

Health Informatics

Rachel L. Richesson · James E. Andrews
Editors

Clinical Research Informatics

Second Edition



Springer

Health Informatics

This series is directed to healthcare professionals leading the transformation of healthcare by using information and knowledge. For over 20 years, Health Informatics has offered a broad range of titles: some address specific professions such as nursing, medicine, and health administration; others cover special areas of practice such as trauma and radiology; still other books in the series focus on interdisciplinary issues, such as the computer based patient record, electronic health records, and networked healthcare systems. Editors and authors, eminent experts in their fields, offer their accounts of innovations in health informatics. Increasingly, these accounts go beyond hardware and software to address the role of information in influencing the transformation of healthcare delivery systems around the world. The series also increasingly focuses on the users of the information and systems: the organizational, behavioral, and societal changes that accompany the diffusion of information technology in health services environments.

Developments in healthcare delivery are constant; in recent years, bioinformatics has emerged as a new field in health informatics to support emerging and ongoing developments in molecular biology. At the same time, further evolution of the field of health informatics is reflected in the introduction of concepts at the macro or health systems delivery level with major national initiatives related to electronic health records (EHR), data standards, and public health informatics.

These changes will continue to shape health services in the twenty-first century. By making full and creative use of the technology to tame data and to transform information, Health Informatics will foster the development and use of new knowledge in healthcare.

More information about this series at <http://www.springer.com/series/1114>

Rachel L. Richesson • James E. Andrews
Editors

Clinical Research Informatics

Second Edition



Springer

Editors

Rachel L. Richesson
Duke University School of Nursing
Durham, NC
USA

James E. Andrews
School of Information
University of South Florida
Tampa, FL
USA

ISSN 1431-1917

Health Informatics

ISBN 978-3-319-98778-1

<https://doi.org/10.1007/978-3-319-98779-8>

ISSN 2197-3741 (electronic)

ISBN 978-3-319-98779-8 (eBook)

Library of Congress Control Number: 2018963302

© Springer International Publishing 2019

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: G  werbestrasse 11, 6330 Cham, Switzerland

Contents

Part I Foundations of Clinical Research Informatics

1	Introduction to Clinical Research Informatics	3
	Rachel L. Richesson, James E. Andrews, and Kate Fultz Hollis	
2	From Notations to Data: The Digital Transformation of Clinical Research	17
	Christopher G. Chute	
3	The Clinical Research Environment	27
	Philip R. O. Payne	
4	Methodological Foundations of Clinical Research	49
	Antonella Bacchieri and Giovanni Della Cioppa	
5	Public Policy Issues in Clinical Research Informatics	87
	Jeffery R. L. Smith	
6	Informatics Approaches to Participant Recruitment	109
	Chunhua Weng and Peter J. Embi	
7	The Evolving Role of Consumers	123
	James E. Andrews, J. David Johnson, and Christina Eldredge	
8	Clinical Research in the Postgenomic Era	147
	Stephane M. Meystre and Ramkiran Gouripeddi	

Part II Data and Information Systems Central to Clinical Research

9	Clinical Research Information Systems	171
	Prakash M. Nadkarni	
10	Study Protocol Representation	191
	Joyce C. Niland and Julie Hom	
11	Data Quality in Clinical Research	213
	Meredith Nahm Zozus, Michael G. Kahn, and Nicole G. Weiskopf	

12 Patient-Reported Outcome Data	249
Robert O. Morgan, Kavita R. Sail, and Laura E. Witte	
13 Patient Registries for Clinical Research	269
Rachel L. Richesson, Leon Rozenblit, Kendra Vehik, and James E. Tcheng	
14 Research Data Governance, Roles, and Infrastructure	291
Anthony Solomonides	

**Part III Knowledge Representation and Discovery: New Challenges
and Emerging Models**

15 Knowledge Representation and Ontologies	313
Kin Wah Fung and Olivier Bodenreider	
16 Nonhypothesis-Driven Research: Data Mining and Knowledge Discovery	341
Mollie R. Cummins	
17 Advancing Clinical Research Through Natural Language Processing on Electronic Health Records: Traditional Machine Learning Meets Deep Learning	357
Feifan Liu, Chunhua Weng, and Hong Yu	
18 Data Sharing and Reuse of Health Data for Research	379
Rebecca Daniels Kush and Amy Harris Nordo	
19 Developing and Promoting Data Standards for Clinical Research	403
Rachel L. Richesson, Cecil O. Lynch, and W. Ed Hammond	
20 Back to the Future: The Evolution of Pharmacovigilance in the Age of Digital Healthcare	433
Michael A. Ibara and Rachel L. Richesson	
21 Clinical Trial Registries, Results Databases, and Research Data Repositories	453
Karmela Krleža-Jerić	
22 Future Directions in Clinical Research Informatics	481
Peter J. Embi	
Index	493

Part I

Foundations of Clinical Research Informatics



Introduction to Clinical Research Informatics

1

Rachel L. Richesson, James E. Andrews,
and Kate Fultz Hollis

Abstract

This chapter provides essential definitions and overviews important constructs and methods within the subdomain of clinical research informatics. The chapter also highlights theoretical and practical contributions from other disciplines. This chapter sets the tone and scope for the text, highlights important themes, and describes the content and organization of chapters.

Keywords

Clinical research informatics definition · CRI · Theorem of informatics · American Medical Informatics Association · Biomedical informatics

Overview

Clinical research is the branch of medical science that investigates the safety and effectiveness of medications, devices, diagnostic products, and treatment regimens intended for human use in the prevention, diagnosis, treatment, or management of a

R. L. Richesson, PhD, MPH, FACMI (✉)
Duke University School of Nursing, Durham, NC, USA
e-mail: rachel.richesson@dm.duke.edu

J. E. Andrews, PhD
School of Information, College of Arts and Sciences, University of South Florida,
Tampa, FL, USA
e-mail: jimandrews@usf.edu

K. Fultz Hollis, MS, MBI
Oregon Health & Science University, Portland, OR, USA
e-mail: fultzhol@ohsu.edu

disease. The documentation, representation, and exchange of information in clinical research are inherent to the very notion of research as a controlled and reproducible set of methods for scientific inquiry. Contemporary clinical research actually represents relatively recent application of statistics to medicine and the acceptance of randomized controlled clinical trials as the gold standard [1] in this last half-century. Clinical research has been characterized as a discipline resting on three pillars of principle and practice related to control, mensuration, and analysis [2], though these can be more modernly interpreted as a triad of expertise in medicine, statistics, and logistics [3].

Clinical research informatics (CRI) is the application of informatics principles and techniques to support the spectrum of activities and business processes that instantiate clinical research. Informatics, as somewhat broadly defined as the intersection of information and computer science with a health-related discipline, has a foundation that has drawn from many well-established, theory-based disciplines, including computer science, library and information science, cognitive science, psychology, and sociology. The fundamental theorem of informatics [4] states that humans plus information technology should function and perform better together than humans alone, and so informatics is a source for supportive technologies and tools that enhance – but not replace – unreservedly human processes.

The US National Institutes of Health offers a comprehensive and widely accepted definition for clinical research that includes a spectrum of populations, objectives, methods, and activities. Specifically, this broad definition states that “clinical research is...patient-oriented research conducted with human subjects (or on material of human origin that can be linked to an individual)” [5]. Under this definition, clinical research includes investigation of the mechanisms of human disease, therapeutic interventions, clinical trials, development of new technologies, epidemiology, behavioral studies, and outcomes and health services research. This definition was used by all authors in this text to scope the content, so readers will see a broad overview of important informatics topics and constructs, as they apply to this wide spectrum of research objectives, participants, stakeholders, and activities.

Given this broad definition, clearly the challenges in clinical research – and the opportunities for informatics support – arise from many different objectives and requirements, including the need for optimal protocol design, regulatory compliance, sufficient patient recruitment, efficient protocol management, and data collection and acquisition; data storage, transfer, processing, and analysis; and impeccable patient safety throughout. Regardless of clinical domain or study design, high-quality data collection and standard, formalized data representation are critical to the fundamental notion of reproducibility of results. In addition to explicit and suitable data collection methods for reliability, strong study design and conduct (sampling in particular) are necessary for the generalizability of research findings. In the age of an electronic data deluge, standards also take on critical importance and can facilitate data sharing, knowledge generation, and new discovery using existing data sets and resources.

Contexts and Attempts to Define Clinical Research Informatics

The driving forces for the rapid emergence of the CRI domain include advances in information technology and a mass of grassroots innovations that are enabling new data collection methods and integration of multiple data sources to generate new hypotheses, more efficient research, and patient safety in all phases of research and public health. While the range of computer applications employed in clinical research settings might be (superficially) seen as a set of service or support activities, the practice of CRI extends beyond mere information technology support for clinical research. The needs and applications of information management and data and communication technologies to support research run across medical domains, care and research settings, and research designs. Because these issues and tools are shared across various settings and domains, fundamental research to develop theory-based and generalizable applications and systems is in order. Original research will afford an evidence base for information and communications technologies that meaningfully address the business needs of research and also streamline, change, and improve the business of research itself. CRI is just at the point where a defined research agenda is beginning to coalesce. As this research agenda is articulated, standards and best practices for research will emerge, as will standards for education and training in the field.

Embi and Payne (2009) present a definition for CRI as “the sub-domain of biomedical informatics concerned with the development, application, and evaluation of theories, methods, and systems to optimize the design and conduct of clinical research and the analysis, interpretation, and dissemination of the information generated” [6]. An illustrative – but nonexhaustive – list of CRI focus areas and activities augment this American Medical Informatics Association (AMIA)-developed definition: evaluation and modeling of clinical and translational research workflow; social and behavioral studies involving clinical research; designing optimal human-computer interaction models for clinical research applications; improving and evaluating information capture and data flow in clinical research; optimizing research site selection, investigator, and subject recruitment; knowledge engineering and standards development as applied to clinical research; facilitating and improving research reporting to regulatory agencies; and enhancing clinical and research data mining, integration, and analysis. The definition and illustrative activities emerged from in-person and virtual meetings and interviews with self-identified CRI practitioners within the AMIA organization. The scope and number of activities, and the information problems and priorities to be addressed, will obviously evolve over time as in any field. Moreover, a single professional or educational home for CRI, and as such a source to develop a single consensus and more precise definition, is lacking at present and likely unachievable given the multidisciplinary and multinational and multicultural scope of CRI activities. However, there is some important work coming out of the AMIA CRI Working Group including an update on Embi and Payne (2009) where the role of the chief research information officer (CRIO) is defined in more detail [7]. What is important to note is that this is all reflective of the

bottom-up development of this area, reflecting the applications of information technology that have been needed and that are in use.

The first references to what is now known as clinical research informatics go back to the 1960s and highlight the inevitable use of computers to support data collection and analysis in research [8]. The use of clinical databases for research inquiry was first established in the late 1960s, and by the next decade, there were at least a handful of clinical information systems being used for research. This history is well described in Collen in a 1990 historical review. In short course, it was clear that structured data entry and data standards would be a critical component of any computerized support or analysis system in research [9]. Bloise first recognized that systems could and should support more than queries about single patient data but rather should be searchable to retrieve many patient records to support research and quality monitoring. The first applications focused on retrieval of clinical information to identify and understand patient subpopulations [10]. Others saw the potential for tapping these clinical databases in observational research and knowledge discovery; by the 1970s, cancer and tumor registries were well established, and cardiovascular disease registries emerged. For the first few decades, computers in clinical research were indeed centered around maintaining a database focused on collecting and querying clinical data. The advent of patient eligibility screening and trial recruitment systems in the 1990s represents the introduction of computers to support clinical research *processes* [11–13]. The regulated nature of human trials, especially since the formal inquiry and establishment of standards for the field in the 1970s, created a critical need for documentation of methods and process, as well as analysis and findings, and we saw systems emerge in the late 1980s that begin to address the conduct of studies. The capabilities of these systems have improved and their use has proliferated. Now, clinical research management systems of various types support the collection of data and the coordination of research tasks. The primary functionality of commercial applications today is essentially concerned with the delivery of valid and accurate data in conformity with the Good Clinical Practice (GCP) guidelines [14], and in most cases, these systems are not well integrated with patient care systems. It is only recently that information management and technologies are forcing the reengineering of work processes and identifying and creating synergies with clinical data documentation. This era is truly an exciting time of massive transformation in the management of clinical research.

The enormity of data generated from new diagnostic and measurement technologies, an increasing ability to collect data rapidly from patients or external data sources, and the scope and scale of today's research enterprises have led to a bewildering array and amount of data and information. Information technology has contributed to the information management problems by generating more data and information, but the techniques and principles derived from informatics promise to purposively utilize IT to address the issues of data collection, information management, process and protocol management, communication, and knowledge discovery and show promise to improve research efficiencies, increase our knowledge of therapeutic evaluation, and impact human health and the global economy. Still, in time these tools will need to be evaluated via more formal means and evolve or be

replaced by the next generation of tools and methods. As original informatics research and proper system evaluations – including randomized trials of various systems with outcome measures related to research efficiency, quality, and patient safety – are conducted, published, and scrutinized, *evidence* to support decision making in health care and research informatics contexts will result.

Perspective, Objectives, and Scope

This book comes during a very exciting time for CRI and biomedical informatics generally. Since the first edition of this text, we have seen new legislation (*21st Century Cures Act*) and new programs including the NIH's *All of Us Research Program* that show promise to leverage CRI to impact human health in unprecedented ways. There is a growing interest around “real-world evidence” for treatments and implementing – and generating evidence in Learning Health Systems (such as that defined by Agency for Healthcare Research and Quality): Learning Health Systems | Agency for Healthcare Research & Quality <https://www.ahrq.gov/professionals/systems/learning-health-systems/index.html>.

This collection of chapters is meant to galvanize and present the current knowledge in the field with an eye toward the future. In this book, we offer foundational coverage of key areas, concepts, constructs, and approaches of medical informatics as applied to clinical research activities, in both current settings and in light of emerging policies, so as to serve as but one contribution to the discourse going on within the field now. We do not presume to capture the entirety of the field (can any text truly articulate the full spectrum of a discipline?), but rather an array of both foundational and more emerging areas that will impact clinical research and, so, CRI. This book is meant for both scholars and practitioners who have an active interest in biomedical informatics and how the discipline can be leveraged to improve clinical research. Our aim is not to provide an introductory book on informatics, as is best done by Shortliffe and Cimino in their foundational biomedical informatics text [15] or Hoyt and Hersh [16].

Rather, this text is targeted toward those who possess a basic understanding of the health informatics field and who would like to apply informatics principles to clinical research problems and processes. Many of these theories and principles presented in this text are, naturally, common across biomedical informatics and not unique to CRI; however, the authors have put these firmly in the context of how these apply to clinical research.

