A depiction of type I and type II error

| | Null hypothesis: No Difference | Alternative hypothesis: There is a difference |
|---|---|---|
| Test conclusion of no evidence of difference | Correct decision (you win) | Incorrect decision (you lose) $\beta$ = type II error |
| Test conclusion of a difference | Incorrect decision (you lose) $\alpha$ = type I error | Correct decision (you win) 1 - $\beta$ = power |

- There are many ways to make a mistake, including
  - Saying there is a difference when there is not one
    - By design, the α level gives the chance of a Type I error
    - The p-value is the chance in the specific study
  - Saying there is not a difference when there is one
    - By design, the β level gives the chance of a type II error, with 1- β being the "power" of the experiment
    - Power is the chance of seeing a difference when one truly exists
    - P-values should be interpreted in the context of the study
    - Adjustments should be made for multiple "peeks" (or interpretations should be made more carefully if there are multiple "peeks") – see Fig. 18.3

## Univariate and Multivariate (Multivariable) Statistics

To understand these analyses one must have an understanding of confounders (also see Chapters 3 and 17). A confounder is a factor that is associated with both the exposure (say a risk factor) and the outcome; and, leads to a false apparent association between the two. Let's use, as an example, the past observational data on the beneficial association of hormone replacement therapy and beta carotene on atherosclerosis, MI and stroke risk (Figure 18.3). When RCTs were performed, these associations not only disappeared, but there was a suggestion that some of these exposures were potentially harmful. Confounders are one of the major limitations of observational studies (recall that for RCTs, randomization equalizes known and unknown confounders between the interventional and control groups so they are not a factor in the

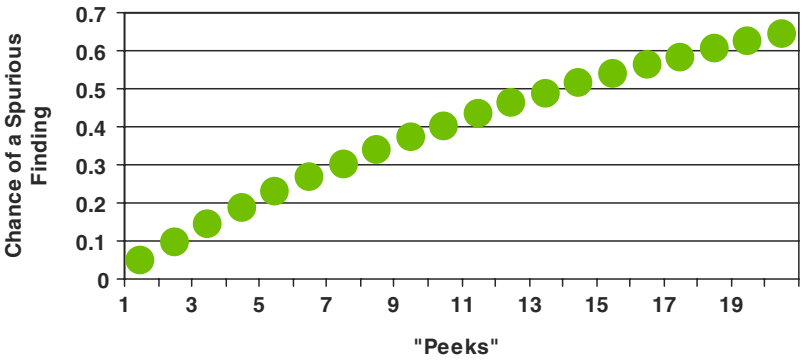



**Fig. 18.3** The chance of spurious findings related to the number of times the data is analyzed during the course of a trial

observed associations). In observational studies, however, it is necessary to "fix" the effect of confounders on the association one is trying to evaluate. In observational studies there are two basic ways of "fixing" confounders: (1) match the interventional and control groups for known confounders, at the start of the study, or (2) to adjust for potential confounders during data analysis. One should note that either of these approaches can only "fix" *known* confounders, which is unlike randomization which also "fixes" any *unknown* confounders (this being one of the major reasons that RCTs result in the highest level of scientific evidence). Remember too, that for something to be a confounder it must be associated with both the exposure and the outcome. In a case-control study, for example, one matches the cases and controls (for example by matching for age and race) so that there can be no association between those confounders (age and race) and the outcome (i.e., the cases and controls have the same distribution of race and age – because they were made to).

A way to mathematically adjust for confounders is multivariate analysis. That is, in case-control, cross-sectional, or cohort studies, differences in confounders between those with and without the "exposure" can be made to be equal by mathematical adjustment. Covarying for confounders is the main reason for multivariate statistics. The interpretation of the exposure variable in a multivariate model is "the impact of a change in the exposure variable at a fixed level of the confounding variable(s)." Saying that the association of the predictor and the outcome "is at a fixed level of the confounding variable" is the same as saying that there is not an association between the exposure and the confounding variable (really, that the relationship has been "accounted for").

Again however, many things can "go wrong" in multivariate analysis. As already mentioned, one must know about the confounders in order to adjust or match for them. In addition, one must be able to appropriately measure confounders (take SES for example, since there is much argument as to what components should make up this variable the full effect of SES may be difficult to account for in the analysis). Not only can one not quantify parts of a confounder, a confounder can never be perfectly measured and as a result confounders can not be perfectly accounted for. Also, even when a potential confounder is identified, the more measurement error there is in the confounder, the more likely that "residual confounding" can still occur.

## Bayesian Analysis

One of the many confusing statistical concepts for the non statistician is the argument over which approach-frequentist or Bayesian-is preferable. With the frequentist approach (this has become the traditional approach for clinical trials) an assumption is made that the difference between treatment groups is unknown and the parameter is fixed (for example, the mean SBP of all British citizens is a fixed number). With the Bayesian approach (some argue becoming a much more common approach in the future) parameters are assumed to be a distribution of potential differences between treatment groups and that there is information existent about

these differences before the trial you are going to perform is done. This latter idea defines one of the major strengths of the Bayesian approach-that is that one can use prior information (prior distribution) known from other studies before one conducts their trial, and this can be "added to the information gained in the current trial (posterior distribution) with the potential benefit of a reduced sample size necessary to show a difference (with the freqeuntist approach one starts statistically with a clean slate). Howard et al argues for the frequentist approach by noting that "we have a difficult time agreeing what we know" – that is the choice of studies guiding the prior knowledge is largely subjective.[1] They also argue that if there is little prior knowledge there would be no meaningful reduction in sample size, while substantial prior knowledge brings into play the ethical need to do a new study. Finally, they argue, that there are at least two reasons why previous studies might provide incorrect information (sampling variation which can be adjusted for, and bias which cannot) and the inclusion of these in the prior distribution then adversely affects the posterior distribution.[1] Berry argues that the Bayesian approach is optimal because it is "tailored to the learning approach", that is as information is accrued one "updates what one knows"; and, that this flexibility makes it ideal for clinical research.[2]

## Selection of Statistical Tools (or Why Are There So Many Statistical Tests?)

Each research problem can be characterized by the type and function of the variable and whether one is doing single or repeated assessments of the data. These are the characteristics of an experiment that determine the statistical tool used in the study. The first characteristic that influences the choice of which statistical "tool" to use, is the *data type*. Data types are categorical, ordinal or continuous. Categorical data (also called nominal or dichotomous if one is evaluating only two groups), is data that are in categories i.e. neither distance nor direction is defined e.g. gender (male/female), ethnicity (AA, NHW, Asian), or outcome (dead/alive), hypertension status (hypertensive, normotensive). Ordinal data, is data that are in categories and direction but not distance, good/better/best; normotensive, borderline hypertension, hypertensive. With continuous (also called interval) data, both distance and direction are defined e.g. age or systolic blood pressure.

*Data function* is another characteristic to consider. With data function, we are dealing with whether the data is the dependent or independent variable. The dependent variable is the outcome in the analysis and the independent variable is the exposure (predictor or risk factor).

Finally, one needs to address whether single or repeated assessments are being performed. That is, a single assessment, is a variable that is measured once on each study participant (for example baseline blood pressure measured on two different participants); while repeated measures (if there are two measures, it is also called "paired measures") are measurements that are repeated multiple times (frequently

**Table 18.3** The Statisticians "Toolbox"

| | | Type of independent data | | | | | |
|---|---|---|---|---|---|---|---|
| | | Categorical | | | | Continuous | |
| | | Two samples | | Multiple samples | | | |
| Type of dependent data | One sample (focus usually on estimation) | Independent | Matched | Independent | Repeated measures | Single | Multiple |
| Categorical (dichotomous) | 1 Estimate proportion (and confidence limits) | 2 Chi-square test | 3 McNemar test | 4 Chi square test | 5 Generalized Estimating Equations (GEE) | 6 Logistic regression | 7 Logistic regression |
| Continuous | 8 Estimate mean (and confidence limit) | 9 Independent t-test | 10 Paired t-test | 11 Analysis of variance | 12 Multivariate analysis of variance | 13 Simple linear regression & correlation coefficient | 14 Multiple regression |
| Right censored (survival) | 15 Kaplan Meier survival | 16 Kaplan Meier survival for both curves, with tests of difference by Wilcoxon or log-rank test | 17 Very unusual | 18 Kaplan-Meier survival for each group, with tests by generalized Wilcoxon or generalized log rank | 19 Very unusual | 20 Proportional hazards analysis | 21 Proportion hazards analysis |

at different times), for example, repeated measures on the same participant at baseline and then 5 years later, or blood pressures of siblings in a genetic study (in this latter case the study is of families not people, and there are two measures on the same family). Why do there have to be so many approaches to these questions? Just as a carpenter needs a saw and a hammer for different tasks, a statistician needs different types of analysis tools from their "tool box" (Table 18.3).

## References

1. Howard G, Coffey C, and Cutter G. Is Bayesian analysis ready for use in Phase III randomized clinical trials? Beware the sound of sirens. Stroke 2005; 36:1622–1623.
2. Berry D. Is the Bayesian approach ready for prime time? Yes! Stroke 2005; 26: 1621–1622.

# Chapter 19
# Grant Writing

**Donna K. Arnett and Stephen P. Glasser**

**Abstract** Perhaps nothing is more important to a new investigator than how to properly prepare a grant to request funding for clinical research. In this chapter we will review the basic elements for successful grant writing, discuss advantages and disadvantages of K versus R applications for National Institutes of Health (NIH) funding, illustrate the "fundamentals" for each section for a standard NIH R-series application, and describe the key components necessary to transition to a successful NIH research career.