The excitement of such a dynamic area is fueled by the significant challenges the field must face. At this stage, there is no consistent or formal reference model (e.g., curriculum models supporting graduate programs or professional certification) that represents the core knowledge and guides inquiry. However several informatics graduate programs across the country offer courses in clinical research informatics (Oregon Health & Science University and Columbia University, to name a couple). Moreover, from these efforts discernible trends are emerging, and research unique to CRI are becoming more pronounced. In this text, we try to cover

both of these and also identify several broad themes that undoubtedly will influence the future of CRI.

In compiling works for this book, we were well aware that our selection of topics and placement of authors, while not arbitrary, was inevitably subjective. Others in CRI might or might not agree with our conceptualization of the discipline. Our goal is not to restrict CRI to the framework presented here; rather, that this book will stir a discourse as this subdiscipline continues to evolve. In a very loose sense, this text represents a bottom-up approach to organizing this field. There is not one exclusive professional venue for clinical research informatics, therefore, no one single place to scan for relevant topics. Numerous audiences, researchers, and stakeholders have emerged from the clinical research side (professional practice organizations, academic medical centers, the FDA and NIH sponsors, research societies like the Society for Clinical Trials, and various clinical research professional and accrediting organizations such as the Association of Clinical Research Professionals) and also from the informatics side (AMIA). Every year since 2011, Dr. Peter Embi does a systematic review of innovation and science of CRI and presents it to AMIA [17]. And virtually every year, he reports a paucity of randomized interventional research of informatics applications in the clinical research domain. Yet, the research base does grow each year. This issue is illustrated in the variety of approaches authors used to cover the chapter topics. Some chapters focus on best practices and are instructional in nature, and some are theoretical (usually drawing from the parent or contributing discipline).

Watching conferences, literature, list serve announcements and discussions, and meetings from these two sides of clinical research informatics for the last few years, we developed a sense of the types of predominant questions, activities, and current issues. We then sought to create chapters around themes, or classes of problems that had a related disciplinary base, rather than specific implementations or single groups. For this reason, readers active in clinical research informatics will possibly be surprised on first glance not to see a chapter devoted exclusively to the BRIDG model or the Clinical and Translational Science Awards program, for instance. While these have been significant movements impacting CRI, we view them as implementations of broader ideas. This is not to say they are not important in and of themselves, but we wanted these topics to be embedded within a discussion of what motivated their development and the attention these initiatives have received.

Authors were selected for their demonstrated expertise in the field. We asked authors to attempt to address multiple perspectives, to paint major issues, and, when possible, to include international perspectives. Each of the outstanding authors succeeded, in our opinion, in presenting an overview of principles, objectives, methods, challenges, and issues that currently define the topic area and that are expected to persist over the next decade. The individual voice of each author distinguishes one chapter from the other; although some topics can be quite discreet, others overlap significantly at certain levels. Some readers may be disappointed at a presumed lack of chapters on specific data types (physiologic and monitoring data, dietary and nutrient data, etc.) or topics. However, to restate, it was impractical for this book to attempt to cover every aspect of the field.

Many of the topics for the book chapters rose rather easily to the surface given the level of activity or interest as reflected in national or international discussions. Others were equally easy to identify, at least to a certain extent, as fundamental concepts. Yet even at this level, it is clear that CRI is a largely applied area, and theory, if drawn from at all, tends to be pulled into different projects in a more or less ad hoc manner. As we have implied, there is a noticeable lack of a single or unifying theory to guide inquiry in CRI (though this is emerging in informatics at large).

Organization of the Book

We have attempted to organize the chapters under unifying themes at a high level using three broad sections: (1) the foundations of clinical research informatics, (2) data and information systems central to clinical research, and (3) knowledge representation and data-driven discovery in CRI, which represents the future of clinical research, health, and clinical research informatics.

Section 1: Foundations of Clinical Research Informatics

The first section addresses the historical context, settings, wide-ranging objectives, and basic definitions for clinical research informatics. In this section, we sought to introduce the context of clinical research and the relevant pieces of informatics that together constitute the *space* for applications, processes, problems, issues, etc., that collectively comprise CRI activities. We start with a historical perspective from Christopher Chute, whose years of experience in this domain, and informatics generally, allow for an overview of the evolution from notation to digitization. His chapter brings in historical perspectives to the evolution and changing paradigms of scientific research in general and specifically on the ongoing development of clinical research informatics. Also, the business aspects of clinical research are described and juxtaposed with the evolution of other scientific disciplines, as new technological advances greatly expanded the availability of data in those areas. Chute also illustrates the changing sociopolitical and funding atmospheres and highlights the dynamic issues that will impact the definition and scope of CRI moving forward. Philip Payne follows this with a chapter focused on the complex nature of clinical research workflows – including a discussion on stakeholder roles and business activities that make up the field. This is a foundational chapter as it describes the people and tasks which information and communication technologies (informatics) are intended to support. Extending the workflow and information needs is an overview of study designs presented by Antonella Bacchieri and Giovanni Della Cioppa. They provide a broad survey of various research study designs (which are described in much more detail in a separate Springer text authored by them) and highlight the data capture and informatics implications of each. Note that while the workflow and study design chapters can be considered fundamental in many respects, the workflows are ever changing in response to new regulations, data types, and study

designs. New study designs are being developed in response to new data collection activities and needs (e.g., small sample sizes). While new research methods and statistical techniques will continue to emerge, the principles of study design and research inquiry will remain constant and are fundamental background for CRI. In this edition, we have added chapter regulations. Here, Jeff Smith describes the history and motivations for regulating clinical research and describes new legislation that will impact informatics aspects of clinical research.

Following a more historical perspective and discussion of fundamentals of clinical research design and conduct, this first section includes two chapters that tackle different perspectives on patients or consumers. Chunhua Weng and Peter Embi address information approaches to patient recruitment by discussing practical and theoretical issues related to patient recruitment for clinical trials, focusing on possible informatics applications to enhance recruitment. Their chapter highlights evolving methods for computer-based recruitment and eligibility determination, sociotechnical challenges in using new technologies and electronic data sources, and standardization efforts for knowledge representation. Given the rapid advances in technology and parallel continued emphasis on patient empowerment and participation in decision making, Jim Andrews, David Johnson, and Christina Eldredge consider the changing role of consumers in health care generally and in clinical research particularly. Traditional treatments of information behaviors and health communication are discussed, building to more current approaches and emerging models. Central to understanding the implications for clinical research are the evolving roles of consumers who are more engaged in their own decision making and care and who help drive research agendas. The tools and processes that support patient decision making, engagement, and leadership in research are also briefly described here, though clearly the chapter can only touch upon them.

Finally, Chap. 8 of this section describes the increasing availability of genetic data that is becoming vital to clinical research and personalized medicine. The discussion provided by Stephane Meystre and Ramkiran Gouripeddi primarily focuses on the relationship and interactions of voluminous molecular data with clinical research informatics, particularly in the context of the new (post) genomic era. The translational challenges in biological and genetic research, genotype-phenotype relations, and their impact on clinical trials are addressed in this chapter as well.

Section 2: Data and Information Systems Central to Clinical Research

Several chapters in this section cover a range of issues in the management of various data and the systems that support these functions. At the crux of clinical research informatics is a variety of information management systems, which are characterized and described by Prakash Nadkarni. His chapter also gives a broad overview of system selection and evaluation issues. His chapter includes brief descriptions of each group of activities, system requirements for each area, and the types and status

of systems for each. Systems are discussed by organizing them by the following broad activities: study planning and protocol authoring, forms design, recruitment, eligibility determination, patient-monitoring, and safety – including adverse events, protocol management, study conduct, analysis, and reporting. Also, a section of this chapter focuses on best approaches in the analysis, selection, and design of information systems that support the clinical research enterprise. Importantly, the authors emphasize needs assessment, user-centered design, organizational features, workflows, human-computer interaction, and various approaches to developing, maintaining, updating, and evaluating software.

The importance of computerized representation of both data and processes – including the formalization of roles and tasks – is underscored by Joyce Niland and Julie Hom in their chapter on Study Protocol Representation. The essence of any clinical study is the *study protocol*, an abstract concept that comprises a study's investigational plan and also a textual narrative documentation of a research study. To date, CRI has primarily focused on facilitating electronic sharing of text-based study protocol documents. Niland and Hom propose a much more powerful approach to leveraging protocol information using a formal representation of eligibility criteria and study metadata.

Common to all clinical research protocols is the collection of data. The quality of the data ultimately determines the usefulness of the study and applicability of the results. Meredith Zozus, Michael Kahn, and Nicole Weiskopf address the idea that central to clinical research are data collection, quality, and management. They focus on various types of data collected (e.g., clinical observations, diagnoses) and the methods and tools for collecting these. Special attention is given to the development as use of case report forms (CRFs), historically the primary mechanism for data collection in clinical research, but also the growing use of EHR data in clinical research. The chapter provides a theoretical framework for data quality in clinical research and also will serve as practical guidance. Moreover, Nahm et al. draw on the themes of workflows presented by Payne in Chap. 3 and advocate explicit processes dedicated to quality for all types of data collection and acquisition.

An important source of data, data reported by patients, is described thoroughly by Robert Morgan, Kavita Sail, and Laura Witte in the next chapter on “Patient-Reported Outcomes.” The chapter describes the important role patient outcomes play in clinical research and the fundamentals of measurement theory and well-established techniques for valid and reliable collection of data regarding patient experiences.

Finally, and also related to patients, is a chapter on patient registries, provided by Rachel Richesson, Leon Rozenblit, Kendra Vehik, and Jimmy Tcheng. Their discussion includes the scientific and technical issues for registries and highlights challenges for standardizing the data collected. In a new chapter on governance, Anthony Solomonides and Katharine Fultz Hollis describe organizational structures and processes that can be used to ensure data quality and patient and institutional protections.

Section 3: Knowledge Representation and Data-Driven Discovery

The premise of clinical research informatics is that the collection (and best representation and availability) of data – and techniques for aggregating and sharing data with existing knowledge – can support discovery of new knowledge leading to scientific breakthroughs. The chapters that comprise this section are focused on state-of-the-art approaches to organizing or representing knowledge for retrieval purposes or use of advanced technologies to discover new knowledge and information where structured representation is not present or possible. While these topics apply across informatics and its subdisciplines, they stand to have a profound influence on CRI, which is inherently (unlike other subdisciplines) focused on data analysis. The ability to use, assimilate, and synergize new data with existent knowledge could potentially identify new relationships that in turn lead to new hypotheses related to causation of disease or potential therapies and biological interactions. Also, the ability to combine and enhance new and old knowledge has a major role in improving safety, speeding discovery, and supporting translational science. Since all new research builds upon what has come before, the ability to access and assimilate current research will accelerate new research.

There is a natural appeal to ideas for transforming and exchanging heterogeneous data, which can be advanced using ontologies (or formal conceptual semantic representations of a domain). Kin Wah Fung and Olivier Bodenreider give us an overview of basic principles and challenges, all tied to examples of use of ontology in the clinical research space. This chapter covers the challenges related to knowledge representation in clinical research and how trends and issues in ontology design, use, and testing can support interoperability. Essential definitions are covered, as well as applications and other resources for development such as the semantic web. Additionally, major relevant efforts toward knowledge representation are reviewed. Specific ontologies relevant to clinical research are discussed, including the ontology for clinical trials and the ontology of biomedical investigation. Organizations, such as the National Center for Biomedical Ontology, that coordinate development, access, and organization of ontologies are discussed. Next, Mollie Cummins' chapter offers an overview of state-of-the-art data mining and knowledge discovery methods and tools as they apply to clinical research data. The vast amount of data warehoused across various clinical research enterprises, and the increasing desire to explore these to identify unforeseen patterns, require such advanced techniques. Examples of how nonhypothesis-driven research supported by advanced data mining, knowledge discovery algorithms, and statistical methods help elucidate the need for these tools to support clinical and translational research.

Last in this section, Feifan Liu, Chunhua Weng, and Hong Yu explain the use of data from electronic healthcare record (EHR) systems to support research activities. This is an area which continues to gain attention since EHRs are widely used and represent *real-life* disease and health-care experiences that are potentially more generalizable than are the results from controlled clinical studies. However, at the current time, much of the important information in EHRs is still narrative in nature. This chapter describes how natural language processing (NLP) techniques can be

used to retrieve and utilize patient information from EHRs to support important clinical research activities.

In this final section of the text, we also include topics that will continue to impact CRI into the future and that build upon the contexts, data sources, and information and knowledge management issues discussed in previous sections. Many of the topics included here are truly multidisciplinary and stand to potentially impact all clinical research studies.

The use of clinical data for research is a tremendous challenge with perhaps the greatest potential for impact in all areas of clinical research. Standards specifications for the use of clinical data to populate research forms have evolved to support a number of very promising demonstrations of the “collect once, use many” paradigm. Rebecca Kush and Amy Nordo cover various scenarios for data sharing, including who needs to share data and why. More importantly, they describe the history and future strategy of cooperation between major standards development organizations in health care and clinical research.

Rachel Richesson, Cecil Lynch, and W. Ed Hammond cover the topic of standards – a central topic and persistent challenge for informatics efforts. Their focus is on the standards development process and relevant standards developing organizations, including the Clinical Data Interchange Standards Consortium (CDISC). They address the collaboration and harmonization between research data standards and clinical care data standards.

Pharmacovigilance is an emerging area that stands to impact the future of CRI, particularly given its relevance to patient safety and potential to impact population health. Informatics methods and applications are needed to ensure drug safety for patients and the ability to access, analyze, and interpret distributed clinical data across the globe to identify adverse drug events. Michael Ibara provides a historical account of its evolution, as well as the increasing need for informatics methods and applications that can be employed to ensure greater patient safety. Various issues are explored in this context, including drug and device safety monitoring, emerging infrastructures for detecting adverse drug events, and advanced database and information sharing approaches.

The full transparency of clinical research is a powerful strategy to diminish publication bias, increase accountability, avoid unnecessary duplication of research, advance research more efficiently, provide more reliable evidence (information) for diagnostic and therapeutic prescriptions, and regain public trust. Trial registration and results disclosure are considered powerful tools for achieving higher levels of transparency and accountability for clinical trials. New emphasis on knowledge sharing and growing demands for transparency in clinical research are contributing to a major paradigm shift in health research that is well underway. This chapter by Karmela Krleža-Jerić discusses the use of trial registries and results databases in clinical research and decision making. International standards of trial registration and their impact are discussed, as are the contribution of informatics experts to these efforts.

The book concludes with a brief chapter by Peter Embi summarizing the challenges CRI researchers and practitioners will continue to face as the field evolves

and new challenges arise. This concluding chapter helps in envisioning the future of the domain of clinical research informatics. In addition to outlining likely new settings and trends in research conduct and funding, the author cogitates on the future of the informatics infrastructure and the professional workforce training and education needs. A focus of this chapter is the description of how clinical research (and supporting informatics) fits into a bigger vision of a learning health systems and of the relationship between clinical research, evidence-based medicine, evidence-generating medicine, and quality of care.

Conclusion

The overall goal of this book is to contribute to the ongoing discourse among researchers and practitioners in CRI as they continue to rise to the challenges of a dynamic and evolving clinical research environment. This is an exciting and quite broad domain, and there is ample room for future additions or other texts exploring these topics more deeply or comprehensively. Most certainly, the development of CRI as a subdiscipline of informatics and a professional practice area will drive a growing pool of scientific literature based on original CRI research, and high-impact tools and systems will be developed. It is also certain that CRI groups will continue to support and create communities of discourse that will address much needed practice standards in CRI, data standards in clinical research, policy issues, educational standards, and instructional resources.

The scholars that have contributed to this book are among the most active and engaged in the CRI domain, and we feel they have provided an excellent starting point for deeper explorations into this emerging discipline. While we have by no means exhausted the range of topics, we hope that readers will see certain themes stand out throughout this text. These include the changing role of the consumer, movement toward transparency, growing needs for global coordination and cooperation on many levels, and the merging together of clinical care delivery and research as part of a changing paradigm in global health-care delivery – all in the context of rapid innovations in technology and explosions of data sources, types, and volume. These forces collectively are the challenges to CRI, but they also show promise for phenomenal synergy to yield unimaginable advances in scientific knowledge, medical understanding, the prevention and cure of diseases, and the promotion of health that can change the lives of all. The use of informatics and computing can accelerate and guide the course of human and global evolution in ways we cannot even predict.

References

1. Mayer D. A brief history of medicine and statistics. In: Essential evidence-based medicine. Cambridge: Cambridge University Press; 2004. p. 1–8.
2. Atkins HJ. The three pillars of clinical research. Br Med J. 1958;2(5112):1547–53.
3. Bacchieri A, Della Cioppa G. Fundamentals of clinical research: bridging medicine, statistics and operations, Statistics for biology and health. Milan: Springer; 2007.

4. Friedman CP. A “fundamental theorem” of biomedical informatics. *J Am Med Inform Assoc.* 2009;16(2):169–70.
5. NIH. The NIH Director’s panel on clinical research report to the advisory committee to the NIH director, December, 1997. 1997. [cited 2011 May 15]. Available from: http://www.oenb.at/de/img/executive_summary%2D%2Dnih_directors_panel_on_clinical_research_report_12_97_tcm14-48582.pdf.
6. Embi PJ, Payne PR. Clinical research informatics: challenges, opportunities and definition for an emerging domain. *J Am Med Inform Assoc.* 2009;16(3):316–27.
7. Sanchez-Pinto L, Mosa ASM, Fultz-Hollis K, Tachinardi U, Barnett WK, Embi PJ. The emerging role of the chief research informatics officer in academic health centers. *Appl Clin Inform.* 2017;8(3):845–53.
8. Forrest WH, Bellville JW. The use of computers in clinical trials. *Br J Anaesth.* 1967;39:311.
9. Pryor DB, et al. Features of TMR for a successful clinical and research database. In: Proceedings of the Sixth Annual Symposium on Computer Applications in Medical Care. Blum BI (ed). New York: IEEE, 1982.
10. Blois MS. Medical records and clinical databases: what is the difference? *MD Comput.* 1984;1(3):24–8.
11. Breitfeld PP, et al. Pilot study of a point-of-use decision support tool for cancer clinical trials eligibility. *J Am Med Inform Assoc.* 1999;6(6):466–77.
12. Carlson R, et al. Computer-based screening of patients with HIV/AIDS for clinical-trial eligibility. *Online J Curr Clin Trials.* 1995.
13. Mansour EG. Barriers to clinical trials. Part III: knowledge and attitudes of health care providers. *Cancer.* 1994;74(9 Suppl):2672–5.
14. ICH. Guideline for good clinical practice E6(R1). International Conference on Harmonisation. 1996.
15. Shortliffe EH, Climo JJ, Hannah KJ, editors. Biomedical informatics: computer applications in health care and biomedicine health informatics. New York: Springer Science+Business Media, LLC; 2006.
16. Hoyt R, Hersh WR, editors. Health informatics: practical guide. 7th ed. North Carolina: Morrisville, Lulu.com, 2018. 486 p.
17. Embi P. AMIA CRI years in review. 2018 [cited 2018 July 1]. Available from: <http://www.embi.net/cri-years-in-review.html>.



From Notations to Data: The Digital Transformation of Clinical Research

2

Christopher G. Chute

Abstract

The history of clinical research precedes the advent of computing, though informatics concepts have long played important roles. The advent of digital signal processing in physiologic measurements tightened the coupling to computation for clinical research. The astronomical growth of computational capacity over the past 60 years has contributed to the scope and intensity of clinical analytics, for research and practice. Correspondingly, this rise in computation power has made possible clinical protocol designs and analytic strategy that were previously infeasible. The factors have driven biological science and clinical research into the big science era, replete with a corresponding increase of intertwined data resources, knowledge, and reasoning capacity. These changes usher in a social transformation of clinical research and highlight the importance of comparable and consistent data enabled by modern health information data standards and ontologies.

Keywords

History of clinical research · Digitalization of biomedical data · Information-intensive domain · Complexity of clinical research informatics · Computing capacity and information processing · Interoperable information · Complexity of design protocol

C. G. Chute, MD, DrPH (✉)

Schools of Medicine, Public Health, and Nursing, Johns Hopkins University,
Baltimore, MD, USA
e-mail: chute@jhu.edu

Historical Perspective

The history of clinical research, in the broadest sense of the term, is long and distinguished. From the pioneering work of William Harvey to the modern modalities of translational research, a common thread has been the collection and interpretation of information. Thus, informatics has played a prominent role, if not always recognized as such. Accepting that an allowable definition of informatics is the processing and interpretation of information that permits analyses or inferencing, the science of informatics can and does predate the advent of modern computing.

Informatics has always been a multidisciplinary science, blending information science with biology and medicine. Reasonable people may inquire whether distinguishing such a hybrid as a science is needed, though this is reminiscent of parallel debates about epidemiology, which to some had merely coordinated clinical medicine with biostatistics; few question the legitimacy of epidemiology as a distinct discipline today (nor biostatistics if I were to nest this discussion yet further). Similarly, in the past two decades, informatics, including clinical research informatics as a recognized subfield, has come into its own.

Nevertheless, common understanding and this present text align informatics, applied to clinical research or otherwise, with the use of digital computers. So when did the application of digital computers overlap clinical research? This question centers on one's notion about the boundaries of clinical research, perhaps more a cultural issue than amenable to rational debate. For the purposes of this discussion, I will embrace the spectrum from physiological measurements to observational data on populations within the sphere of clinical research.

Analog Signal Processing

In its simplest form, the use of an analog measurement can be seen in the measurement of distance with a ruler. While not striking most as a predecessor of clinical informatics, it does illustrate the generation of quantitative data. It is the emphasis on the quantification of data that distinguishes ancient from modern perspectives on biomedical research.

The introduction of signal transducers, which enabled the transformation of a myriad of observations ranging from light, pressure, velocity, temperature, or motion into electronic signals, such as voltage strength, demarcated the transition from ancient to modern science. This represents yet another social transformation attributable to the harnessing of electricity. Those of us old enough to remember the ubiquitous analog chart recorder, which enabled any arbitrary voltage input to be continuously graphed over time, recognize the significant power that signal transduction engendered.

The ability to have quantified units of physiologic signals, replete with their time-dependent transformations as represented on a paper graph, enabled the computation, albeit by analog methods, of many complex parameters now taken

for granted. These include acceleration constants, maximum or minimum measures, inflection points, a host of continuous data properties, and most importantly an ability to observe and quantitate covariance between and among such measures. These in turn enabled the creation of mathematical models that could be inferred, tested, validated, and disseminated on the basis of continuous quantitative data.

Departments of physiology and biomedical research saw a huge progress in the evolution and sophistication of physiologic models arising from increasing quantities of continuous quantitative data over time. Early work invoking signal transduction and quantified analog signals could be found in the 1920s but became much more common in the 1930s and was a standard method in the 1940s and 1950s. This introduced unprecedented precision, accuracy, and reproducibility in biomedical research.

The novel capability of complex quantitative data capture, analysis, and utilization presaged the next great leap in clinical informatics: the digitalization of data.

Digital Signal Processing

The advent of digital signal processing (DSP), first manifested in analog to digital converters (ADCs), has fundamentally transformed clinical research. In effect, it is the marrying of quantitative data to computing capability. ADCs take analog input, most typically a continuous voltage signal, and transform it into a digital number. Typically, the continuous signal is transformed into a series of numbers, with a specific time interval between the generations of digital “snapshots.” The opposite twin of ADCs are digital to analog converters (DACs), which can make digital data “move the needle” proverbially.

DSPs were first practically used during the Second World War, when they were experimented to carry telephonic signals over long distances without degradation by putting ADCs and DACs in series. The telephony industry brought this capability into the civilian world, and commercial DSP began to appear in the 1950s. At that time, the numerical precision was crude, ranging from 4 to 8 bits. Similarly, the frequency of digital number generation was relatively slow, on the order of one number per second.

The appearance of transistors in the 1960s, and integrated circuits in the 1970s, ushered in a period of cheap, reliable, and relatively fast DSP. While case reports exist of physiologic researchers using DACs in the 1950s, this did not become a common practice until the cost and performance characteristics of this technology became practical in the early 1970s. Today, virtually all modern smartphones have highly sophisticated DSP capabilities, some of which is starting to be used for remote physiological monitoring of clinical research participants and the general public through fitness apps.

The Digitalization of Biomedical Data

The early 1970s was also coincident with the availability of affordable computing machinery for routine analysis to the same biomedical research community. Because DSP is the perfect partner for modern digital computing, supporting moderately high-bandwidth data collection from a myriad of information sources and signals, they enabled a practical linkage of midscale experimental data to computing storage and analysis in an unprecedented way. Prior to that time, any analysis of biomedical data would require key entry, typically by hand. Again, many of us can recall rooms of punch card data sets, generated by tedious key-punch machinery.

While it is obviously true that not all biomedical data or clinical informatics arose from transducer-driven DSP signals, the critical mass of biomedical data generated through digitalization of transducer-generated data culturally transformed the expectation for data analysis. Prior to that time, small data tables and hand computations would be publishable information. The advent of moderate-volume data sets, coupled with sophisticated analytics, raised the bar for all modalities of biomedical research. With the advent of moderate-volume data sets, sophisticated computing analytics, and model-driven theories about biomedical phenomenon, the true birth of clinical research informatics began.

Dimensions of Complexity

Informatics, by its nature, implies the role of computing. Clinical research informatics simply implies the application of computational methods to the broad domain of clinical research. With the advent of modern digital computing, and the powerful data collection, storage, and analysis that this makes possible, inevitably comes complexity. In the domain of clinical research, I assert that this complexity has axes, or dimensions, that we can consider independently. Regardless, the existence and extent of these complexities have made inexorable the relationship between modern clinical research, computing, and the requirement for sophisticated and domain-appropriate informatics.

Computing Capacity and Information Processing

Biomedical research and, as a consequence, clinical research informatics are by their nature within a profoundly information-intensive domain. Thus, any ability to substantially increase our capacity to process or manage information will significantly impact that domain. The key-enabling technology of all that has been described in clinical research informatics is the advent of ever-increasing computational capabilities. This has been widely written about, but I submit its review is germane to this introduction. I will frame these advances in four dimensions: computational power, network capacity, local memory, and data storage.

Computational Power

The prediction of Gordon Moore in 1965 that integrated circuit density would double every 2 years is well known. Given increasing transistor capabilities, a corollary of this is that computing performance would double every 18 months. Regardless of the variation, the law has proved uncannily accurate. As a consequence, there has been roughly a ten trillion-fold increase in computing power over the last 60 years. The applications are striking; the supercomputing resources that national spies would kill each other to secure 20 years ago now end up under Christmas trees as game platforms for children. The advent of highly scalable graphical processing units (GPU) has correspondingly transformed our capacity to feasibly address many problems previously beyond practical limits.

Network Capacity

Early computing devices were reliant on locally connected devices for input and output. The most primitive interface devices were plugboard and toggle switches that required human configuration; the baud rates of such devices are perhaps unimaginably slow. Today, Tb network backbones are not uncommon, giving yet nearly another trillion-fold increase in computational capabilities.

Local Storage

Early computers used electromechanical relays later replaced by speedy vacuum tubes. The advent of the transistor, and subsequently the integrated circuit, enabled the dramatic reduction in space with an increase in density for local storage. It is clear that at least a trillion-fold increase in common local storage capability in terms of speed and size has been achieved.

Data Storage

The advent of high-density, high-performance disk drives, compared to early paper tape or punch card, yields perhaps the most dramatic increase in data processing capability and capacity. Petabyte drive complexes are not uncommon, and with the advent of cloud storage, there is no practical upper limit. For the purposes of this exercise, and to make a relatively round number, we can assert a 10^{15} increase in data storage capacity.

Taken together, these advances total an approximate 10^{60} increase in computational power (albeit we are cheating somewhat adding exponents, which is really multiplying in non-logarithmic space) over the past 60 years. Regardless, there has been an astronomical increase in our ability and capacity to manage, process, and inference about data and information. In an information-intensive industry such as clinical research, the consequences cannot be other than profound.

Data Density

The most obvious dimension of data complexity is its sheer volume. Historically, researchers would content themselves with a data collection sheet that might have been enumeration of subjects or objects of study and at most a handful of variables. The advent of repeated measures, metadata, or complex data objects was far in the future, as were data sets that evolved from the scores to the thousands.

Today, it is not uncommon in any domain of biomedical research to find vast, rich, and complex data structures. In the domain of genomics, this is most obvious with not only sequencing data for the genome but also the associated annotations, haplotype, pathway data, and sundry variants with clinical or physiological import, as important attributes. The advent of whole genome sequences (WGS) increases volume and complexity, while the application of WGC to discrete cells within tumors further raises the bar.

This complexity is not unique to genomic data. Previously humble clinical trial data sets now have highly complex structures and can involve vectors of laboratory data objects each with associated normal ranges, testing conditions, and important modes of conclusion-changing metadata. Similarly, population-based observational studies may now have large volumes of detailed clinical information derived from electronic health records.