## Basic Tenets of Grant Writing

The three fundamental principles involved in the successful preparation of an NIH grant are to understand the mission of the particular NIH branch from which you wish to secure funding, to know the peer review process, and to build the best team possible to accomplish the work proposed. It is very important, particularly to new investigators, to secure collaborators for areas in which you lack experience and training. While this often proves to be challenging for the new investigator since it is difficult to secure the attention of busy senior investigators, it is a critical step toward securing funding for the work you propose. Finally, grant writing, like any skill, can only be optimized by doing it repeatedly. You can read all about the physics of learning to ride a bicycle, but until one does it repetitively, one will not be good at it. The same is true with respect to grant writing: writing, editing, and rewriting of the grant should occur on a regular basis.

Having all the tools described above in your toolbox, however, will not necessarily lead to a successful grant. The ideas must be presented, or "marketed" in such a way as to show the review team the importance of the proposed work as well as its innovative elements. The grant proposal must be presented in an attractive way and the placed information where reviewers expect to find it. Complex writing styles are also ill advised for grants. It is important to use clear and simple sentence structures, and to avoid complicated words. Also avoid the temptation to use abbreviations to save space since many abbreviations, or unusual abbreviations, make a grant difficult to read. Instead, use a reviewer friendly approach where the formatting is simple and the

S.P. Glasser (ed.), *Essentials of Clinical Research*,

font is readable. Organize and use subheadings effectively (e.g., like a blueprint to the application), and use topic sentences for each section that build the "story" of your grant in a logical and sequential way. Use spell-checking programs before submission, and also, ask a colleague to read through the final draft before submission. Most importantly, be consistent in specific aims and format throughout the application.

## The Blueprint of a Research Grant

For the scientist, the most important content of the NIH grant for which the proponent is fully responsible consists of the

Abstract
Budget for initial period
Budget for 5 year period
Introduction (revised or supplemental applications)
Research Plan which includes:

- Specific aims
- Background and significance
- Preliminary studies/progress report
- Research design and methods
- Use of human subjects
- Use of vertebrate animals
- Literature cited
- Data sharing plan

There are many administrative forms that also must be included from your agency (such as the face page and the checklist, to name a few), but the items described above are where you will spend the majority of your time. It is important to carefully read the instructions, and also to check with your agency's grants and contracts officer to resolve any questions early in the process of preparing your application.

## Writing the Research Grant

In writing the research grant, start with strength by clearly articulating the problem you will address and how it relates to the present state of knowledge. Find the gap in knowledge and show how your study will fill that gap and move the field closer to the desired state of knowledge. Pick the "right" question, knowing that the question should have potential to get society closer to an important scientific answer while at the same time knowing that there are many more questions than one can answer in an individual career. In other words, get the right question, but don't spend much time figuring out what the right question is that you don't move forward. The question should lead you to research that have the potential for being fun.

While securing NIH funding is an important milestone in your career, remember if your study is funded, you will be doing it for at least the next 2–5 years and it will impact your future area of research. Don't propose any research question that you really do not think you will enjoy for the "long term". Aside from the fun aspect (which is an important one), the "right" research question should lead to a hypothesis that is testable, that is based upon existing knowledge and fills and existing gap in specific areas of knowledge. Finally, the "right" research question is a question that can be transformed into a feasible study plan. How does one find the "right" research question? Open your eyes and observe: patients often provide clues into what is known and unknown about clinical practice. This approach formed the basis of one of the authors R01 ("does the variable left ventricular hypertrophy response in the context of hypertension have a genetic basis?"). Another way of coming by the "right" research question is through teaching and through new technologies.

## *Abstract*

The abstract and specific aims (described below) are the two most important components of any grant application and must provide a cohesive framework for the application. The abstract provides an outline of the proposed research for you and the reviewer. Include in the abstract the research question that the study will address with a brief justification to orient the reviewer, the overall hypotheses to be tested, the study population you will recruit, the methods you will use, and the overall research plan. These details are important so that study section personnel can decide which study section best fits the grant. The final statement in the abstract should indicate how the proposed research, if, successful, will advance your field of research. Always revise the abstract after your complete proposal has been writin so that it agrees with what you have written in the research section.

## *Developing a Research Question and Specific Aims*

In developing a research question, one needs to choose a "good" or the "right" question as discussed above (also see Chapter 2). The "right" research question should lead you towards a testable hypothesis about the mechanisms underlying the disease process you are studying. A testable hypothesis will also require a feasible experimental design such that you can test the various predictions of your hypotheses in the most rigorous way so that your study does all that it can to fail to refute the null hypothesis if it is true. Once you have a testable hypothesis and feasible and rigorous design to translate the research question into the hypothesis, there are certain necessary components that one needs to consider. Certainly, the hypothesis should define the study purpose, but should also address: the patient/subject eligibility (i.e., characterize the study population); the exposure (or the intervention); the comparison group; and the endpoints (outcomes, dependent variable). As

described by Hulley et al. the criteria of a good hypothesis is that it is feasible, interesting, novel, ethical, manageable in scope, and relevant.[1] It is helpful to engage colleagues to respond to how novel and interesting the hypothesis is and to address whether the results of your study will confirm extend, or refute prior findings, or provide new knowledge. Arguably, the most common mistake a new investigator makes it to not have narrowly focused the question such that it is feasible to answer with the research proposed. That is, is the question is too broad or vague to be reasonably answered. Finally, include only experiments that you and your colleagues and you're your institution have the expertise and resources to conduct.

For the NIH grant, the hypotheses are written in Section A of the proposal, named "Specific Aims." Specific aims are extensions of your research questions and hypotheses, and they should generally be no more than one page and should include (i) a brief introduction that underscores the importance of the proposed research, (ii) the most important findings to date, and (iii) the problem that the proposed research will address. Using the example of the genetic determinants of ventricular hypertrophy mentioned above, the aims section began with "(i) LVH is a common condition associated with cardiovascular morbidity and mortality….(ii) we have shown that LVH is, at least in part, genetically determined…. (iii) we anticipate these strategies will identify genetic variants that play clinically significant roles in LVH. Such knowledge may suggest novel pathways to be explored as targets for preventive or therapeutic interventions".

Even though the specific aims should be comprehensive in terms of the proposed research, the aims should be brief, simple, focused, and limited in number. Draft the specific aims like you would a novel such that you create a story that builds logically (i.e., each aim should flow logically into the next aim). The aims should be "realistic", that is, they should represent one's capacity for completing the work you propose and within the budget and the time requested. Use a variety of action verbs, such as characterize, create, determine, establish, delineate, analyze, or identify, to name a few. Most importantly, keep the aims simple, at the appropriate level of your team's expertise, and where you have supporting preliminary data.

Writing specific aims can take on a variety of models. One model might be to have each aim present a different approach that tests a central hypothesis. Another model may be to have each aim develop or define the next logical step in a disease process. You should avoid a model in which an aim is dependent of the successful completion of an earlier aim. In other words, do not have aims that could only successfully move when and if the earlier aim is successful. Such contingent aims reduce the scientific merit of the grant since reviewers cannot assess their probability of success.

## *The Background and Significance Section*

The background and significance section must convince your reviewers that your research is important; in other words, you must market your idea to reviewers in such a way that it engages them intellectually and excites them in terms of the

potential for impact on clinical practice, and ultimately, health. You must also provide the foundation for your research, and show your knowledge of the literature. To provide the reviewer evidence of your ability to critically evaluate existing knowledge, the background and significance section should not only clearly state and justify the hypotheses, but should also justify variables and measurements to be collected, and how the research will extend knowledge when the hypotheses are tested. The wrap-up paragraph should discuss how your proposed research fits into the larger picture and demonstrate how the work proposed fills an important gap in knowledge. Some key questions to address are:

- What is the current state of knowledge in this field?
- Why is this research important? Does it fill a specific gap in knowledge?
- What gaps in knowledge will this project fill?
- More generally, why is this line of research important?

Captivate the reviewer by emphasizing why the research question is fascinating. For instance, what is known? What question is still unanswered? And why do we want to answer this particular question? Finally, you must address what your proposed project has to do with the public health or clinical medicine.

Background and significance sections will be read by experts in your field since reviewers are selected based on their matched expertise with your project. Therefore, you must be both factual and provide "readable" material. Whenever possible, use cartoons or diagrams to clarify concepts and to visually break up the page. It is also useful to create a "road map" for your application in the introductory paragraph (e.g. in one of the author's section, the following was used: "in this section, we review (1) the epidemiology of hypertension; (2) the pathophysiology of hypertension; (3) other medical consequences of hypertension; (4) the clinical treatment of hypertension; (5) the genetics of hypertension, and (6) implications for proposed research").

Having this roadmap is particularly important for the reviewer, since often a busy reviewer may only skim headings. Your headings within the background and significance section should lead the reviewer to know fully why that section is in the application. Like the specific aims, it is important to keep the background and significance section simple, to avoid jargon, to define acronyms, to use "sound bites", and repeatedly use these "sound bites" throughout the application. Finally, engage a colleague from a close but unrelated field to read the background section to test the ease of understanding of its structure and content to a non-expert…

## *Preliminary Studies Section*

*The best predictor of what you will do tomorrow is what you did yesterday*

The NIH has specific Instructions for the preliminary studies section, and "suggest" this section should provide an account of the principal investigator's preliminary studies relevant to the work proposed and/or any other information – from the

investigator and/or the research team – that will help to establish their experience and competence to pursue the proposed project. Six to eight pages are recommended for this section. Content should include previous research, prior experiments that set the stage for proposal and build the foundation for the proposed study. The pilot data provided should be summarized using tables and figures. Interpretation is also important so that you demonstrate your ability to articulate accurately the relevance of your pilot data and correctly state the impact of your prior work. In a related way, this section also uses the previous results to demonstrate the feasibility of your proposed project. To convince reviewers of your research feasibility, you should discuss your own work – and that of your collaborators – on reasonably related projects in order to convince reviewers that you can achieve your research aims. Pilot studies are required for many (but not all) R-series grants, and are extremely important to show your project is "do-able".