The historical model of relying on human-extracted or entered data is long past for most biomedical investigators. High data volumes and the asserted relationships among data elements comprise information artifacts that can only be managed by modern computing and informatics methods.

Design Complexity

Commensurate with the complexity of data structure and high volume is the nature of experimental design and methodology. Today, ten-way cross-fold validation, bootstrapping techniques for various estimates, exhaustive Monte Carlo simulation, and sophisticated experimental nesting, blocking, and within-group randomization afford unprecedented complexity in the design, specification, and execution of modern-day protocols.

Thus, protocol design options have become inexorably intertwined with analytic capabilities. What was previously inconceivable from a computational perspective is now a routine. Examples of this include dynamic censoring, multiphase crossover interventions, or imputed values.

Analytic Sophistication

Paralleling the complexity of design is the sophistication of analysis. As implied in the previous section, it is difficult to say which is causal; no doubt analytic capabilities push design, as design innovations require novel analytic modalities.

The elegant progression from simple parameter estimation, such as mean and variance, to linear regressions, to complex parametric models, such as multifactorial Poisson regression, to sophisticated and nearly inscrutable machine learning techniques such as multimodal neural networks or deep learning, demonstrates exponentially more intensive numerical methods demanding corresponding computational capacity. Orthogonal to such computational virtuosity is the iterative learning process now routinely employed in complex data analysis. It is rare that a complete analytic plan will be anticipated and executed unchanged for a complex protocol. Now, preliminary analysis, model refinement, parameter fitting, and discovery of confounding or effect modification are routinely part of the full analysis process. The computational implications of such repeated, iterative, and computationally complex activities are entirely enabled by the availability of modern computing. In the absence of this transformative resource, and the commensurate informatics skills, modern data analysis and design would not be possible.

The Emergence of Big Science

What then are the consequences of unprecedented computational capabilities in an information-intensive enterprise such as clinical research? It is useful to examine where this or similar activities have occurred previously. An evolutionary change for many disciplines is a transition from an exclusively independent-investigator-driven suite of agendas across a discipline (small-science or bottom-up foci) to a maturation where interdependency of data and methods, multidisciplinary teams of talent and interest, and large-scale, cross-discipline shared resources, such as massive machines or databases, predominate (big-science or top-down coordination).

Evolution of Astronomy and Physics

The practice of modern astronomy relies upon large groups, large data sets, and strong collaboration between and among investigators. The detection of a supernova in a distant galaxy effectively requires a comparison of current images against historical images and excluding any likely wandering objects, such as comets. Similarly, the detection of a pulsar requires exhaustive computational analysis of very large radio telescope data sets. In either case, the world has come a long way from the time when a single man with a handheld telescope, in the style of Galileo, could make seminal astronomical discoveries.

In parallel, the world of high particle physics has become a big science given its requirements for large particle accelerators, massive data-collection instrumentation, and vast computational power to interpret arcane data. Such projects and initiatives demand large teams, interoperable data, and collaborative protocols. The era of tabletop experiments, in the style of Rutherford, has long been left behind.

What is common about astronomy and physics is their widely recognized status as big-science enterprises. A young investigator in those communities would not

even imagine or attempt to make a significant contribution outside the community and infrastructure that these fields have established, in part due to the resource requirements, but equivalently because of the now-obvious multidisciplinary nature of the field.

Biology and Medicine as a Socially Interdependent Process

I return to the assertion that biology and medicine have become information-intensive domains. Progress and new discovery are integrally dependent on high-volume and complex data. Modern biology is replete with the creation of and dependency on large annotated data sets, such as the fundamental GenBank and its derivatives, or the richly curated animal model databases. Similarly, the annotations within and among these data sets constitute a primary knowledge source, transcending in detail and substance the historically quaint model of textbooks or even the prose content in peer-reviewed journals.

The execution of modern studies, relying as it does on multidisciplinary talent, specialized skills, and cross-integration of resources, has become a complex social process. The nature of the social process at present is still a hybrid across bottom-up, investigator-initiated research and team-based, program project-oriented collaborations.

The Social Transformation of Clinical Research

The conclusion that biology and medicine, and as a consequence clinical research informatics, are evolving into a big-science paradigm is unavoidable. While this may engender an emotional response, the more rational approach is to understand how we as a clinical research informatics community can succeed in this socially transformed enterprise. Given the multidisciplinary nature of informatics, the clinical research informatics community is well poised to contribute importantly in the success of this transformed domain.

A consequence of such a social transformation is the role of government or large foundations in shaping the agenda of the cross-disciplinary field. One role of government, in science or any other domain, is to foster the long-term strategic view and investments that cannot be sustained in the private marketplace or the agendas of independent investigators. Further, it can encourage and support the coordination of multidisciplinary participation that might not otherwise emerge. In the clinical trials world, the emergence of modest but influential forces such as [ClinicalTrials.gov](#) illustrates this role.

Standards

If biology and medicine, and by association clinical research informatics, are entering a big-science paradigm, what does this demand as an informatics infrastructure?

Comparable and Consistent Information

Given the information-intensive nature of clinical research informatics, the underlying principle for big science is the comparability and consistency of data. Inferencing across noncomparable information, by definition, cannot be done. Anticipating or accounting for inconsistent data representations is inefficient and non-scalable. The obvious conclusion is that within biology and medicine, a tangible contribution of clinical informatics is to ensure that genomic, clinical, and experimental data conform to frameworks, vocabularies, and specifications that can sustain interoperability. This also raises the profile and critical requirement for robust ontologies to mediate data and knowledge integration. Emergent projects that demonstrate large-scale inferencing and reasoning, using ontological annotations of basic science and clinical data, are beginning to bridge the historical “chasm of semantic despair” that inhibited rapid translation of basic science knowledge and information into clinical care.

Interoperable Systems and Constructs

The hallmark of big science, then, is interoperable information. The core of interoperable information is the availability and adoption of standards. Such standards can and must specify data relationships, content, vocabulary, and context. As we move into this next century, the great challenge for biology and medicine is the definition and adoption of coherent information standards for the substrate of our research practice.

The present volume outlines many issues that relate to data representation, inferencing, and standards – issues that are crucial for the emergence of large-scale science in clinical research. Readers must recognize that they can contribute importantly through the clinical research informatics community to what remains an underspecified and as yet immature discipline. Yet there is already tremendous excitement and interest at the intersection between basic science and clinical practice, manifested by translational research, that has well-recognized dependencies on clinical research informatics. I trust that the present work will inspire and guide readers to consider and hopefully undertake intellectual contributions toward this great challenge.



The Clinical Research Environment

3

Philip R. O. Payne

Abstract

The conduct of clinical research is a data- and information-intensive endeavor, involving a variety of stakeholders spanning a spectrum from patients to providers to private sector entities to governmental policymakers. Increasingly, the modern clinical research environment relies on the use of informatics tools and methods, in order to address such diverse and challenging needs. In this chapter, we introduce the major stakeholders, activities, and use cases for informatics tools and methods that characterize the clinical research environment. This includes an overview of the ways in which informatics-based approaches influence the design of clinical studies, ensuing clinical research workflow, and the dissemination of evidence and knowledge generated during such activities. Throughout this review, we will provide a number of exemplary linkages to core biomedical informatics challenges and opportunities and the foundational theories and frameworks underlying such issues. Finally, this chapter places the preceding review in the context of a number of national-scale initiatives that seek to address such needs and requirements while advancing the frontiers of discovery science and precision medicine.

Keywords

Clinical research funding · Clinical research design · Clinical research workflow · Clinical research data management · Data sharing · Discovery science · Precision medicine

P. R. O. Payne, PhD (✉)

Institute for Informatics (I2), Washington University in St. Louis School of Medicine,
St. Louis, MO, USA

e-mail: prpayne@wustl.edu

Overview

We describe here the clinical research environment, including an overview of common activities and processes, as well as the roles played by various stakeholders involved throughout the life cycle of clinical studies, including both interventional and observational study designs. This discussion summarizes data and information management requirements incumbent to the clinical research domain. This chapter concludes with a review of the state of knowledge concerning clinical research workflow and communication patterns as well as prevailing trends in clinical research funding and the evolving range of settings in which clinical research is taking place. In addition, the chapter includes an introduction to the relationship between clinical research and the pursuit of both discovery science and precision medicine paradigms.

This chapter is organized into three general sections describing:

1. The basic processes, actors, settings, and goals that serve to characterize the physical and sociotechnical clinical research environment.
 2. A framework of clinical research data and information management needs.
 3. The current understanding of the evolving body of research that seeks to characterize clinical research workflow and communications patterns. This understanding can be used to support the optimal design and implementation of informatics platforms for use in the clinical research environment.
-

Clinical Research Processes, Actors, and Goals

In the following section, we introduce the major processes, stakeholders, and goals that serve to characterize the modern clinical research environment. Taken as a whole, these components represent a complex, data- and information-intensive enterprise that involves the collaboration of numerous professionals and participants in order to satisfy a set of tightly interrelated goals and objectives. Given this complex environment and the role of informatics theories and methods in terms of addressing potential barriers to the efficient, effective, high-quality, and timely conduct of clinical research, this remains an area of intensive research interest for the biomedical informatics community [1–6].

Common Clinical Research Processes

At a high level, the processes and activities of the life cycle of a clinical research program can be divided into eight general classes, as summarized below. Of note, we will place particular emphasis in this section on describing those processes relative to the conduct of interventional clinical studies (e.g., studies where a novel treatment strategy is being evaluated for safety, efficacy, and comparative effectiveness if an alternative treatment strategy exists). However, similar processes generally apply to observational or retrospective studies, with the exception of processes

related to the tracking and execution of study-related participant encounters and interventions. An example of this clinical research life cycle, its major phases, and constituent processes and activities, relative to the context of an interventional clinical trial, is illustrated in Fig. 3.1. Key processes and activities include the following.

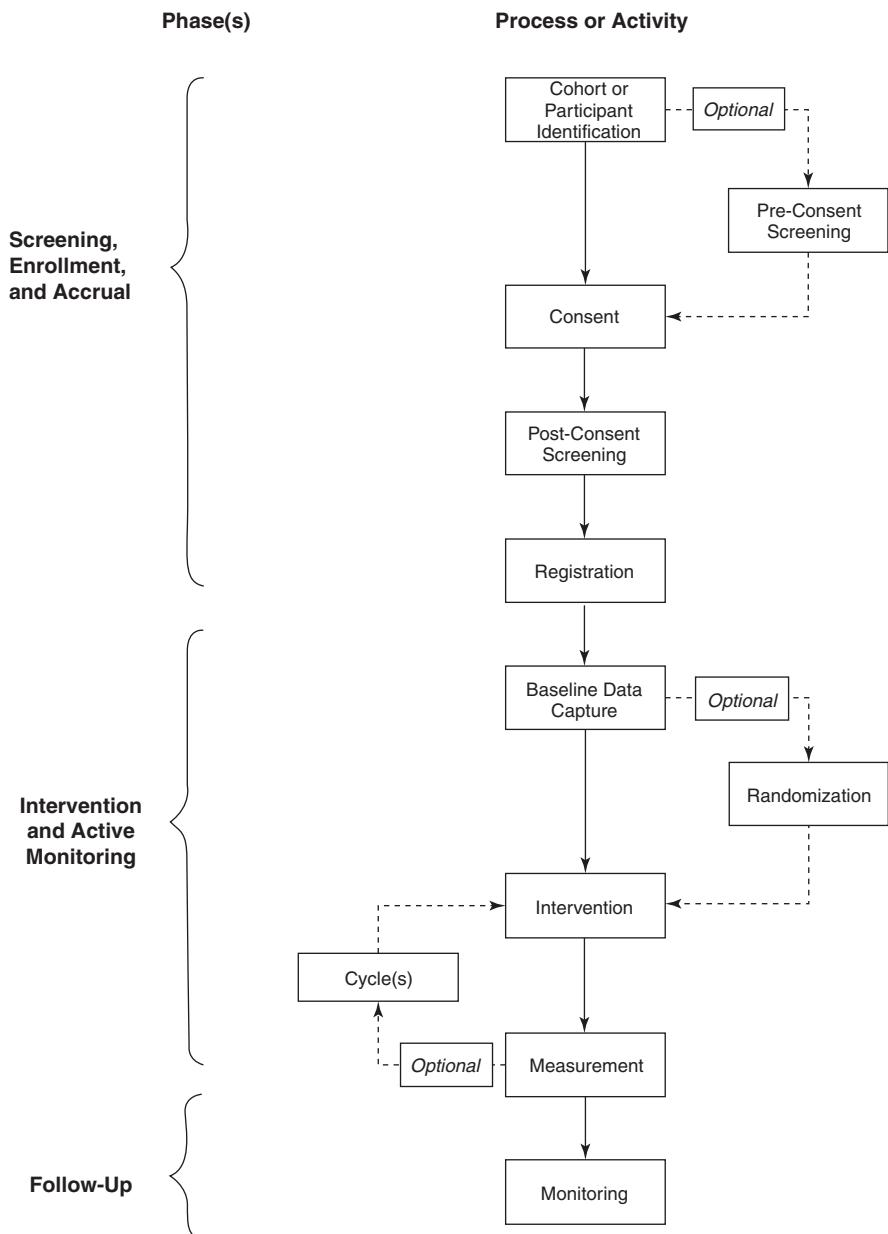


Fig. 3.1 Interventional clinical trial phases and associated execution-oriented processes

Identifying Potential Study Participants

This process usually involves either (1) the pre-encounter and/or point-of-care review of an individual's demographics and clinical phenotype in order to determine if they are potentially eligible for a given research study, given a prescribed set of eligibility criteria concerned with those same variables (also referred to as inclusion and exclusion criteria), or (2) the identification of a cohort of potential study participants from whom data can be derived, via a retrospective review of available data sources in the context of a set of defining parameters. In many cases, the data elements required for such activities are either incomplete or exist in unstructured formats, thus complicating such activities. This usually makes it necessary for potential participants to be identified via automated methods that provide a partial answer as to whether an individual is or is not eligible for a trial, which is then further explored via screening activities such as physical examinations, interviews, medical record reviews, or other similar labor-intensive mechanisms (see section "[Screening and Enrolling Participants in a Clinical Study](#)" for more details). Due to prevailing confidentiality and privacy laws and regulations, if the individual performing such eligibility screening is not directly involved in the clinical care of a potential study participant and eligibility is determined through secondary use of primarily clinical data, then the individual performing such screening must work in coordination with an individual who is involved in such clinical care in order to appropriately communicate that information to a potential study participant.