The preliminary study section is particularly important for junior investigators where there may be inadequate investigator experience or training for the proposed research, a limited publication record, and/or a team that lacks the skill set required for the research proposed. The quality of the preliminary study section is critically important for junior investigators as the quality of the presentation of the pilot work is evidence of your ability to complete the work you propose.

## *Research Design and Methods*

The research design and methods section is the place where you cover all the materials and methods needed to complete the proposed research. You must leave adequate time and sufficient space to complete this section. Many applicants run out of time and page requirements before the last aim is addressed in sufficient detail, significantly weakening the application. As concordant with the aims, it is important to not be overly ambitious. In the opening paragraph of this section it is also an important time to re-set "the scene" by refreshing the reviewer regarding the overview for each specific aim. Sometimes, this is the section where reviewers began to read the application. As you progress, use one paragraph to overview each specific aim, and then to deal with each sub-aim separately.

You should be clear, concise, yet detailed regarding how you will collect, analyze, and interpret your data. As stated in the specific aims section, it is important to keep your words and sentence structure simple because if the reviewer is confused and has to read your proposal numerous times, your score will suffer. At the end of this section give your projected sequence or time table. This is the section to convince reviewers that have the skills, knowledge and resources to carry out the work, and that you have considered potential problems and pitfalls and considered a course of action if your planned methods fail. Finally, by providing data interpretation and conclusions based on the expected outcome, or on the chance that you find different results than expected (a not uncommon occurrence), it demonstrates that you are a thoughtful scientist.

One should provide a bit of detail for each section, such as addressing the design chosen for your research project and why you chose that design rather than another, what population you will study and why, what will be measured and how it will be operationalized in the clinical setting, and on what schedule. Develop each specific aim as a numerical entity by reiterating it, and using **BOLDING** or a text box in order to highlight it. Briefly re-state the rationale for your each aim.

### Patient Enrollment

Convey to the reviewer your appreciation for the challenges in recruiting. Discuss from where the population will be recruited, what the population characteristics (gender, age, inclusion and exclusion criteria) will be, how subjects will be selected and the specific plans for contact and collaboration with clinicians that may assist you. Provide any previous experience you have with recruitment and include some numbers of subjects, and response rates, from previous or preliminary studies. Provide strategies to remedy any slow recruitment that might occur. Be cognizant of NIH policies in order to properly address issues related to gender, minority, and children inclusions and exclusions.

One also needs to consider and address the participant burden for the proposed research in order to properly weigh the benefits and costs of participation… In many studies, research subjects should be paid but not to the degree that it is coercive.

### Methods

One should provide details for the most important techniques to be used in your research. For commercially available methods you need only to briefly describe or reference the technique; but, for methods crucial to your aims, you need to provide adequate description such as referencing published work, abstracts, or preliminary studies.

In the author's experience, there are some common weaknesses of the Methods Section. These weaknesses include such issues as an illogical sequence of study aims and experiments; that subsequent aims (also known as contingent aims) rely on previous aims such that if the previous aims fail, the study comes to a halt. Inadequate descriptions of contingency plans, or poorly conceived plans, or plans that are not feasible significantly weaken a proposal. Other weaknesses include not adequately describing or constructing the control groups; and/or underestimating the difficulty of the proposed research.

## *Tips for Successful Grants*

A successful grant proposal generally "tells a story" and engages the reviewer. The proponent should anticipate questions that are likely to occur and present a balanced

view for the reviewers. To be successful, you must not take things for granted, and you must deliver a clear, concise, and simply stated set of aims, background, preliminary studies, and experimental methods that has addressed threats to both internal and external validity. You must be able to follow directions precisely and accurately, and target your grant to the expected audience (i.e., your reviewer). Your timeline and budget must align with your aims. As stated earlier, you should obtain an independent review both from your mentors and collaborators, but from external reviewers if possible. And finally, and perhaps most importantly, remember, not every proposal gets FUNDED!, in fact only a minority get funded so it is prudent to submit a number of different proposals, understanding that you won't get funded unless you submit proposals. When resubmitting proposals you should be careful to revise it based upon the critique and realize that reviewers are attempting to help you make your study better. There is no use getting mad – get funded instead! Every application must be above any level of embarrassment (i.e., do not submit anything that is not your best work). Develop a game face after submission, and be confident about your proposal. To maintain your sanity through the process, convince yourself that your grant won't get funded while concurrently reminding your colleagues it is tough to get funded.

## Types of NIH Research Funding

There are a number of types of NIH research funding, but of most relevance to clinical research are:

Grant (investigator initiated)
Cooperative agreement (NIH is a partner; assistance with substantial involvement)
Contract (purchaser)
Training awards
Research career development awards
Mentored NIH career development awards
K01/K08 research/clinical scientist
K23 and K24 patient oriented research
Mentored research scientist development award (K01)

These awards provides support for intensive, supervised career development experience, leading to research independence for early or mid-career training, as well as to provide for a mechanism for career change (K24). The K24 requires that the applicant have a substantial redirection, appropriate to the candidate's current background and experience, or that the award provides for a significant career enhancement. "Unlike a postdoctoral fellowship, the investigator must have demonstrated the capacity for productive work following the doctorate, and the institution sponsoring the investigator must treat the individual as a faculty member."

The characteristics of the ideal candidate may vary. For example, the candidate may have been a past PI on an NIH research or career development award; but, if the

proposed research is in a fundamentally new field of study or there has been a significant hiatus because of family or other personal obligations, they may still be a candidate for one of these awards. However, the candidate may not have a pending grant nor may they concurrently apply for any other career development award.

**Summary** Remember; logically develop your aims, background, preliminary studies and research design and methods into a cohesive whole. Clearly delineate what will be studied, why it is important, how you will study it, who(m) you will study, and what the timeline is to complete the research. When writing, say what you're going to say, then say it, and finally summarize what you said. Write a powerful introduction if you are constructing a revised application. Develop your "take-home messages" and reiterate them throughout your application. Finally, be tenacious: learn from your mistakes, pay careful attention to critiques, collaborate with smart people and find a good mentor. Keep it simple.

## Reference

1. Hulley SB, Cummings SR, Browner WS, et al. *Designing Clinical Research*. 2nd ed. Philadelphia, PA: Lippincott Williams & Wilkins; 2000.

# Part IV

Now that the research has been done, how is it presented? That is, how is it presented to the media and to colleagues? This Part also discusses the mentoring process that is necessary for the optimal development of a junior faculty member into an independent researcher.

*Before I give my speech, I have something important to say.*

Grocho Marx

# Chapter 20
# The Media and Clinical Research

**Stephen P. Glasser**

*Media is a word that has come to mean bad journalism.*
*http://thinkexist.com/search/*

**Abstract** The news media are an increasingly important source of information about new medical treatments. The media can be persuasive, pervasive, and can influence health care beliefs and behaviors. This chapter briefly addresses the maturation process of medical controversy, discusses some of the reasons for the "tension" that develops between scientists and the media, and hopefully allows the reader when they are asked to discuss their research findings, to develop some strategies for dealing with the media.

The media (whether we like it or not) is playing an increasing role in helping or confounding the transmission of knowledge to patients. The news media are an increasingly important source of information about new medical treatments. The media can be persuasive, pervasive, and can influence health care beliefs and behaviors.[1] Caspermeyer et al. investigated nine large newspapers to determine how often the coverage of neurological illness contained errors and stigmatizing language.[2] They determined that medical errors occurred in 20% and stigmatizing language in 21% of the articles evaluated. In another report, seven stories regarding three preventative treatments (cholesterol, osteoporosis, and aspirin) were analyzed.[3] Of those media reports, 40% did not report benefits quantitatively; of those that did, 83% reported relative (not absolute) benefits only, while 98% reported potential harm.

In 1997 Weber reviewed the "natural history" of reports on medical controversies (approximately a 10 year process) which I believe are instructional.[4] The first phase in the natural history of media reports about medical innovations, he entitled the Genesis Phase. During the Genesis Phase new information is identified. The next phase in the natural history of media reporting is the Development Phase, where questions of safety and/or efficacy about the innovation arise; print and broadcast publicize the debate; and, complex issues tend to be oversimplified and/or sensationalized. This is followed by the Maturation Phase where more data and studies become available, but public interest by this time tends to be waning and media

S.P. Glasser (ed.), *Essentials of Clinical Research*,

coverage is less intense. Finally, there is the Resolution Phase where objective re-evaluations are published, and a more fair-balance of the pros and cons of the innovation are presented. Weber presents two examples of this natural evolution process: the silicone gel breast implant; and, the calcium channel blocker (CCB) controversies, the latter of which is discussed below.

The genesis of the CCB controversy began in 1995 when Psaty et al. presented a Case Control Study from a single center suggesting that short-acting nifedipine could harm patients treated for hypertension (specifically they reported an increased risk of myocardial infarction).[5] The RR for harm was reported as 1.6. The Development Phase was evident after the American heart Association published a press release which was hyped by the media. Many who were treating patients with hypertension at that time will recall being inundated with telephone calls from concerned patients. Examples of the news reports are shown in Fig. 20.1.

The CCB controversy that arose was followed by a meta-analysis (see Chapter 10) of 16 studies also suggesting the same harm.[6] Subsequently, all CCBs were said to be harmful and furthermore were said to be associated with cancer and GI bleeding.[7,8] During the Maturation Phase of this controversy, the FDA and NIH reviewed the CCB data and gave them a clean bill of health (with the exception of short-acting CCBs). Reanalysis of the data began to show the flaws in the methodology of studies impugning the CCBs. The methodological flaws included selection bias and prescription bias, that is, sicker patients were more likely to be given CCBs. In the Resolution Phase (8–10 years after the controversy began), the CCB controversy was "put to rest" most recently by ALLHAT.[9] It should be noted that during this process another issue surfaced relative to the Multicenter Isradipine Diuretic

**Blood pressure drugs may increase heart risk**

An Associated Press Report

SAN ANTONIO — Six million Americans who are taking a class of drugs to lower blood pressure may instead be increasing their risk of heart attacks by 60 percent, researchers reported Friday.

The drugs, called calcium channel blockers, are nifedipine, diltiazem and verapamil. They are sold under various brand names, including Adalat, Calan, Cardizem, Dilacor, Isoptin, Procardia and Verelan.

National guidelines recommend the use of beta blockers and diuretics to lower high blood pressure.

Calcium channel blockers also have been widely used to treat high blood pressure, but in recent years studies have shown they don't work very well and are not considered the best treatment.

However, many American cardiologists still prescribe them for that purpose, said study author Bruce Psaty, an epidemiologist at the University of Washington in Seattle.

"We are very concerned about these results," he said.

Previous studies have suggested that calcium channel blockers can be dangerous when given to patients immediately after a heart attack, or when they are given some time later to prevent a second heart attack.

This is the first suggestion that they are dangerous in the large group of outwardly healthy people who are trying to prevent heart attacks by lowering their blood pressure, Psaty said.

About 50 million Americans suffer from high blood pressure, greatly increasing their chances of a heart attack. Half of those are being treated, and about 25 percent of the patients being treated are getting calcium channel blockers, Psaty said. That amounts to about 6 million Americans.

Paul Ridker, a cardiologist and epidemiologist at the Brigham and Women's Hospital in Boston, who wasn't involved in the study, said doctors continue to use calcium channel blockers in part because drug companies have "marketed them beyond what the data shows."

Psaty studied 623 people who had had heart attacks and a control group of 2,032 who hadn't.

When he determined the drugs they had been taking, he found that those on calcium channel blockers had a heart disease risk 60 percent higher than those taking beta blockers or diuretics. The study was funded by the National Institutes of Health.

**Report links margarine, heart disease**

Associated Press Report

SAN ANTONIO — Men who ate six teaspoons of margarine per day had twice the heart disease risk of men who didn't eat any margarine, a study has found.

The study failed to find any harmful effects from the consumption of butter, but researchers said that finding was probably due to the study's limitations.

The most recent findings were based on information gathered from 1966-1969 in the Framingham Heart Study. The risks are probably much lower now, because margarine sticks sold in the 1960s had higher levels of trans fatty acids than most of the tubs of soft margarine and low-fat spreads sold today.

Matthew Gillman of Harvard Medical School reported his findings Thursday at the American Heart Association's epidemiology meeting.

***Heart Attacks May Have Tie To Drug Type***

***Scientist Describes Risk as 'Minuscule'***

By GINA KOLATA

A new study suggests that people who take one type of high blood pressure medication may be more likely to have heart attacks.

The drugs, known as calcium channel blockers, may increase the risk of a heart attack by 60 percent compared with other types of medication, the researchers found. But heart disease experts cautioned that the study did not prove an effect and some questioned the findings.

The researchers, led by Dr. Bruce Psaty of the University of Washington in Seattle, reported their findings on Friday at a meeting of the American Heart Association in San Antonio. Dr. Psaty and others agreed that the study was not definitive but said that even if the finding was correct, the day-to-day risk to an individual taking a calcium channel blocker was "minuscule."

**Fig. 20.1** Two examples of media reports on the CCB controversy

Atherosclerosis Study (MIDAS), a large multi-center study that compared the effects of isradipine (a short-acting CCB) compared to the diuretic hydrochlorothiazide on the course of carotid artery disease in hypertensive patients.[10] The investigators found that the progression of carotid atherosclerosis did not differ between the two treatment groups, but that there was an increased incidence of vascular events in patients treated with the CCB. A side issue in this study was the withdrawal of some of the investigators from the manuscript preparation due to what they perceived as "undue influence" exerted by the sponsor of the study. Needless to say, this resulted in some interesting media reporting such as "a high-tension drug study has been reported".

Why the media publicized this controversy and deemed it newsworthy while another controversy is not so publicized seems to be a mystery to most readers and listeners. In great part the publicizing of such studies depends upon what the media editors think will have "headline potential". As Semir noted, "…news of killer bacteria, exterminating viruses, and miraculous therapies tend to have greater appeal because such stories compete with murders, rapes, ecologic catastrophes, and declarations from famous people…"[11] In fact, this author had a personal experience following publication of 13 subjects who underwent a roll-a-coaster ride.[12] The heart rate response (by ambulatory ECG monitoring) was quite impressive; but, let's face it, 13 healthy subjects with no adverse outcomes? Yet this became a story for national media attention, probably because there had been a few recent deaths on similar rides throughout the country. Marilyn Chase reported in the Wall Street Journal ways of putting hyped study results under the microscope.[13] Every week, she noted, medical science makes headlines with a promising new study or "cure", and it is "often hard to tell ephemeral findings from epochal breakthroughs-especially when distilled into a few paragraphs or sound bites spiced with hype."[13] Interestingly, she cites a number of questions that need to be addressed in media reports, questions that should sound familiar from reading chapters in this book, regarding clinical trial methodology. Some of the questions Chase cited were: Was the study large enough to make it significant? Was the study fair i.e. were the two groups equally matched? Who paid for the study? Who was the control group? Were volunteers randomly assigned? Was there appropriate blinding?

Deary et al. report their media experience with a study that had been reported in Lancet.[14] The Lancet report concluded that women with more submissiveness were less likely to have myocardial infarction compared to those women who were less submissive. The Lancet publication was under embargo (a topic to be discussed shortly); however, a newspaper ran the story prematurely under the headline "put down that rolling pin, darling, its bad for your heart". Other headlines included "do as you're told girls…and live to be old", "stay home and you'll live longer", "do what hubby says and you will live longer", and "meekness is good for a women's heart…" The authors further note that one phone interview included questions like: "So these feminists are all barking up the wrong tree?" and, Should women be getting back to the kitchen sink?" Of course, these questions did not accurately represent what the study in fact showed, and I recommend reading Deary's editorial, as it should be instructive to all researchers interested in communicating their studies results.

The importance of the media in providing the public with health information should not be underestimated. Timothy Johnson (in the 108th Shattuck Lecture) noted a survey in which 75% of the respondents said they pay either a great deal or moderate amount of attention to the medical and health news reported by the media; and, 58% said that they have changed their behavior or have taken some type of action based upon what was reported (read, seen, or heard).[15] Thus, the role of the clinical researcher in providing news to the media is important. Some basic tenants for the researcher to follow are: be certain you are the best person to provide the media with the necessary information; do not digress – start with your main conclusion first and then do not wander; consider the two to three points that are important about ones study, and keep returning to those points; do not become defensive or argumentative; and, be concise – particularly with television interviews. As an example of the above let us assume that you have hypothetically just published a study on the benefits of a new drug and the interview proceeds with a question such as "what were your primary findings?" Having briefly discussed the outcomes with great pride, the reporter than asks "but doctor weren't there three deaths in your study and do you really think it was ethical to perform such a trial?" The response by most of the uninitiated would go something like this – "yes there were 3 deaths, but in this population we expected there to be deaths, and blah blah blah". In general it is best not to repeat the negative, and the answer perhaps could have been better shaped with something like "the important thing is that we found a significant overall benefit of our new drug treatment, and this was in a very sick population. In addition we did everything possible to protect the safety of our patients." Many might remember the very funny interview in the Bob Newhart comedy television series, when off camera a very pleasant reporter pumped up Newhart's ego, and when they went live totally blind-sided him with embarrassing and demeaning questions such as "since psychologists hardly ever cure anyone, don't you think the fees that you charge them are outrageous?". In actuality, this type of blind-siding is rare with health reporting, the reporter is generally your colleague, and is attempting (with their limited knowledge) to impart accurate information, but being prepared for that occasional problem is not a bad idea.

## Control of Information (The Embargo Rule)

Perhaps the most important issue that results in researcher-media conflicts is the long struggle over the "Ingelfinger rule" since it involves the control of information, a control the media despises. The pressure to be the first or to be able to claim to be the exclusive report of a story results in significant tension when they are asked to hold (embargo) a story until it is published in a scientific journal.