Screening and Enrolling Participants in a Clinical Study

Once a potential participant is identified, they are often subjected to evaluation, as introduced above, in order to satisfy all applicable study eligibility criteria. If they do so successfully, the participant is "enrolled" or "registered" in a study (note that both of these activities depend upon a documented informed consent process or equivalent mechanism for human subjects protection). During this process, it is common for a study-specific enrollment identifier to be assigned to the participant. Of note, study staff usually maintain a set of records (often known as a "screening log") that summarize numbers of potential participants who were identified via such screening processes and how many of those individuals were successfully enrolled in a given study. Such screening logs may also include de-identified or abstracted data that details the reasons why some individuals were not successfully enrolled in a study, which can be used to help inform the pursuit of recruitment efforts for the investigation in process as well as the design of future studies.

Scheduling and Tracking Study-Related Participant Events

Once participants have been identified, screened, and enrolled in a study, they are usually scheduled for a series of encounters as defined by a study-specific calendar of events, which is also referred to as the study protocol. Sometimes, the scheduling

of such events is sufficiently flexible (allowing for windows of time within which a given task or event is required to take place) that individuals may voluntarily adjust or modify their study calendar. In other cases, the temporal windows between study-related tasks or events are very strict and therefore require strict adherence by investigators and participants to the requirements defined by said calendars. Such participant- and study-specific calendars of events are tracked at multiple levels of granularity (e.g., from individual participants to large cohorts of participants enrolled in multiple studies) in order to detect individuals or studies that are “off schedule” (e.g., late or otherwise noncompliant with the required study events or activities specified in the research protocol).

Executing Study Encounters and Associated Data Collection Tasks

For each task or activity specified in a study protocol, there is almost always a corresponding study encounter (e.g., visit or phone call), during which the required study activities will be executed and the resulting data collected using either paper forms (i.e., case report forms or CRFs) or electronic data capture (EDC) instruments that replicate such CRFs in a computable format. While EDC tools are preferable for a number of reasons (e.g., quality, completeness, and auditability of data capture and management, as well as maintaining the security and confidentiality of study data) and access to computational resources has become commonplace in many study environments, there still remain large numbers of studies that are conducted using paper CRFs.

Ensuring the Quality of Study Data

Throughout a given study, study investigators and staff will usually engage in a continuous cycle of reviewing and checking the quality of study-related data. Such quality assurance (QA) usually includes reconciling the contents of CRFs or EDC instruments with the contents of supporting source documentation (e.g., electronic health records or other legally binding record-keeping instruments). It is common for such QA checks to be triggered via automated or semiautomated reports or “queries” regarding inconsistent or incomplete data that are generated by the study sponsor or other responsible regulatory bodies (a more thorough characterization of data quality and quality assurance activities specific to clinical research is presented in Chap. 10).

Regulatory and Sponsor Reporting and Administrative Tracking/Compliance

Throughout the course of a study, there are often prescribed reports concerning study enrollment, data capture, and trends in study-generated data that must be submitted to regulatory agencies, study-specific and/or institutional monitoring bodies, and/or the study sponsor. As was the case with study-encounter-related data capture,

such reports can be submitted on paper or electronically. In addition, for studies regulated by government agencies (such as the FDA) or local institutional review boards (IRBs), further study-related reporting requirements must be tracked and complied with, often using proprietary or locally developed reporting instruments or tools. A primary example of such tracking/compliance is the preparation, submission, and approval of institutional review board (IRB) protocols that define how participants will be recruited and enrolled in studies and subsequently how data will be collected from them and how any physical or other risks (such as those related to security and confidentiality) are to be identified, reported, and mitigated. Additional activities included in this particular class of processes include seeking and retrieving information related to study protocols and any changes (or amendments) made to those documents throughout the course of their execution.

Budgeting and Fiscal Reconciliation

At the outset of a study, throughout its execution, and after its completion, an ongoing process of budgeting and fiscal reconciliation is conducted. The goal of these processes is to ensure the fiscal stability and performance of the study, thus making it possible to maintain necessary overhead and support structures in what is ideally a revenue or cost neutral manner.

Human Subjects Protection Reporting and Monitoring

As mentioned previously, compliance with human subject-related reporting and the monitoring of such compliance are a central part of the conduct of clinical research. This type of compliance can include obtaining IRB or equivalent approval for a study protocol and its associated practices and the execution of informed consent (a process by which potential participants are informed of the nature of a study, its risks, and benefits, in a way that allows them to weigh such factors before voluntarily engaging in a study). In addition, suspected adverse events must be collected and reported periodically to the institutional, sponsor, and regulatory organizations. The definition of “reportable” adverse events can vary by protocol, sponsor, and institution and can include local events (called internal AEs) and those occurring at other research sites (called external AEs). Similarly, actions taken in response to an AE (e.g., an amendment to a protocol reflecting changes or elimination of study procedures, adding new risks to informed consent documents) must be communicated, documented, and tracked for compliance.

Common Tasks and Barriers to Successful Study Completion

According to several recent studies concerned with clinical research workflow and the tasks executed by investigators and study staff, the most common tasks performed by those individuals relative to the preceding activity areas include

[2, 7–11] (1) completing paper or electronic case report forms; (2) seeking source documentation to validate the contents of such case report forms; (3) identifying, screening, and registering new study participants; and (4) responding to various reporting and monitoring requirements. In an analogous group of studies, the most common barriers encountered by investigator and study staff to the successful completion of clinical research program include [3, 10, 12, 13] (1) an inability to identify and recruit a sufficient number of study participants; (2) the attrition of participants in a study due to non-compliance with the study calendar or protocol; and (3) missing, incomplete, or insufficient high-quality data being collected such that planned study analyses cannot be performed using such data.

Clinical Research Stakeholders

As was noted previously, the clinical research environment involves the collaboration of a broad variety of stakeholders fulfilling multiple roles. Such stakeholders can be classified into six major categories, which apply across a spectrum from community practice sites to private sector sponsors to academic health centers (AHCs) and ultimately to governmental and other regulatory bodies. In the following discussion, we will briefly review the roles and activities of such actors, relative to the following six categories [3, 9, 14, 15]. *It is important to note that much of the data and information intensity of modern clinical research is a function of the need for these diverse stakeholders to interact and coordinate their activities in near real time, often in settings that span organizational, geographic, and temporal boundaries.*

Patients and Advocacy Organizations

The first and perhaps most important stakeholder in the clinical research domain is the patient, also known as a study participant, and as an extension, advocacy organizations focusing upon specific disease or health states. Study participants are the individuals who either (1) receive a study intervention or therapy or (2) from whom study-related data are collected. Participants most often engage in studies due to a combination of factors, including:

- The availability of novel therapies as a result of participation, which may provide better clinical or quality of life outcomes and that are not available via standard-of-care models
- The exhaustion of standard-of-care options for a given disease state, thus leaving interventional clinical studies as the only viable treatment modality
- A desire to support the advancement of the understanding of a specific uncharacterized or *under*-characterized disease or condition via an observational or natural history study or the advancement of understanding of biological processes, life sciences more generally, or public health

Unfortunately, identifying participants who are motivated by one or more of the preceding factors and that meet appropriate demographic or clinical criteria for enrollment in a study (e.g., eligibility or inclusion/exclusion criteria) is a difficult task. In fact, in a recent report, it was found that less than 4% of the adult US population who could have participated in a clinical research study actually did so. Such low participation is a significant impediment to our collective ability to advance the state of human health and disease treatments. It is also important to note in any discussion of clinical research participants that family and friends play an equally important role as the participants themselves, providing the encouragement, information, support, and environment that may lead to or support such individual's participation in a given study [3, 15–17].

As mentioned previously, patient advocacy organizations also play a major role in clinical research, largely through a combination of (1) promoting policy and funding initiatives intended to motivate and support clinical research efforts in targeted disease states and (2) providing a medium by which potentially large cohorts of study participants may be recruited. In recent years, patient advocacy organizations have been taking increasingly active roles in shaping the agenda of the clinical research community, especially in rare and genetic diseases [6, 11, 14, 15, 18].

Academic Health Centers

Any number of sites can serve as the host for a given clinical research program, including individual physician practices, for-profit or not-for-profit clinics and hospitals, academic health centers (AHCs), colleges or universities, or community-based institutions such as schools and churches (to name a few of many examples). However, by far, the most common site for the conduct of clinical research in the United States is the AHC [3, 5, 15, 19]. During the conduct of clinical studies, AHCs or equivalent entities may take on any number or combination of the following responsibilities:

- Obtaining local regulatory and human subjects protection approval for a research study (e.g., IRB approval)
- Identifying, screening, and enrolling or registering study participants
- Delivery of study-specific interventions
- Collection of study-specific data
- Required or voluntary reporting of study outcomes and adverse events

As part of these responsibilities, study sites such as AHCs take on significant fiscal and ethical liabilities and risks related to a study's aim and objectives. Such fiscal risks are most often times shared with study sponsors, while ethical liabilities must be mitigated through the provision and maintenance of appropriate training and oversight structures for site-specific investigators or research staff.

Within an AHC, it is common for clinical studies to be motivated by a champion, who most often serves as the study investigator. Such investigators take primary responsibility for the clinical, scientific, and ethical design and conduct of a study

within their immediate or otherwise defined scope of control and influence (e.g., at a site or across a network of sites in the cases of a study site and sponsor-affiliated investigator, respectively). Study investigators may be engaged in a number of study-related activities for a given clinical research program, including:

- Development of preclinical or other pilot data as required to support study objectives and design
- Authoring and approval of study protocol documents
- Securing local or broader-scale regulatory and ethical approval
- Interactions with study participants in order to either deliver study-based interventions or collect study-related data elements
- Analysis and reporting of study outcomes and adverse events
- Analysis and reporting of data and knowledge generated during the course of a study (both regulatory reporting and scholarly communication, such as articles or presentations)

In addition to these activities, investigators are also responsible for overseeing the activities of research staff involved in a study and ensuring that the actions of those staff comply with applicable best practices and regulatory or ethical frameworks. In some studies, investigators may also serve as a type of study sponsor, usually when the hypotheses or interventions being evaluated are the result of the investigator's own scientific discoveries or research questions. We refer to such studies as being "investigator initiated." Most investigator-initiated studies are of a small scale and are funded using a combination of institutional and grant-related resources [9, 13–15].

Another recurring feature of AHCs is the engagement of research staff in the conduct of studies. Such research staff can be either fully focused upon research activities or only partially focused on such efforts, depending on their organization and role. Examples of research staff members include research coordinators/associatesassistants, data managers, statisticians, nurses, allied healthcare professionals, and information technology professionals. Such individuals usually serve as investigator "extenders," performing the detailed and day-to-day work required to satisfy the range of study-related tasks and activities attributed to investigators in the preceding discussion. There are numerous professional groups and certifications for such individuals, who normally serve as the true implementers of the vast majority of clinical research projects.

Clinical or Contract Research Organizations

Clinical or contract research organizations (CROs) are agencies that administer and facilitate clinical research processes and activities, most often on a contract basis that is funded by the study sponsor. Such CROs often provide study monitoring or regulatory support (acting as a proxy for sponsors and/or regulatory bodies) as well as study-specific research staffing relative to conduct research encounters and/or manage study-related data sets. The use of CROs is most prevalent in

studies involving multiple sites that must adhere to and administer a common research protocol across those sites. In this role, the CRO can ensure consistency of study processes and procedures and support participating sites, such as community-based practices, that may not nominally have the research experience or staff usually seen in AHCs.

Sponsoring Organization

Sponsoring organizations are primarily responsible for the origination and funding of clinical research programs (except in the case of investigator-initiated clinical trials, as discussed earlier). Examples of sponsors include pharmaceutical and biotechnology companies, nonprofit organizations, as well as government agencies, such as the National Institutes of Health. Sponsors may be responsible for some combination of the following tasks or activities during the clinical research life cycle:

- Conducting preclinical studies (e.g., animal models, *in silico* evaluations) of therapeutic interventions
- Developing or securing therapeutic agents or devices that are appropriate for use in human subjects
- Preparing a study protocol and informed consent documents and obtaining necessary regulatory approvals
- Identifying and engaging sites and/or investigators to execute a trial
- Negotiation and funding of protocol contracts, grants, or other fiscal and operational agreements as required to scope, inform, and fund a given study
- Training investigators concerning study procedures and activities
- Coordinating and monitoring data collection, including the performance of data quality assurance checking (often referred to as monitoring)
- Preparation and submission of required or otherwise necessary reports concerning trial activities, outcomes, and adverse events
- Aggregation, analysis, and dissemination of study data, outcomes, and findings

As can be surmised from the preceding exemplary list of sponsor tasks and activities, the nature of such items is broadly variable given the type of clinical research program being executed. For example, in the case of a trial intended to evaluate a novel therapy for a specified disease state, a private sector sponsor could be responsible for all of the preceding tasks (any of which could theoretically be outsourced to a CRO). In contrast, in the case of an epidemiological study being conducted by a government agency, such a sponsor may only be engaged in a few of these types of tasks and activities (e.g., preparing a protocol, identifying and engaging sites, funding participation, and aggregating or analyzing study results or findings). Ultimately and in the vast majority of clinical research programs, the sponsor possesses the greatest fiscal or intellectual property “stake” in the design, conduct, and outcomes of a study [9, 13–15].

Federal Regulatory Agencies

Federal regulators are primarily responsible for overseeing the safety and appropriateness of clinical research programs, given applicable legal frameworks, community-accepted best practices, and other regulatory responsibilities or requirements. Examples of federally charged regulators can include institutional review boards (IRBs, who act as designated proxies for the US Department of Health and Human Services (DHHS) relative to the application and monitoring of human subjects protection laws) as well as agencies such as the Food and Drug Administration (FDA). Such regulators can be responsible for numerous tasks and activities throughout the clinical research life cycle, including:

- Approving clinical research studies in light of applicable legal, ethical, and best practice frameworks or requirements
- Performing periodic audits or reviews of study data sets to ensure the safety and legality of interventions or other research activities being undertaken
- Collecting, aggregating, and analyzing voluntary and required reports concerning the outcomes of or adverse events associated with clinical research activities

Broadly characterized, the overriding responsibility of regulators is to ensure the safety of study participants as well as monitor the adherence of study investigators and staff with often times complex regulatory and ethical requirements that define the responsible and appropriate conduct of a given research model or approach [4, 6].