Scientists also expect that they are the ones to control the flow of information, and view the media as but a pipeline to inform the public about recent discoveries.[1] Most journalists, however, do not view themselves merely as a spokesperson for the scientist, but rather they view their role as raising probing questions about the research. In

fact, both scientists and journalists are committed to communicating accurate information, but the media aims for brevity, readability, simplicity; and, are usually pressured by time constraints; whereas the scientist has been working on the research that is being reported for years, are interested in precautionary qualifications, and are aware that their scientific readership can assimilate the nuances of their research.[1]

In summary, the media is playing an increasing role in the reporting of health news. Most health reporters are attempting to write a credible and accurate story. The enduring tensions between medicine and the media are largely due to the different perspectives between researchers and journalists. As Nelkin noted, "these tensions arise because of perceived differences in defining science news, conflicts over styles of science reporting, and most of all disagreement about the role of the media".[16] It is incumbent upon the researcher, if they are going to accept a media interview, to know how to present clear concise answers to question about their research.

## References

1. Fishman JM, Casarett D. Mass media and medicine: when the most trusted media mislead. *Mayo Clin Proc*. Mar 2006; 81(3):291–293.
2. Caspermeyer JJ, Sylvester EJ, Drazkowski JF, Watson GL, Sirven JI. Evaluation of stigmatizing language and medical errors in neurology coverage by US newspapers. *Mayo Clin Proc*. Mar 2006; 81(3):300–306.
3. Moynihan R, Bero L, Ross-Degnan D, et al. Coverage by the news media of the benefits and risks of medications. *N Engl J Med*. June 1, 2000; 342(22):1645–1650.
4. Psaty BM, Heckbert SR, Koepsell TD, et al. The risk of myocardial infarction associated with antihypertensive drug therapies. *JAMA*. Aug 23–30, 1995; 274(8):620–625.
5. Weber MA. The Natural History of Medical Controversy Consultant 1997.
6. Furberg C, Psaty B, Meyer J. Nifedipine. Dose-related increase in mortality in patients with coronary heart disease. *Circulation*. 1995; 92:1326–1331.
7. Jick H. Calcium-channel blockers and risk of cancer. *Lancet*. June 7, 1997; 349(9066): 1699–1700.
8. Pahor M, Guralnik J, Furbert Cea. Risk of gastrointestinal hemorrhage with calcium antagonists in hypertensive patients over 67. *Lancet*. 1996; 347:1061–1066.
9. Major outcomes in high-risk hypertensive patients randomized to angiotensin-converting enzyme inhibitor or calcium channel blocker vs diuretic: the Antihypertensive and Lipid-Lowering Treatment to Prevent Heart Attack Trial (ALLHAT). *JAMA*. Dec 18, 2002; 288(23):2981–2997.
10. Borhani NO, Mercuri M, Borhani PA, et al. Final outcome results of the Multicenter Isradipine Diuretic Atherosclerosis Study (MIDAS). A randomized controlled trial. *JAMA*. Sept 11, 1996; 276(10):785–791.
11. de Semir V. What is newsworthy? *Lancet*. Apr 27, 1996; 347(9009):1163–1166.
12. Glasser SP, Clark PI, Spoto E. Heart rate response to "Fright Stress." Heart Lung. 1978; 7:1006–1010.
13. Chase M. How to put hyped study results under a microscope. *Wall Street J*. 1995; 16:B-1.
14. Deary IJ, Whiteman MC, Fowkes FG. Medical research and the popular media. *Lancet*. June 6, 1998; 351(9117):1726–1727.
15. Johnson T. Shattuck lecture–medicine and the media. *N Engl J Med*. July 9 1998; 339(2):87–92.
16. Nelkin D. An uneasy relationship: the tensions between medicine and the media. *Lancet*. June 8 1996; 347(9015):1600–1603.

# Chapter 21
# Mentoring and Advising

**Stephen P. Glasser and Edward W. Hook III**

*Advice is like mushrooms. The wrong kind can prove fatal.*

–Unknown

**Abstract** Mentorship refers to the development of a relationship between a more experienced individual (the mentor) with a less experienced individual (the mentee or protégé). The role of the mentor in the development of the junior faculty member's academic relationship is extremely important. As such, this chapter discusses the expectations of the mentor, mentee, and the mentor-mentee relationship.

## Mentoring vs. Advising

Mentorship refers to the development of a relationship between a more experienced individual (the mentor) with a less experienced individual (the mentee or protégé). The word itself was inspired by the character of Mentor in Homer's Odyssey. Historically, mentorship goes back to ancient Greek and Hindu times. Today, the definition of mentor continues to encompass 'a trusted counselor or guide', and a 'wise, loyal advisor or coach.'

Mentoring in the research sense developed mostly in the basic science laboratories, where an experienced researcher would literally take a junior person 'under their wing' and would help them develop research independency. This concept has been taken up by the NIH through its K23 and K24 programs, but this has been a relatively recent development (see below). The problem has always been, that there is little in the way of formal training in how to be a good mentor, and there is usually no external reward for the time spent in mentoring.

In academics, mentoring and academic advising are frequently used synonymously, but we view advising as a lesser responsibility than mentoring. One can over-simplistically say that advising is an 'event' while mentoring is a 'process'. A mentor has both a professional and personal relationship with the mentee, an advisor, in general, does not, to the same degree, have a personal relationship. Also, mentoring is more dynamic, in that there is a distinct change over time.

S.P. Glasser (ed.), *Essentials of Clinical Research*,

Although there is no single formula for good mentoring, most would agree that a good mentor is approachable and available, and this is where good mentoring most often comes up short, since in a busy academicians life (who has multiple demands, and has requirements for promotion, research grants, manuscripts, etc.); little academic reward is provided for mentoring. Although perhaps more empathetic with the role of the mentee, junior faculty are often ill-equipped to serve as mentors. Factors militating against effective mentorship by junior faculty include an (appropriate) emphasis on one's own career advancement, limited resources to devote to the mentee, and limited opportunities to promote the mentee's career by virtue of limited personal recognition as a result of being early in one's career. Students, for their part, must recognize the professional pressures and time constraints faced by their mentors, but still must insist on obtaining adequate time and availability from their mentors, or be willing to change who their mentor is. Much misunderstanding can be circumvented with a well intentioned discussion about these issues prior to choosing a given mentor. As such, both the mentor and mentee should be clear about their respective expectations, have a clear agreed upon career development plan, with regular meetings a priority. On the one hand, the mentor cannot be to busy, otherwise they should not have accepted the responsibility, but the mentee cannot expect unlimited access.

## Guidelines for Faculty/Student Interactions

Faculty members often develop a close working relationship with students, especially advisees. Often a relationship is formed that provides benefits to both the faculty member and the student. Faculty should be cognizant of the power differential in these types of relationships and set appropriate boundaries. Although faculty members may not intend a favor or request to be an obligation, they should be aware that this may place some students in a difficult position. Some students are intimidated by faculty members and may not feel free to decline such requests. http://www.epi.umn.edu/academic/pdf/FacAdvising.pdf. It is recognized that many situations are ambiguous. Examples are of some of these ambiguous situations include:

- **Asking a student to drive you someplace, including the airport, home, or main campus**. Such a request does not fall under a student's duties. A situation when this may be acceptable is when the student has the same destination.
- **Asking a student to work extra hours or late hours**. Students should be expected to work the hours they are paid for. Students may volunteer to put in extra hours to gain more experience (e.g. grant writing) or gain authorship on a paper or help meet a deadline – but these extra hours should not be an expectation.
- **Asking an advisee to housesit, take care of your children or pets, or help you move**. While some students may not mind house sitting, taking care of children

or pets, or helping someone move, others may only agree to do this because they feel obligated or worry that saying no will somehow affect their relationship with the faculty member. To avoid this situation, faculty members may post a request for a sitter or mover for pay without any faculty names attached to the flyer – ensuring that respondents really want this job.

## Advising

Expectations for advising vary between institutions but mainly in terms of frequency of meetings. It seems to these authors that minimal expectations should include:

(1) Academic advisors should meet with their advisees at least twice per semester, but more often is preferable. These meetings should be scheduled, but there should also be opportunities for ad hoc meetings to deal with acute problems.
(2) Academic advisors should respond in a timely manner to requests from advisees for meetings or responses by telephone or e-mail, even if this is to schedule the requested meeting.
(3) Academic advisors should provide general guidance to students about coursework, fieldwork, project selection, and career planning.
(4) Academic advisors should make students feel welcome to the Division.
(5) Academic advisors should act as a contact person for the student and help direct them to the appropriate resources in the Division given whatever issues or problems the students may have.
(6) Academic advisors should act as a resource for the student when bureaucratic or political problems in the University, School or Division may be interfering with the student's effective progress toward his or her degree.
(7) Although the advisors role is to help the advisee to not over-extend themselves, they should also help them see what an important opportunity is.

Advising may include a number of diverse activities such as procedural advising (e.g. should the student drop a course), academic advising e.g. how satisfied are they with the program, career planning, selecting course work), and advising 'students' on the conduct of their research. Excellent advising requires a significant time commitment.

What are the mentor's responsibilities? They should find out what are the junior investigators career goals, how often formal meetings should take place, what the mentor's expectations are (this should be spelled out in terms of frequency of meetings, metrics, and outcomes), devise the best way to communicate (face to face, e-mail, telephone). The advisee also has responsibilities. They should take the lead in scheduling meetings, and contacting the advisor if there are problems. Finally, there should be clear expectations of what protected time will be provided for the mentee's career development. If this is not under the control of the mentor, the mentor should aid the mentee in establishing protected time with whoever the responsible person is. There are many pitfalls in the term 'protected time'. One of the most important is

the denominator for calculating it. For example, is the percentage of protected time based upon a 40, 60, or 80 hour-week. What other responsibilities will the mentee have (i.e. clinics, ward rotations, committee meetings, teaching, conferences etc.). When there are multiple mentors, who will have the overall 'big picture'

## K23 and K24 Awards

The NIH has developed a number of Career Development Programs (K awards; Figs. 21.1–21.3), in fact there are 13 different awards available and these are dependent upon such factors as one's career stage and how they may interact with other NIH Awards. However, there are common features of NIH career awards, such as salary, fringe benefits, and research/development costs, salary caps, research/development costs, and award duration. In addition, entry level awards require a mentor, and at least 75% protected time for the awardees to spend on research and other career development activities. For non-mentored senior awards 25–50% is required. Eligibility for NIH awards requires a Doctoral Degree (generally), that the applicant be a US citizen, Non-Citizen National, or a Permanent Resident. Should the awardee change their Institution or Mentor prior approval of the NIH awarding component must be advised.

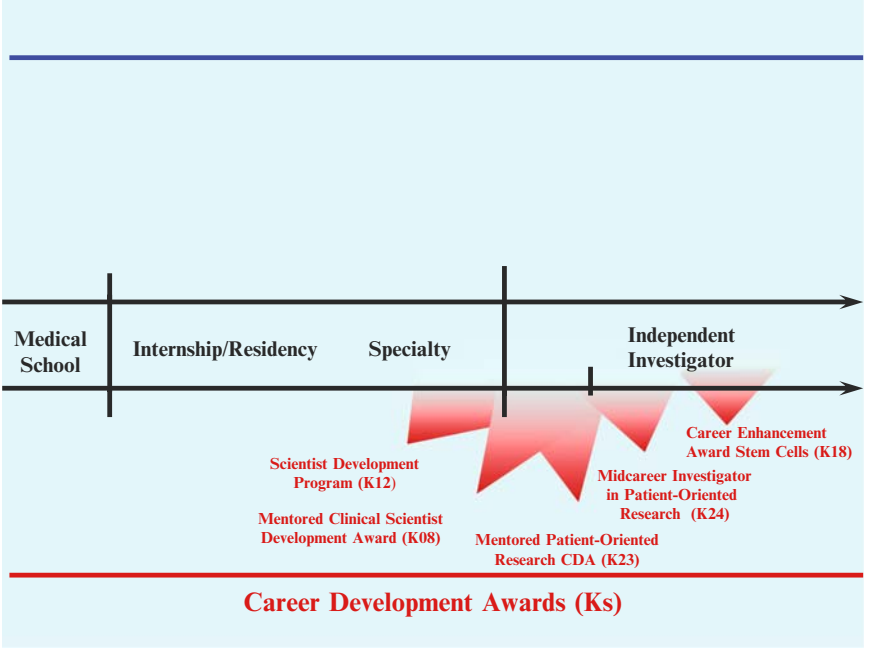



**Fig. 21.1** Career development awards (Ks)

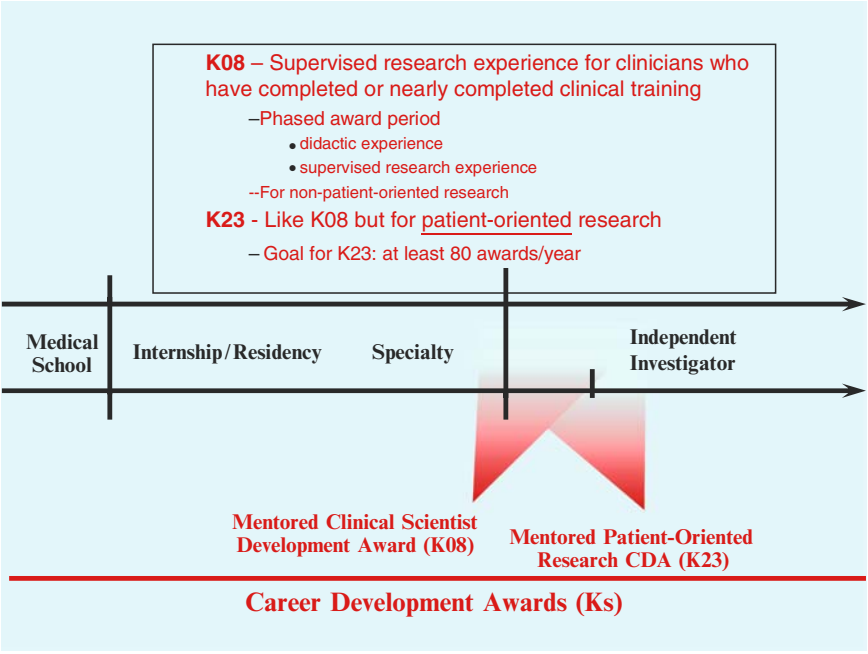



**Fig. 21.2** Career development awards (Ks)

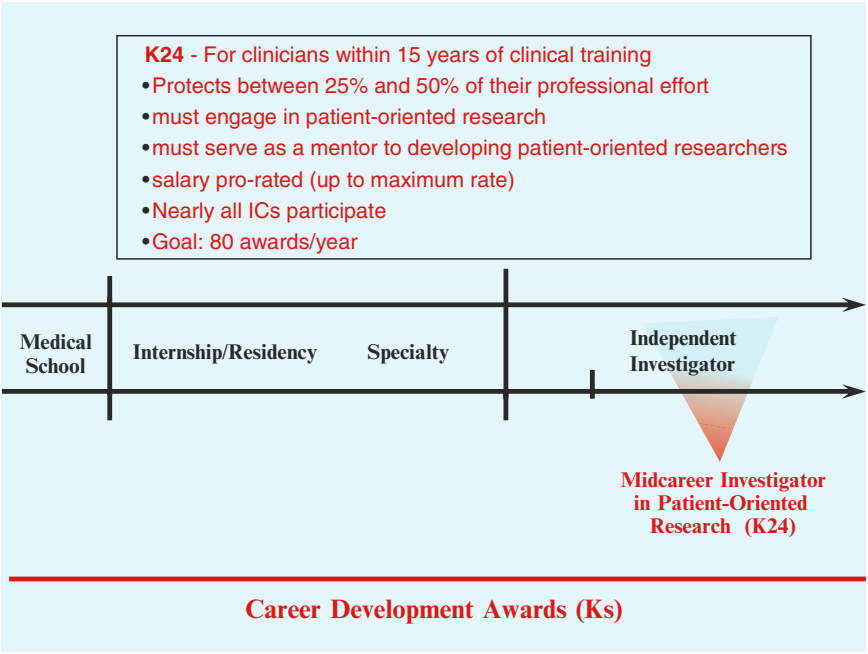



**Fig. 21.3** Career development awards (Ks)

For most of the readers of this book, the K23 award is likely to be the most appropriate. The guidelines for K23 Awards include an application that includes information about the nature and extent of supervision that will occur during the award period (co-mentors must supply similar information), and there must also be a career development plan that incorporates a systematic approach towards obtaining the necessary skills necessary to become an independent researcher. This plan should include course work appropriate to the experience of the candidate. The mentors research qualifications in the area of the project and the extent and quality of his/her proposed role in guiding and advising the mentee, as well as previous experience in mentoring is critical. The application must include the applicant's career goals and objectives with a detailed description of what the candidate wants to achieve following the completion of the award.

The K23 application should be very detailed about the mentor's role and responsibilities, how the mentor's area of expertise relates to the research interests of the applicant, how often the applicant will meet with the mentor (and co-mentors), what will happen during those meetings, and how short-comings in the applicant's performance will be addressed. The mentor, on the other hand, should provide the same information, as well as extol the mentor's virtues with prior mentoring activities.

The application should also contain information about formal coursework that will be taken in support of the applicant's career plan, and ideally one that will lead to a degree, such as a Master of Science Degree in Clinical Research (a K30 supported Program). Ideally, the applicants plan will include both an Internal as well as an External Advisory Committee which is formed to provide an objective review of the candidate's progress. More details are spelled out in the grant description, but these are the key components that have been problematic in K23 grants that I have reviewed.

The K24 is a senior non-mentored award that is a natural extension once the K23 is completed. It allows for funded protected time to mentor junior investigators, particularly those seeking a K23 award.

In summary, a number of pitfalls face the junior faculty member interested in a career in patient oriented research. A good mentor/advisor can be of enormous help in guiding young researchers toward their career goals. Unfortunately, many mentors/advisors, acting as role models have fallen into the same traps that they should be preventing in a new researcher, so the mentors role-modeling is somewhat tarnished. We agree with Grigsey that five of the most important pitfalls in the mentor-mentee relationship are: committing to excessive service time; 'diffusion and confusion' i.e. a new faculty member has no clue as to what is or is not a priority without a good advisor guiding them; lack of mentoring/advising; exploitation by other faculty; and, lack of discipline and perseverance.

# Chapter 22
# Presentation Skills: How to Present Research Results

**Stephen P. Glasser**

*Speech is power; Speech is to persuade, to convert, to compel*

Ralph Wald Emerson

**Abstract** This book is about designing, implementing and interpreting clinical research. This chapter is aimed at a discussion of how to present the research that has been performed. Although almost no one currently disagrees that a formal curriculum in research methodology is critical for a new investigator, the manner in which the results of a study are presented is presumed to be obvious, and training in the art of presentations is much less common. The belief is that good speakers are born, not made, and this is no more true than good researchers are born and not made. And so, the methodology of presentations should be an important part of a young investigators training. This chapter provides an introduction to delivering an effective presentation.