Healthcare and Clinical Research Information Systems Vendors

Software developers and vendors play a number of roles in the clinical research environment, including (1) designing, implementing, deploying, and supporting clinical trial management systems and/or research-centric data warehouses that can be used to collect, aggregate, analyze, and disseminate research-oriented data sets; (2) providing the technical mechanisms and support for the exchange of data between information systems and/or sites involved in a given clinical research program; and (3) facilitating the secondary use of primarily clinical data in support of research (e.g., developing and supporting research-centric reporting tools that can be applied against operational clinical data repositories associated with electronic health record systems) [1, 8, 10, 20, 21]. Given the ever-increasing adoption of healthcare information technology (HIT) platforms in the clinical research domain and the corresponding benefits of reduced data entry, increased data quality and study protocol compliance, and increased depth or breadth of study data sets, the role of such healthcare and clinical research information systems vendors in the clinical research setting is likely to increase at a rapid rate over the coming decades. Further, with the advent of open standards for the interoperability of data across and between such HIT platforms, entirely new modalities for the capture, integration,

QA, and reporting of data relevant to the conduct of clinical research are becoming possible and helping to overcome numerous resource barriers that may have otherwise impeded the conduct of large-scale and/or complex studies [21–23].

Other Clinical Research Actors

Additional actors who play roles in the clinical research setting include the following [9, 15]:

- Administrative managers/coordinators: Administrative managers and coordinators are often responsible for multiple aspects of regulatory or sponsor reporting, administrative tracking/compliance, budgeting and fiscal reconciliation, and human subjects protection reporting and monitoring.
- Data safety and monitoring boards (DSMBs): DSMBs are usually comprised of individuals without a direct role in a given study and who are charged with overseeing the safety and efficacy of study-related interventions. The members of a DSMB are usually empowered to halt or otherwise modify a study if such factors are not satisfied in a positive manner. A related mechanism for patient safety oversight in observational research studies is the Observational Study Monitoring Board (OSMB).

Common Clinical Research Settings

As was noted in the earlier sections of this chapter, clinical research programs are most commonly situated in AHCs. However, such institutions are not the sole environment in which clinical research occurs. In fact, as will be discussed in greater detail in section “[Identifying Potential Study Participants](#)”, there are significant trends in the clinical research community toward the conduct of studies in community practice and practice-based network (e.g., organized networks of community practice sites with shared administrative coordinating processes and agents) settings as well as global-scale networks. The primary motivations for such evolution in the practice of clinical research include (1) an access to sufficiently large participant populations, particularly in rare diseases or studies requiring large-scale and diverse patient populations; (2) reduced costs or regulatory overhead; and (3) increasing access to study-related therapies in underserved or difficult to access communities or geographic environments [1, 16, 24, 25].

Common Clinical Research Goals

In a broad sense, the objectives or goals of most clinical research programs can be stratified into one or more of the design patterns summarized in Table 3.1. These patterns serve to define the intent and methodological approach of a given study or program of research.

Table 3.1 Summary of clinical research design patterns

Pattern description	Goals/objectives	Exemplary methodological approaches
Evaluation of the safety of a new or modified therapy	Establish safety of therapy as prerequisite for efficacy testing	Phase I clinical trial ^a
Evaluation of the efficacy (ability to positively effect a targeted disease state) of a new or modified therapy	Establish efficacy of therapy relative to targeted disease state as prerequisite for comparison to existing therapies	Phase II clinical trial ^a
Comparison of new or modified therapy to existing therapies	Establish benefits or equivalency of new or modified therapy relative to existing therapies	Phase III clinical trial ^a
Observation of the longitudinal effects of a new, modified, or existent therapy	Identify long-term effects of therapies and population level	Phase IV clinical trial ^a
Collection of observational data to identify clinical, behavioral, or other manifested phenomena of interest	Identify phenomena of interest that serve to inform basic science, clinical, or population-level studies and interventions	Observational study Ethnography Surveys Interviews
Collection of biospecimens and/or correlative clinical data	Identify and collect biospecimens and data that can support retrospective studies and/or hypothesis generation activities	Biospecimen banking Remnant tissue capture

^aThe gold standard for such methodological approaches is the randomized controlled trial (*RCT*)

A Framework for Data and Information Management Requirements in Clinical Research

In order to better understand the relationships between the information needs of clinical researchers and available data management and informatics tools or platforms, it is helpful to conceptualize the conduct of clinical research programs as a multiple-stage sequential model [26]. At each stage in this model, a combination of general purpose, clinical, and research-specific HIT systems may be utilized. Examples of general purpose and clinical systems that are able to support the conduct of clinical research include:

- Literature search tools such as the National Library of Medicine's PubMed can be used to assist in conducting the background research necessary for the preparation of protocol documents.
- Electronic health records (EHRs) can be utilized to collect clinical data on research participants in a structured form that can reduce redundant data entry.
- Data mining tools can be used in multiple capacities, including (1) determining if participant cohorts meeting the study inclusion or exclusion criteria can be practically recruited given historical trends and (2) identifying specific participants and related data within existing databases (also see Chap. 16).

- Decision support systems can be used to alert providers at the point of care that an individual may be eligible for a clinical trial.
- Computerized physician order entry (CPOE) systems, which collect data describing the therapies delivered to research participants, can be used in both participant tracking and study analyses.

In addition to the preceding general purpose and clinical systems, research-specific IT systems have been developed that include:

- Simulation and visualization tools can streamline the preclinical research process (e.g., disease models) and assist in the analysis of complex data sets.
- Protocol authoring tools can allow geographically distributed authors to collaborate on complex protocol documents.
- Participant screening tools can assist in the identification and registration of research participants.
- Research-specific web portals provide researchers with a single point of access to research-specific documents and information.
- Electronic data collection or capture tools (EDC) can be used to collect research-specific data in a structured form and reduce the need for redundant and potentially error-prone paper-based data collection techniques.
- Research-specific decision support systems provide protocol-specific guidelines and alerts to researchers, for example, tracking the status of participants to ensure protocol compliance.

Clinical Research Workflow and Communications

Despite the critical role of workflow in determining both operational efficiencies and effective tactics for the deployment and adoption of information technology in the biomedical domain, there is a paucity of literature describing systematic clinical research workflow paradigms. However, a small body of literature does provide some insight into the basic workflows engaged in or experienced by clinical research investigators and staff and associated challenges and opportunities. In the following section, we will highlight a number of salient features of such findings, in order to provide a general overview of prevailing clinical research workflow characteristics.

Workflow Challenges

There are a number of workflow challenges that serve to characterize the clinical research environment [5, 10, 15, 21], including the four broad categories of such issues as summarized below:

Paper-Based Information Management Practices

As was noted previously, a majority of clinical research tasks and activities are completed or otherwise executed using some combination of paper-based information management practices. As with all such scenarios involving the use of paper-based information management, inherent limitations associated with paper, including its ability to only be accessed by one individual at one time in one location, severely limit the scalability and flexibility of such approaches. Furthermore, in many clinical research settings, with the number of ongoing studies that regularly co-occur, the proliferation of multiple paper-based information management schemes (e.g., study charts, binders, copies of source documentation, faxes, print-outs) leads to significant space and organizational challenges and inefficiencies.

Complex Technical and Communications Processes

In recent studies of clinical research workflow, it has been observed that most research staff conduct their activities and processes using a mixture of tools and methods, including the aforementioned paper-based information management schemas, as well as telephones, computers, and other electronic mediums, and interpersonal (e.g., face-to-face) communications. The combined effects of such complex combinations of tools and methods is an undesirable increase in cognitive complexity and corresponding decreases in productivity, accuracy, and efficiency, as described later in this chapter.

Interruptions

Again, as has been reported in recent studies, upwards of 18% of clinical research tasks and activities are interrupted, usually by operational workflow requirements (e.g., associated with the environment in which a study is occurring, such as a hospital or clinic) or other study-related activities. Much as was the case with the preceding issues surrounding complex technical and communication processes, such interruptions significantly increase cognitive complexity, with all of the associated negative workflow and efficiency implications.

Single Point of Information Exchange

One of the most problematic workflow challenges in the clinical research environment is the fact that, in many instances, a single staff member (most often a CRC) is the single point of research-related information management and exchange. In such instances, the physical and cognitive capacities, as well as availability of

such individuals, serve as a primary rate limiting component of overall research productivity and workflow. This phenomenon is most often associated with the scarcity of individuals with the necessary training to conduct clinical research activities and/or the availability of funding and resources to support such positions.

Cognitive Complexity

As was briefly introduced in the preceding discussion, many of the characteristics of the current clinical research environment lend themselves to increased cognitive complexity. At a high level, the concept of cognitive complexity refers to scenarios in which the frequent use of multiple methods and artifacts to accomplish a given task exceeds inherent human cognitive capacities for information retention and recall. In such instances, increased errors and reduced efficiencies are usually observed. Ideally, such cognitive complexity is alleviated through the implementation or optimization of workflows and tools that minimize the need to switch between modalities and artifacts in order to accomplish a task [27, 28]. A small number of studies in the clinical research setting, including efforts focusing on clinical trial management systems and, in particular, clinical trial participant calendaring applications, have demonstrated that the use of rigorous, human-centered design principles can reduce cognitive complexity and increase the speed and accuracy of task completion in commonly occurring clinical study tasks and events (such as scheduling and/or rescheduling protocol-related events) [29]. However, the proliferation of paper-based information management and manually oriented workflows in the modern research environment, largely as a result of slow or incomplete information technology adoption, continues to preclude large-scale reengineering efforts intended to tackle the important problem of cognitive complexity.

Emergent Trends in Clinical Research

In the preceding sections of this chapter, we have outlined the basic theories and methods that serve to inform the design and conduct of clinical research programs, as well as the stakeholders and their workflow characteristics that define the domain and current state of clinical research practice. Throughout these discussions, we have described the ways in which informatics theories and methods can enable or enhance such processes and activities. Building on this background, in the following section, we will explore some of the emergent trends in clinical research that will serve to drive future innovation in healthcare, the life sciences, and the role of informatics as it relates to the research activities needed to support and enable such innovation.

Precision or Personalized Medicine

The advent of national-scale research programs focusing on precision or personalized medicine has served to draw increased attention to the critical role of data and computation in terms of pursuing some of the most complex research questions in the health and life science domains. At its most basic level, precision (or personalized) medicine involves:

...the tailoring of medical treatment to the individual characteristics of each patient. It does not literally mean the creation of drugs or medical devices that are unique to a patient, but rather the ability to classify individuals into sub-populations that differ in their susceptibility to a particular disease, in the biology and/or prognosis of those diseases they may develop, or in their response to a specific treatment. (National Academies of Medicine, ‘Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease’, 2011)

As can be seen from this definition, being able to achieve the vision of precision medicine requires that we establish an evidence based that can serve to link a deep understanding of a patient’s individual biomolecular and clinical phenotype with the best available scientific evidence that may in turn inform an optimal therapeutic strategy given those characteristics. Building this knowledge base is an intrinsically clinical research focused endeavor, and it is through which large numbers of research participants will need to be recruited to participate in studies where such data and outcomes will be collected and analyzed either retrospectively or prospectively. Doing so introduces numerous challenges relative to the design and execution of such studies, including being able to recruit sufficient numbers of participants or finding alternative strategies for the design of studies that can overcome the need to recruit large numbers of individuals but instead focus on generating more targeted data that can quickly prove or disprove a hypothesized connection between phenotype and treatment outcomes [11, 18, 30–32]. Programs such as the “All of Us” initiative, sponsored by the US National Institute of Health (NIH), serve as prime examples of this emergent area of activity [33, 34].

Learning Healthcare Systems and Evidence Generating Medicine

In a manner that is closely aligned with the emergence of precision and personalized medicine as a national and international research priority, there is also an increasing awareness of the need to instrument the healthcare delivery environment such that every patient encounter becomes an opportunity to learn and improve the collective

biomedical knowledge base. Such activities are often referred to as the creation of “learning healthcare systems” that can support or enable “evidence generating medicine.” In this context, we can define a learning healthcare system as a system in which:

science, informatics, incentives, and culture are aligned for continuous improvement and innovation, with best practices seamlessly embedded in the delivery process and new knowledge captured as an integral by-product of the delivery experience. (National Academies of Medicine, ‘The Learning Healthcare System’ 2015)

In such a model, we move beyond a unidirectional relationship between evidence generation (e.g., clinical research) and practice, toward a model in which there is a continuous cycle of learning that feeds data from the point of care to researchers for analysis, with ensuing knowledge products being delivered for clinical decision-making via rapid-cycle innovation [20]. Achieving this type of outcome requires a number of clinical research innovations, including (1) the creation of data capture instruments within EHRs and other clinical systems that are compatible with both standard of care and research activities (e.g., delivering sufficient data while not impeding clinical workflow), (2) the establishment of pragmatic clinical research designs that can produce empirically defensible results with incomplete or otherwise “messy” data resulting from clinical care activities, and (3) the implementation of mechanisms for returning actionable knowledge generated via the analysis of such data to the point of care in short time frames, often via computable guidelines and/or decision support rules [7, 13, 16, 20, 21, 24, 32].

Real-World Evidence Generation (RWE)

Finally and again in a manner that is synergistic with the two preceding themes (e.g., precision or personalized medicine and learning healthcare systems or evidence generating medicine), there is also an increasing focus being placed by the biotechnology and pharmaceutical industries on the pursuit of what is known as real-world evidence (RWE) generation. Such RWE extends beyond traditional post-market surveillance of drug safety and efficacy, toward the collection of “real-world” data that can help to identify new uses for existing therapeutics and/or identify potential toxicities and adverse events associated with the use of predictive modeling methods before such issues become widespread. In a formal sense, RWE is the product of analyses applied to real-world data (RWD), which can be defined as:

the data relating to patient health status and/or the delivery of health care routinely collected from a variety of sources. RWD can come from a number of sources, for example: 1) Electronic health records (EHRs); 2) Claims and billing activities; 3) Product and disease registries; 4) Patient-related activities in outpatient or in-home use settings; and 5) Health-monitoring devices. <https://www.fda.gov/sciencceresearch/specialtopics/realworldevidence/default.htm>

One of the most common examples of leveraging RWD to generate RWE is the retrospective analysis of collections of disease-specific registries generated during the course of either prospective trials or observational studies [6, 12, 35]. In such instances, informaticians, data scientists, and statisticians find ways to link and integrate such data so that longitudinal or outcome-oriented hypotheses can be tested with large amounts of data within short time frames. Such study designs represent new models for defining and conducting clinical studies, particularly when the therapeutic agent of interest is already FDA approved and in widespread use or when seeking to conduct the sorts of analyses needed to establish a precision medicine knowledge base.

Conclusion

As stated in the introduction to this chapter, the primary learning objectives to be addressed were associated with following three aims:

1. To describe the basic processes, activities, stakeholders, environments, and goals that serve to characterize the modern physical and sociotechnical clinical research environment
2. To introduce a framework of clinical research information management needs
3. To summarize the current state of an evolving body of research and knowledge that seeks to characterize clinical research workflow and communications patterns, in order to support the optimal design and implementation of informatics platforms in the clinical research environment

We have addressed these objectives and aims by reviewing common processes, activities, stakeholders, environmental settings, and goals that characterize the contemporary clinical research environment. We have also introduced a conceptual model by which the information needs incumbent to the clinical research domain can be satisfied by a combination of general purpose and research-specific information systems. Finally, we have introduced the major workflow activities and challenges that exist in the clinical research setting, as well as emerging trends in the broad health and life sciences research domain that are helping to advance the state of clinical research design and practice. Taken as a whole, this overview should equip readers with a solid grounding by which they can place the content in the remainder of this text in context. Furthermore, this background should serve as the basis for educating clinical research informatics researchers and professionals about the basics of clinical research design and practice, thus catalyzing their acculturation to this critical and rapidly evolving domain.

References

1. Embi PJ, Payne PR. Clinical research informatics: challenges, opportunities and definition for an emerging domain. *J Am Med Inform Assoc.* 2009;16(3):316–27.
2. Embi PJ, Payne PR. Advancing methodologies in Clinical Research Informatics (CRI). *J Biomed Inform.* 2014;52(C):1–3.
3. Johnson SB, Farach FJ, Pelphrey K, Rozenblit L. Data management in clinical research: synthesizing stakeholder perspectives. *J Biomed Inform.* 2016;60:286–93.

4. Kahn MG, Weng C. Clinical research informatics: a conceptual perspective. *J Am Med Inform Assoc.* 2012;19(e1):e36–42.
5. Payne PR, Pressler TR, Sarkar IN, Lussier Y. People, organizational, and leadership factors impacting informatics support for clinical and translational research. *BMC Med Inform Decis Mak.* 2013;13(1):20.
6. Weng C, Kahn M. Clinical research informatics for big data and precision medicine. *IMIA Yearb.* 2016;(1):211–8.
7. Embi PJ, Kaufman SE, Payne PR. Biomedical informatics and outcomes research. *Circulation.* 2009;120(23):2393–9.
8. Goldenberg NA, Daniels SR, Mourani PM, Hamblin F, Stowe A, Powell S, et al. Enhanced infrastructure for optimizing the design and execution of clinical trials and longitudinal cohort studies in the era of precision medicine. *J Pediatr.* 2016;171:300–6. e2.
9. Prokscha S. Practical guide to clinical data management. Boca Raton: CRC Press; 2011.
10. Richesson R, Horvath M, Rusincovitch S. Clinical research informatics and electronic health record data. *Yearb Med Inform.* 2014;9(1):215.
11. Saad ED, Paoletti X, Burzykowski T, Buyse M. Precision medicine needs randomized clinical trials. *Nat Rev Clin Oncol.* 2017;14(5):317–23.
12. Nelson EC, Dixon-Woods M, Batalden PB, Homa K, Van Citters AD, Morgan TS, et al. Patient focused registries can improve health, care, and science. *BMJ.* 2016;354:i3319.
13. Pencina MJ, Peterson ED. Moving from clinical trials to precision medicine: the role for predictive modeling. *JAMA.* 2016;315(16):1713–4.
14. Friedman LM, Furberg C, DeMets DL, Reboussin DM, Granger CB. Fundamentals of clinical trials. Cham: Springer; 1998.
15. Hulley SB, Cummings SR, Browner WS, Grady DG, Newman TB. Designing clinical research. Philadelphia: Lippincott Williams & Wilkins; 2013.
16. Brightling CE. Clinical trial research in focus: do trials prepare us to deliver precision medicine in those with severe asthma? *Lancet Respir Med.* 2017;5(2):92–5.
17. Browner WS. Publishing and presenting clinical research. Philadelphia: Lippincott Williams & Wilkins; 2012.
18. Vicini P, Fields O, Lai E, Litwack E, Martin AM, Morgan T, et al. Precision medicine in the age of big data: the present and future role of large-scale unbiased sequencing in drug discovery and development. *Clin Pharmacol Ther.* 2016;99(2):198–207.
19. Korn EL, Freidlin B. Adaptive clinical trials: advantages and disadvantages of various adaptive design elements. *JNCI J Natl Cancer Inst.* 2017;109(6):djx013.
20. Embi PJ, Payne PR. Evidence generating medicine: redefining the research-practice relationship to complete the evidence cycle. *Med Care.* 2013;51:S87–91.
21. Murphy SN, Dubey A, Embi PJ, Harris PA, Richter BG, Turisco F, et al. Current state of information technologies for the clinical research enterprise across academic medical centers. *Clin Transl Sci.* 2012;5(3):281–4.
22. Mandel JC, Kreda DA, Mandel KD, Kohane IS, Ramoni RB. SMART on FHIR: a standards-based, interoperable apps platform for electronic health records. *J Am Med Inform Assoc.* 2016;23(5):899–908.
23. Mandel KD, Mandel JC, Kohane IS. Driving innovation in health systems through an apps-based information economy. *Cell Syst.* 2015;1(1):8–13.
24. Chambers DA, Feero WG, Khoury MJ. Convergence of implementation science, precision medicine, and the learning health care system: a new model for biomedical research. *JAMA.* 2016;315(18):1941–2.
25. Embi PJ. Future directions in clinical research informatics. *Clinical research informatics.* New York: Springer; 2012. p. 409–16.
26. Payne PR, Johnson SB, Starren JB, Tilson HH, Dowdy D. Breaking the translational barriers: the value of integrating biomedical informatics and translational research. *J Investigig Med.* 2005;53(4):192–201.
27. Patel VL, Arocha JF, Kaufman DR. A primer on aspects of cognition for medical informatics. *J Am Med Inform Assoc.* 2001;8(4):324–43.

28. Zhang J, Patel VL. Distributed cognition, representation, and affordance. *Pragmat Cogn.* 2006;14(2):333–41.
29. Payne PR. Advancing user experience research to facilitate and enable patient-centered research: current state and future directions. *eGEMS.* 2013;1(1):1026.
30. Ashley EA. Towards precision medicine. *Nat Rev Genet.* 2016;17(9):507–22.
31. Hunter DJ. Uncertainty in the era of precision medicine. *N Engl J Med.* 2016;375(8):711–3.
32. Tenenbaum JD, Avillach P, Benham-Hutchins M, Breitenstein MK, Crowgey EL, Hoffman MA, et al. An informatics research agenda to support precision medicine: seven key areas. *J Am Med Inform Assoc.* 2016;23(4):791–5.
33. Collins FS, Varmus H. A new initiative on precision medicine. *N Engl J Med.* 2015;372(9):793–5.
34. Sankar PL, Parker LS. The precision medicine initiative’s all of us research program: an agenda for research on its ethical, legal, and social issues. *Genet Med.* 2017;19(7):743.
35. Ekins S. Pharmaceutical and biomedical project management in a changing global environment. Hoboken: Wiley; 2011.



Methodological Foundations of Clinical Research

4

Antonella Bacchieri and Giovanni Della Cioppa

Abstract

This chapter focuses on clinical experiments, discussing the phases of the pharmaceutical development process. We review the conceptual framework and classification of biomedical studies and look at their distinctive characteristics. Biomedical studies are classified into two main categories, observational and experimental, which are then further classified into subcategories of prospective and retrospective and community and clinical, respectively. We review the basic concepts of experimental design, including defining study samples and calculating sample size, where the sample is the group of subjects on which the study is performed. Choosing a sample involves both qualitative and quantitative considerations, and the sample must be representative of the population under study. We then discuss treatments, including those that are the object of the experiment (study treatments) and those that are not (concomitant treatments). Minimizing bias through the use of randomization, blinding, and a priori definition of the statistical analysis is also discussed. Finally, we briefly look at innovative approaches, for example, how adaptive clinical trials can shorten the time and reduce the cost of classical research programs or how targeted designs can allow a more efficient use of patients in rare conditions.

Keywords

Phase I, II, III, and IV trials · Classification of biomedical studies · Observational study · Experimental study · Equivalence/non-inferiority studies · Superiority versus non-inferiority studies · Crossover designs · Parallel group designs · Adaptive clinical trials · Targeted designs

A. Bacchieri, MS (✉)
CROS NT srl and Clinical R&D Consultants srls, Verona, Rome, Italy
e-mail: anto.bacchieri@alice.it

G. Della Cioppa, MD
Clinical R&D Consultants srls, Rome, Italy

The Development of Pharmaceuticals: An Overview

The development of a pharmacological agent (preventive, diagnostic, or therapeutic) from start to first launch on the market typically lasts in excess of 10 years, at times considerably longer, and thereafter continues throughout its life cycle, often for decades postmarketing [1] (see Chap. 20).

Clinical experiments, the focus of this chapter, are preceded by many years of preclinical development. In very broad terms, the preclinical development process can be summarized in a sequence of seven phases [2]:

1. Screening of hundreds of candidates by means of biological assays. Candidates may be produced chemically or through biological systems.
2. Selection of the lead compound.
3. Physicochemical characterization of the lead compound.
4. Formulation of the drug product, consisting of drug substance, excipients, and delivery system.
5. Scale-up of production and quality control.
6. Toxicology experiments.
7. Preclinical pharmacology, which includes pharmacokinetics (what the body does to the drug: absorption, distribution, metabolism, and excretion – ADME) and pharmacodynamics (what the drug does to the different organs and body systems).

There is a considerable chronological overlap between phases with multiple iterations and parallel activities, many of which continue well into the clinical stages. As the clinical experiments proceed and the level of confidence on the potential of a new compound grows, experiments also proceed in nonclinical areas, from toxicology to production, in preparation for the more advanced clinical phases and finally for commercialization.

Conventionally, the clinical development process is divided into four phases, referred to as Phases I, II, III, and IV.

Phase I begins with the first administration of the compound to humans. The main objectives of Phase I investigation are:

1. Obtain indications on the safety and tolerability of the compound.
2. Study its pharmacokinetics in humans, when appropriate.
3. Obtain preliminary indications on pharmacodynamics.

Typically Phase I trials are conducted over a large range of doses. Whereas traditionally Phase I is conducted in healthy volunteers, increasingly Phase I studies are carried out directly in patients.

Phase II studies are carried out on selected groups of patients suffering from the disease of interest, although patients with atypical forms and concomitant diseases are excluded. Objectives of Phase II are:

1. Demonstrate that the compound is active on relevant pharmacodynamic endpoints (*proof of concept*).

2. Select the dose (or doses) and dosing schedule(s) for Phase III (dose-finding).
3. Obtain safety and tolerability data.

Sometimes Phase II is divided further into two *subphases*: IIa, for proof of concept, and IIb, for dose-finding.

The aim of *Phase III* is to demonstrate the clinical effect (therapeutic or preventive or diagnostic), safety, and tolerability of the drug in a representative sample of the target population, with studies of sufficiently long duration relative to the treatment in clinical practice. The large Phase III studies, often referred to as *pivotal* or *confirmatory*, are designed to provide decisive proof in the registration dossier.

All data generated on the experimental compound, from the preclinical stage to Phase III, and even Phase IV (see below), when it has already been approved in other countries, must be summarized and discussed in a logical and comprehensive manner in the *registration dossier*, which is submitted to health authorities as the basis for the request of approval. In the last 30 years, a large international effort took place to harmonize the requirements and standards of many aspects of the registration documents. Such efforts became tangible with the guidelines of the International Conference on Harmonisation (ICH) (www.ich.org). These are consolidated guidelines that must be followed in the clinical development process and the preparation of the registration dossiers in all three regions contributing to ICH: Europe, the United States, and Japan. An increasing number of regulatory authorities, including Chinese, Canadian and Australian, have adopted guidelines similar to ICH. With regard to the registration dossier, the ICH process culminated with the approval of the Common Technical Document (CTD). The CTD is the common format of the registration dossier recommended by the European Medicines Agency (EMA), the US Food and Drug Administration (FDA), and the Japanese Ministry of Health, Labour and Welfare (MHLW). The CTD is organized in five modules, each composed of several sections. Critical for the clinical documentation are the Efficacy Overview, the Safety Overview, and the Conclusions on Benefits and Risks. The overviews require pooling of data from multiple studies into one or more integrated databases, from which analyses on the entire population and/or on selected subgroups are carried out. In the assessment of efficacy, pooling may be necessary for special groups such as the elderly or subjects with renal or hepatic impairment. In the assessment of safety and tolerability, large integrated databases are critical for the evaluation of infrequent adverse events and for subgroup analyses by age, sex, race, dose, etc. The merger of databases coming from different studies requires detailed planning at the beginning of the project. The more complete the harmonization of procedures and programming conventions of the individual studies, the easier the final pooling. Vice versa, the lack of such harmonization will cause an extenuating ad hoc programming effort at the end of the development process, which will inevitably require a number of arbitrary assumptions and coding decisions. In some cases, this can reduce the reliability of the integrated database.

Clinical experimentation of a new treatment continues after its approval by health authorities and launch onto the market. Despite the approval, there are always many questions awaiting answers. *Phase IV* studies provide some of the answers. The expression *Phase IV* is used to indicate clinical studies performed after the approval of a new drug and within the approved indications and restrictions imposed by the Summary of Product Characteristics and the Package Insert.

The sequence of clinical development phases briefly described above is an oversimplification, and many departures occur in real life. For example, Phases I and II are frequently combined. Phases II and III may also be merged in an adaptive design trial (described later). Further, compounds in oncology have many peculiarities in their clinical development, mainly concerning Phases I and II. These differences are determined mostly by the toxicity of many compounds, even at therapeutic or subtherapeutic doses, combined with the life-threatening nature of the diseases in question. Rare diseases are another broad field where the above sequence of phases is not followed, due to the limited number of patients.

As mentioned above, the clinical development process for a new diagnostic, preventive, or therapeutic agent is extremely long and the costs correspondingly high, often exceeding 10 years and 800 million USD, respectively [3]. Therefore, faster and cheaper development has always been a key objective for pharmaceutical companies, academic institutions, and regulatory agencies alike. Clearly, there is no magic solution, and no method is universally applicable. However, new methodological and operational solutions have been introduced, which contribute in selected situations to reducing the overall time of clinical development and/or lowering costs. Among the most efficient tools are the following:

- Modeling and simulation statistical techniques aimed at evaluating the consequences of a variety of assumptions, i.e., answering “what happens if...” questions. Simulations are used for many purposes, including detection of bias, comparison of different study designs, contribution to dose and schedule selection, and evaluation of the consequences of different decision-making rules in determining the success or failure of a study or an entire study program.
- Strategies that combine different phases of development, mainly Phases II and III, such as adaptive designs (described later).
- Technological innovations such as electronic data capture (EDC), which allows data entry directly by the study staff at the site into a central database without the intermediate step of traditional paper case report forms (CRFs) or direct download from measurement instruments into the central database without any manual intervention.
- Special regulatory options made available for the very purpose of accelerating clinical development of lifesaving and essential treatments. Prominent among these are the *Treatment IND* (FDA) and the *mock application* (EMA) for the approval of vaccines in outbreak situations, such as the H1N1 swine flu pandemic.