## Introduction

This book is about designing, implementing and interpreting clinical research. This chapter is aimed at a discussion of how to present the research that has been performed. Although almost no one currently disagrees that a formal curriculum in research methodology is critical for a new investigator, the manner in which the results of a study are presented is presumed to be obvious, and training in the art of presentations is much less common. The belief is that good speakers are born, not made, and this is no more true than good researchers are born and not made. And so, the methodology of presentations should be an important part of a young investigators training. The ability to communicate effectively is a key to professional success. The investigator who wants to express complex ideas, inform, and educate realizes that effective presentations are an important skill. If you are relatively inexperienced and suffer from stage-fright, relax – you are not alone. Public speaking ranks at the top of the list of peoples fears surpassing even the fear of death. But like any skill, public speaking takes training, experience, persistence, motivation and practice.

In a handbook by Foley and Smilansky[1] the authors quote Frost as follows, 'in a lecture given by a brilliant scholar, with an outstanding topic, and a highly

S.P. Glasser (ed.), *Essentials of Clinical Research*,

competent audience, ten percent of the audience displayed signs of inattention within 15 minutes. After 18 minutes, one third of the audience and 10% of the platform guests were fidgeting. At 35 minutes everyone was inattentive; at 45 minutes trance was more notable than fidgeting; and at 48 minutes some were asleep and at least one was reading. A casual check 24 hours later revealed that the audience recalled only insignificant details, and these were generally wrong.' How long should a talk be? 'A speech, like a bathing suit, should be long enough to cover the subject-but short enough to be interesting'.[2]

What is the least efficient way of communicating a lot of information, particularly technical information? Think about it, and the answer will probably be the oral presentation. Why? for a number of reasons, the most important being that the ear is a limited learning tool. Additionally, the oral lecture is of low efficiency, is associated with low audience recall, and forces the audience to assimilate the information on the speakers schedule, in contrast to a written document or an audio tape or DVD, where a 'student' can review the information at a time when there are no other deadlines that have to be met, or an upcoming appointment for which they do not want to be late etc. Also, the information can be reviewed and re-reviewed at their leisure, important points underlined, and so on. So what is it about the oral presentation that makes it so valuable? Two things: the rapport the speaker can gain with the audience, and the ability of the audience to ask the 'expert' (defined as someone who lives more than 50 miles away and has slides) questions. In fact, some studies have shown that how a lecture is perceived is 55% visual, 38% related to how the speaker sounds, and 7%, the content. The cliché goes that a famous professor is introduced, and with much fanfare walks to the podium, calls for the lights to be dimmed, and says 'for my first slide….' thereby removing the 55% visual component needed to gain the necessary rapport that renders the oral presentation so valuable in the first place. If the lights go down, and you can no longer see the speaker, you might as well have an audio tape playing. Standing behind the podium (a protective mechanism) or leaning on it (a message of disinterest), also takes away from the presentation, so when possible it is to be avoided.

## The Structure of a Presentation

The old adage for the outline of a talk is the Introduction to the talk - tell them what you are going to tell them; the Body of the talk - Tell them; and, the Conclusion - tell them what you've told them. Because your audience is most attentive during the introduction and conclusion, those are really the most important parts of the presentation, and of the two probably the introduction is the key in gaining their attentiveness, and the conclusion is most important for the take home messages. Thus, if possible, memorize the conclusion so you do not have to look at the slide, but rather you can look directly at the audience while you make your concluding remarks. During the introduction you have a free ride for about 2 minutes and it is during this time, if you use it wisely, that you need to catch the audience's attention. This author likes to use 'hooks' or 'grabbers' during the introductory comments, such as a joke-

but be careful in this era of political correctness this can backfire (I have had it happen to me!) or the use of a short video clip relevant to the topic which can engage the audience and demonstrate to them that you have given thought to the presentation. Self-effacing humor (if not overdone) can be useful, a speaker who can laugh at him or herself gains rapport with the audience.

Some examples of 'grabbers' follow: Grocho Marx's famous quote of 'Before I speak, I have something important to say'; Or, for a presentation about a drug that caused sinus bradycardia, but had no other hemodynamic effect, this author once began a presentation by asking the audience what they thought the most important anti-ischemic mechanism of beta adrenergic blockers was. Most of the audience answered 'sinus bradycardia' after which I responded 'that was my thought as well, but now I am going to tell you about a drug that slows the sinus rate but has no anti-ischemic effects'. Catchy titles for your talk also demonstrate to the audience that you have given some thought to your presentation. Some examples I have used were: 'What do the first flight of the Columbia and quinidine have in common?' (for a talk on re-entry as a mechanism of arrythmias), or 'What do the Japanese puffer fish and silent ischemia have in common? Alliterations can be catchy also, such as 'Palpitations, Prolapse, and Palpating the Pachyderm' (for this talk on mitral valve prolapse-by the way, I began this talk with the famous poem of the blind man palpating different parts of the pachyderm and coming away with different impressions about what the animal might look like; in order to make the simile of the many ways that mitral valve prolapse can present clinically). Posing the title of your talk as a question can get the audience thinking, and changes them from taking a passive to becoming an active role in your presentation, thereby gaining more attentiveness. Or, posing a question in the opening of your introduction such as 'how many of you have patients who have suffered an MI despite the LDL being at goal?' During the author's first exposure to formal training in presentation skills, I was asked to prepare a 5 minute presentation. I entitled it 'What do exercise testing and stratigraphy have in common? Digging for answers' – the thesis of the talk being stratification (layers) of risk based on exercise test results, just as a stratigrapher tries to make interpretations based upon the rock layers they observe. In addition to the 'grabber' one should also begin with the thesis of the talk, that is, the 'what's in it for the audience question'. One should also cover the outline of the presentation. The outline should have no more than five points and ideally three points, because studies have shown that after a 10 minute presentation, the average listener forgets 25% of what was said within the first 24 hours and 80% within 4 days.[3] By highlighting the three main points of your presentation and repeating them in the conclusion, you increase the chances that your audience will at least remember the most important points that you wanted to communicate. However, the outline of your presentation should be specific rather than broad. I have heard speakers who have picked up on the point that an outline is important, but unknowingly have 'gotten around it' by using broad general topics. As an example, I heard one speaker, talking on the metabolic syndrome, have an outline that included outline points like: 'I will cover lipid metabolism, the different definitions of metabolic syndrome, and all the treatment options; when the focus of the talk was really to discuss whether the metabolic syndrome was a precursor to diabetes.

## Stages of a Speaker

Almost all speakers have to go though three stages before they become accomplished presenters. The speed with which they traverse these stages depends upon their personalities and whether one follows the precepts outlined in this chapter.

*Stage 1* is the fear centered stage. Novice speakers are almost always more nervous than the situation dictates, but being nervous (stage fright) is common to even the most experienced speaker. I remember when Johnny Carson was doing his umpteenth monologue and it was being telemetered as part of the show. Before he went on stage and as he was being introduced his pulse rate surged to 120 bpm! Many novice speakers read from a prepared text to help deal with nerves, but a speech that reads well does not necessarily 'listen' well.

*Stage 2* is the speaker-centered stage which is characterized by imparting the points you as a speaker want to make. You have now given enough presentations that there is the appropriate amount of nervousness, you know your subject well, and then you go about presenting everything you know about it. The underlying motivation is probably to impress upon your audience how much you do know, and it is your job to tell them everything! The fact is that for most audiences you will know more about the subject you are presenting then they will (exceptions might be at a national specialty meeting), and this is where another major mistake is made by the stage 2 speaker-assuming a level of knowledge that is really not present and thereby leaving the audience in the dark. This fly's in the face of what a good speech should be- clarity, simplicity, and repetition (it is a good idea in talks over 15 or 20 minutes that after each point you have elaborated in your outline, that you repeat what you just said in one sentence-this entrenches the bullet point that you want them to 'take home'); that is, present a small number of essential ideas, simplicity, and being conversational (see, I just did it) are the attributes of a good presentation. You should strive for keeping your message simple for three reasons: (1) so that you can remember it, (2) so that the audience will understand it, and (3) so that the audience will remember it. Novice speakers and speakers frozen in stage 2 are also notorious for apologizing – apologizing about not having enough time to cover the subject, for not having had time to prepare adequately, for the time of day, month, or season; and, for anything else they can think of. I remember one speaker apologizing for something, then catching himself and apologizing for apologizing! My advice is never apologize! Deal with what you are dealt and go on with it!

*Stage 3*. It is the third stage that every good speaker should strive for-this is the audience-centered stage characterized by understanding the audience, having a feel for what they really need to know; and, that is dependent upon who the audience is. The fact is, that expectations among most audiences, accustomed to the general inadequacy of speakers, are so low that almost any well-intentioned bumbler is, at the very least, accepted – provided that the speaker doesn't drone on too long. With this knowledge, the speaker should now be confident enough in their knowledge of the subject, and relaxed enough that they can control their nervousness. They can now focus on what the specific audience to whom they are presenting absolutely needs to know about the subject-and with almost every subject this can be accomplished with

three to five main points. It is the integration of the last two stages that makes an excellent speaker, and the approach to message building is fundamental to the art of 'getting to the point'. It is also the stage where you know when to stop! Never, never, never, go over the allotted time, you will not impart any additional information to the audience, and you will antagonize them. I have heard many complaints about talks that have gone on to long, but I have never heard anyone complain about a talk that is to short. One characteristic of the presenter still frozen in stage 2, but knowledgeable enough that he or she knows not to go over time, is to simply take the same amount of material but talk faster; rather than reducing the number of points to be covered. These latter presenter's are sometimes dubbed the 'speed demon' or the 'talking encyclopedia', and this should obviously be avoided.

## Audiovisuals

Audiovisuals should be used-but not overused. Most speakers use audiovisuals as a crutch rather than the stepping stones that helps an audience understand the message the speaker is trying to make. Many (most?) speakers also crowd too much information on a slide, and some, knowing that the slides are too crowded, even apologize for it. Comments such as 'I know you cannot see it because the print is too small, but the point I am trying to make is…' If you know it cannot be seen why are you using it for? Epidemiologists are renown for using to much detail in their slides (I can say this because I am one). One of my mentors (Dr. Roy Behnke-referred to as 'Revrend Roy' behind his back because of the way he preached his presentations) used three to five slides for an entire Grand Rounds presentation-and those slides had at the most three lines on each. My suggestion is to synthesize the information as is shown in Tables 22.1 and 22.2. In general, three bullet points per slide is ideal and each slide should have only one unifying idea. The other common mistake speakers make with slides is related to the use of the pointer. As an experiment one day, watch the eyes of the audience as the speaker uses the pointer like a weapon and is roaming all over the slide instead of holding it steady on the point that they wish to emphasize. As the eyes follow the pointer the listener is distracted from the point that is being made. In fact, if you use a limited number of lines per slide, you can also minimize your use of the pointer, minimize pointer wander, and for those of us who are red-green color blind, it will not matter that one cannot see the red dot from the pointer in the first place.

An accomplished speaker arrives at the venue early enough to become familiar with the AV equipment so that they do not stumble around trying to control the lights (remember to keep the lights as high as possible while ensuring that the slides can be seen by the audience). Reviewing the slide advancement mechanism (hopefully on a PowerPoint or related computer presentation format) is also important so that when their actual presentation begins there is not a lot of stumbling (recall the importance of the opening impression one makes on the audience).

**Table 22.1** Table of eight studies which would be good for a manuscript but difficult for an audience to digest during a presentation

| | | Ability of exercise ST depression to predict subsequent CAD events (%) | | | |
|---|---|---|---|---|---|
| Study | N | Mortality w/+ETT | Events w/+ETT | Mortality w/−ETT | Events w/−ETT |
| Ref 1 | 210 | 25 | – | 1 | – |
| Ref 2 | 85 | 22 | – | 0 | 0 |
| Ref 3 | 130 | 10 | 40 | 7 | 4.5 |
| Ref 4 | 46 | 4 | 41 | 0 | 0 |
| Ref 5 | 195 | – | 38 | 0 | 0 |
| Ref 6 | 62 | 20 | – | 2 | – |
| Ref 7 | 236 | 13 | 5 | 7.5 | 3.8 |
| Ref 8 | 48 | 23 | – | 4 | – |

**Table 22.2** Ability of exercise ST depression to predict subsequent CAD events (pooled analysis) (Ref 1, Ref 2 etc)

| N | Mortality w/+ETT | Mortality w/−ETT | Events w/+ETT | Events w/−ETT |
|---|---|---|---|---|
| 817 | 17% (4-25) | 4% (1-7.5) | 31% (5-41) | 1.6% (0-4.5) |

## The Question and Answer Period

The two main fears about the Q and A are that no questions will be asked, or that questions will be asked for which you do not know the answer. To elicit questions, be invitational such as 'I have been looking forward to your questions' or 'I would be happy to answer any questions'. If there are none, try jumping in with something like 'I am almost always asked about…', and this frequently gets the Q and A going. When a question is asked, keep the answer brief (this is not the time for a mini-talk); and, if you do not know the answer, it is fine to say something like 'I do not know-do you have experience in this area?' – no-one expects you to know everything even if you are 'the expert'. Also, ALWAYS repeat the question so members of the audience who did not hear it are not left out. You can also sometimes rephrase the question so that it is clearer. If the question has nothing to do with the presentation, one can either very briefly address it and then segue into the points you feel are important, or say you would be happy to answer it individually after the Q and A period.

There are a number of other things a speaker can learn about presentations, such as how to answer questions, how to deal with an audience member who is carrying on a conversation during the presentation, the heckler, the know-it-all, the media etc. One should take advantage of courses, seminars etc. that teach these skills. As an example, during a formal seminar on presentation skills, our talks were videotaped and then played back. One of my colleagues – an accomplished speaker – (fortunately it was not me – I had plenty of my own affectations) had his finger in

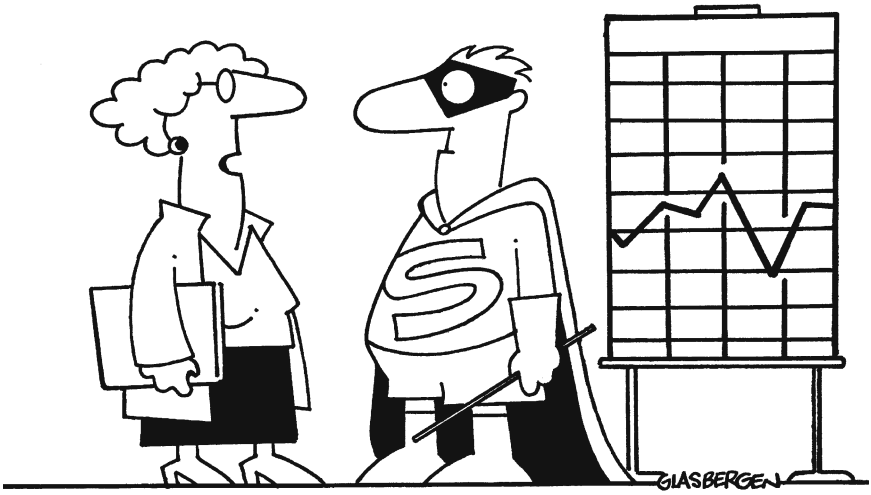
his ear during the entire 5 minute mock talk. He was totally unaware that he had done that and even questioned whether the tape had been altered.

As a researcher, it is becoming more and more common to interact with the media about research that you have done (see Chapter 20). Answers to the media have to be even more carefully thought out, because journalists are not only interested in getting the information correctly, but want the 'headline grabber' to get people to read about it. They also unknowingly (sometimes knowingly) take things out of context. Despite my experience, I can not think of an instance where what I intended to be the message of the interview actually came out to my total satisfaction (you might want to think about this when reading a newspaper article of someone else who has been interviewed and 'quoted'). Almost never will a reporter allow you to review beforehand what they are going to print (or edit, if it is a television interview) because they feel they want to maintain their autonomy (by the way, in my view this is more important to them than getting it right). Also, there is a famous (among the presentation skills people) clip form the Bob Newhart show (the one in which he portrayed a psychologist). When he was about to be interviewed before airtime, the reporter was as sweet as sugar, telling him how wonderful his reputation was, what a great field psychology was etc. Then the lights came on, and the interviewer's first question went something like, 'Since your field never cures anyone, how can you justify the outrageous fees you charge?' – and it went downhill from there. Hopefully, if you have watched that series, you can imagine how the bumbling Newhart responded.

## Conclusion

I have found the following points to be critical for a good presentation.

1. A speech that reads well does not necessarily listen well
2. A good speech consists of a surprisingly small number of ideas – do not saturate the audience
3. A secret of effective speech is simplicity, another is the use of conversational language
4. Content alone will not insure a successful talk
5. Do not apologize about the topic, time etc.
6. Vary the volume of your voice, rate of speaking, etc.
7. Use pauses and inflection along with body movement to emphasize key points
8. Do not exceed your time limit
9. Stand up, speak up, and then shut up
10. Always repeat the question asked, and answer the question briefly
11. Like your presentation, keep audiovisuals simple with a limited number of points on each slide
12. Keep the room lighting as bright as possible




**"Fear of public speaking is quite common. If dressing up as Speaker Man makes you feel more confident, then so be it."**

**Fig. 22.1**


**"Always start your presentation with a joke, but be careful not to offend anyone! Don't mention religion, politics, race, money, disease, technology, men, women, children, plants, animals, food..."**

**Fig. 22.2**


**"What software would you recommend to give my presentation so much flash and sizzle that nobody notices that I have nothing to say?"**

**Fig. 22.3**

## References

1. Teaching Techniques, A Handbook for Health Professionals, R Foley and J Smilansky, McGraw-Hill, New York, 1980
2. The majority of this chapter was taken from personal experience and extensive notes that I had taken from a large number or Presentation Skills Workshops that I have attended. Although I cannot give specific credit for individual pieces of information, I can credit the Instructors of those workshops as follows:

   (a) Sue Castorino, President, The Speaking Specialist, Chicago, IL, 1993
   (b) Gerald Kelliher Ph.D., Associate dean, Medical College of Pennsylvania
   (c) Eleanor Lopez, Let's Communicate Better, www.eleanorlopez.com
   (d) Power Speaking, and More, Joyce Newman Communications Inc
   (e) Jerry Michaels-Senior Consultant CommCore Communication Strategies
   (f) Science and Medicine Canada, Presentation & Platform Skills Workshop, 1992
   (g) Wyeth Ayerst Laboratories, Ciba-Geigy, Schering, Pfizer, and KOS Pharmaceuticals for sponsoring many of the Presentation Skills Workshops that I attended

3. Garson A, Gutgesell H, Pinsky WW, McNamara DG, The 10-minute talk, slides, writing, and delivery, Am Heart J, 111:193–203, 1985.

# Index