Conceptual Framework and Classification of Biomedical Studies

Variability of Biological Phenomena

All biological phenomena as we perceive them are affected by variability. The overall goal of any biomedical study is to separate the effect related to an intervention (the *signal*) from the background of variability of biological phenomena unrelated to the intervention ([1], Chap. 1).

Variability of biological phenomena can be divided into three main components:

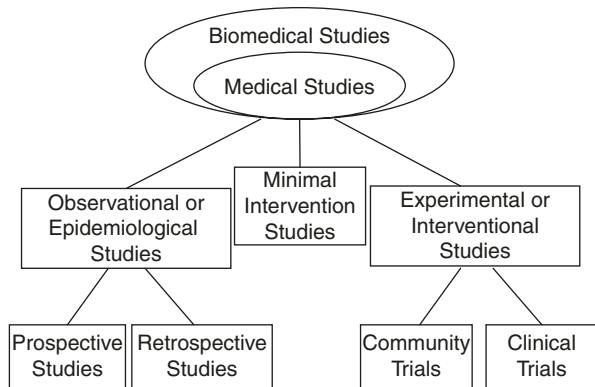
1. *Phenotypic variability*, i.e., differences between individuals at a given point in time.
2. *Temporal variability*, i.e., changes in a given individual over time. Temporal variability can be predictable and cyclical (e.g., hormonal changes during the menstrual cycle), predictable and noncyclical (e.g., age-related changes of height), or erratic and unpredictable. An element of unpredictability is always superimposed to any biological phenomenon undergoing predictable temporal changes; for example, the hormonal changes during the menstrual cycle, although predictable quantitatively and chronologically, can still be very different from month to month.
3. *Measurement-related variability*, due to the use of measurement instruments. External phenomena exist for us only to the extent they are detected by our senses and understood by our intellect. To understand an external phenomenon, we first have to recognize it and then measure it. Measurement is the process of assigning a quantity and/or symbol to a variable according to a predefined set of rules. The set of rules is often implicit: for example, the statement “my friend Ann died young at age 40” implies the assignment of a quantity (*young*) to Ann’s age at the time of death, based on the implicit rule that the *normal* time of death is much later than age 40, say, 85 or more. In scientific measurements, the set of rules is explicit and defined by the measurement scale used. Variability related to the measuring process becomes an integral part of the variability of biological phenomena as we perceive them. Errors made in the process of measuring can be of two types: random and systematic.
 - A *random error* generates measurements that oscillate unpredictably about the true value. Example: rounding off decimals from two digits to one.
 - A *systematic error*, also referred to as *bias* or distortion, generates measurements that differ from the true value always in the same direction. Example: measuring weight with a scale that is not correctly calibrated and, therefore, always underestimates or overestimates weight.

Both random error and bias have an impact on the reliability of results of biomedical studies. Random error causes greater variability. This can be rescued to some extent by increasing the sample size of a study. Bias may simulate or obscure the treatment effect. This cannot be rescued: bias can only be prevented by a proper design of the study (see below).

Biomedical Studies: Definitions and Classification

Biomedical studies are experiments with the objective of establishing a relationship between a characteristic or intervention and a disease or condition. The relationship of interest is one of cause-effect. The element which makes the biomedical studies different from deterministic experiments is the variability of the phenomenon under

Fig. 4.1 Classification of biomedical studies.
(Adapted from Bacchieri and Della Cioppa [1])



study. As mentioned above, all methods and techniques used in biomedical studies have the overall goal of differentiating a true cause–effect relationship from a spurious one, due to the background noise of biological variability and/or to bias.

Biomedical studies must have four critical distinctive characteristics:

1. Rationale, methods, and conclusions must be based on comparisons between groups of subjects.
2. The groups of subjects being compared must be similar, i.e., must have similar distribution of important demographic and clinical characteristics.
3. An adequate probabilistic model “tailored” exactly to the problem under study must allow the conclusions from the specific study to be applied to the underlying population (inference).
4. All aspects of the study must be planned in advance, in most cases before the study starts and in all cases before the data are analyzed.

Biomedical studies can be classified as shown in Fig. 4.1 [1]. Medical studies are the subset of biomedical studies which involve human subjects. These studies are classified in two main categories: observational and experimental.

Observational or Epidemiological Studies

In observational studies, also referred to as epidemiological studies, the association between a characteristic and an event is investigated without any type of intervention. When the entity of the association is relevant, a causal relationship is assumed. The characteristic being studied can be a pharmacological treatment or a demographic, behavioral, or environmental factor. The event can be the occurrence or recrudescence of a disease, hospitalization, death, etc. If the characteristic modifies the event in a favorable way, it is called *protective factor*; if it modifies the event in a negative way, it is called *risk factor* [4].

There are two main types of design for observational studies: prospective (or cohort) and retrospective (or case control) ([1], Chap. 3). In *prospective studies*, subjects are selected on the basis of the presence or absence of the characteristic. Prospective studies are also referred to as cohort studies. In a prospective study, the researcher selects two groups of subjects, one with the characteristic under study (exposed) and the other without (non-exposed). For example, exposed could be subjects who are current cigarette smokers and non-exposed those who never smoked cigarettes or have quit smoking. With the exception of the characteristic under study, the two groups should be as similar as possible with respect to the distribution of key demographic features (e.g., age, sex, socioeconomic status, health status). Each enrolled subject is then observed for a predefined period to assess if, when, and how the event occurs. In our example, the event could be a diagnosis of lung cancer. Prospective studies can be classified based on time in three types: *concurrent* (the researcher selects exposed and non-exposed subjects in the present and prospectively follows them into the future), *non-concurrent* (the researcher goes back in time, selects exposed and non-exposed subjects based on exposure in the past, and then traces all the information relative to the event of interest up to the present), and *cross-sectional* (the researcher selects subjects based on the presence/absence of the characteristic of interest in the present and searches the event in the present).

In *retrospective studies*, subjects are selected on the basis of the presence or absence of the event. Retrospective studies are often referred to as case-control studies. In a retrospective study, the researcher selects two groups of subjects, one group with the event of interest (cases) and the other without (controls). In order to increase comparability between cases and controls, each case is often matched to one or more controls for a few key demographic features (e.g., sex, age, ethnicity). In our example, cases are subjects with a diagnosis of lung cancer; each case could be matched with one or more controls, similar for important characteristics, for example, sex, age, work exposure to toxic air pollutants, and socioeconomic status. The medical history of each enrolled subject is then investigated to see whether, during a predefined period of time in the past, he/she was exposed (and when and how much) to the characteristic under study, in our example cigarette smoking.

Retrospective studies can be classified based on time in two types: *true retrospective* (the researcher selects the subjects with and without the event and goes back in time to search for exposure) and *cross-sectional* (the researcher selects subjects based on the presence/absence of the event but limits the investigation about the exposure to the present).

Experimental or Interventional Studies

In experimental studies, also referred to as interventional, the researcher has the control of the conditions under which the study is conducted. The intervention, typically a therapeutic or preventive treatment, also referred to as an experimental factor, is not simply observed; the subjects are assigned to the intervention by the researcher, generally by means of a procedure called *randomization* (see below). The assignment of

the study subjects to the intervention can be done by groups of subjects (community trial) or, more frequently, by individual subject (clinical trial). Many other factors besides the experimental factor can influence the study results. These are referred to as sub-experimental factors. Some are known (e.g., age, sex, previous or concomitant treatments, study site, degree of severity of the disease), but most are unknown. In experimental studies, the investigator not only controls the assignment of the experimental factor but also attempts to control as much as possible the distribution of sub-experimental factors, by means of (a) randomization; (b) predefined criteria for the selection of study subjects (inclusion/exclusion criteria); (c) precise description, in the study protocol, of the procedures to which study subjects and investigators must strictly adhere; and (d) specific study designs (see below). Nevertheless, sub-experimental factors, known and unknown, cannot be fully controlled by the above-mentioned techniques. The influences that these uncontrollable factors exercise on the study results are collectively grouped in a global factor referred to as *chance*.

There are two main types of design for experimental studies: between-group and within-group.

1. In *between-group studies*, different subjects are assigned to different treatments. The conclusions are drawn by comparing independent groups of subjects. The most important design of this class is the randomized parallel group design.
2. In *within-group studies*, different subjects are assigned to different sequences of treatments, i.e., each subject receives more than one treatment. The conclusions are drawn by comparing subjects with themselves. The most important design of this class is the randomized crossover design.

Minimal Intervention Studies

This common type of studies somewhat falls in between the observational and the interventional approach. The overall framework is that of an observational study. However, the investigator is not completely hands off: a small degree of intervention is imposed by the study design, such as a blood draw or collection of other biological fluid, a noninvasive diagnostic procedure, or a questionnaire, hence the definitions “minimal intervention studies” or “low intervention clinical trials” [5]. These studies are often assimilated to observational studies, but individual informed consent is necessary outlining the risks and benefits of the additional procedure.

In the rest of this chapter, we will focus on clinical trials, which are the most commonly used type of experimental studies.

The Logical Approach to Defining the Outcome of a Clinical Trial

Let us assume we are the principal investigator of a clinical trial evaluating two treatments against obesity: A (experimental treatment) vs. B (control treatment). The sample size of the trial is 600 subjects (300 per treatment group). The primary

outcome variable (or end-point; see below), as defined in the protocol, is the weight expressed in kilograms after 1 month of treatment and is summarized at the group level in terms of mean. After over 1 year of hard work to set up the trial, recruit the patients, and follow them up, results finally come. These are as follows:

- Experimental treatment (A), mean weight: 104 kg
- Control treatment (B), mean weight: 114 kg

To simplify matters, we assume no imbalance of the average weight of the subjects at baseline and ignore the variability of the measurements, expressed by the standard deviation (clearly, in real life, both aspects are considered in the analysis and interpretations of results). After only 1 month of treatment, the group receiving the new treatment lost on average 10 kg, compared to the group receiving the traditional treatment. Most likely investigators would be inclined to rejoice at this finding. We want to believe that the observed difference is attributable to the new treatment and that we are on the verge of an important advancement in the management of obesity.

Unfortunately, this is not necessarily the case. In fact, three factors may contribute to different degrees to the observed difference: chance, bias, and treatment. The first two must be ruled out with a reasonable degree of certainty before attributing the outcome to the treatment.

The first question when confronting any observed difference between treatment groups must always be: can chance be the main reason for the observed difference? In clinical trials, the answer is given by a properly conducted statistical analysis. The famous *p* value expresses the probability of obtaining a difference as large as the one observed, or even larger, simply by chance, i.e., under the hypothesis of no true difference between groups (*null hypothesis*). If this probability is lower than a predefined (and totally arbitrary) threshold, traditionally fixed at 5% ($p < 0.05$), then the likelihood of chance being responsible for the result is considered small enough to be dismissed. Thus, the null hypothesis is rejected, and the alternative hypothesis of a true difference between groups is accepted.

Once chance is ruled out, the second question must be asked: can bias be the main reason for the observed difference? Bias is a systematic error that always favors one group over the other, thus potentially simulating a treatment effect. If two different scales were used for the two treatment groups, and the scale used for group A was malfunctioning and underestimating weight by 5–15 kg, then the observed difference between group A and group B would not be due to a treatment effect but to a measurement effect. This would be a typical, easily detectable example of bias. In most cases, the influence of bias is much more subtle and difficult to detect. The antidote against bias is in the study design features, including randomization and blinding (see below). In our example, clear rules on the validation and use of the scale(s) should be given in the protocol. The expert investigator will be reassured or concerned on the potential impact of bias by a careful review of the trial design and the way it was implemented. In addition, mathematical procedures exist to help detect bias.

Only after chance and bias have been excluded with reasonable certainty can the observed difference be attributed to the treatment. However, the logical approach to interpreting the study results is not over yet. A final, crucial question must be asked: is the observed treatment effect clinically or biologically meaningful? The clinically meaningful difference is an essential ingredient in the calculation of the sample size of a properly designed clinical trial. However, not all trials have a proper sample size calculation, and anyway the choice of the threshold for clinical significance (superiority or non-inferiority margin) is a highly subjective one. Biomedical journals are full of statically significant results of well-conducted trials which are of questionable clinical relevance.

Defining the Treatment Effect: From Measurement to Signal

The definition of the effect of a treatment is a conceptually complex process that starts with defining the aspects of interest of the disease and then proceeds in progressive steps to define, for each aspect of interest, the measurements to be performed on each patient, the variable that summarizes the measurements at the individual patient level (end-point), the variable that summarizes the measurements at the group level (group indicator), and, finally, the overall effect expressed in comparative terms between two treatment groups (signal) [1].

This process has several key contributors including physicians/biologists, statisticians, and regulatory, marketing, and pharmacoeconomic experts.

An example will help to understand the many choices that the researchers must make in this process. Suppose we are planning a clinical trial testing a new antihypertensive agent. The main objective of the study is to show the blood pressure-lowering capacity of the new agent (as opposed, e.g., to showing its impact on clinical outcomes such as myocardial infarction or stroke, a much more difficult task). We focus here on the main (primary) objective of the trial, but clearly the process should be repeated for each of the secondary objectives as well.

Step 1. Define the measurements (individual subject level). The researcher must painstakingly describe in the protocol the *what*, *how*, and *when* of each of the measurements selected to meet the objectives:

- For the *what*, we could choose diastolic blood pressure (DBP) or systolic blood pressure (SBP) or one of many other more sophisticated indicators of blood pressure. We choose DBP as the measurement to meet the main objective of the study.
- The *how* is equally important. Mechanical or electronic sphygmomanometer? Any particular brand? How far back is the last validation acceptable? Furthermore, the measurement procedure should be described in detail. Our decision is as follows: mechanical sphygmomanometer; one of three models deemed acceptable; calibration of instruments no more than 6 months before study starts; and DBP measurement to be taken on subject seated for at least 10 min, using dominant arm, each step precisely described in the protocol (e.g., inflate cuff, stop when no

pulse is detectable, then slowly deflate, stop when pulse detectable again, continue to deflate, stop deflation when pulse is again undetectable).

- Finally, the *when*. We decide that DBP is to be taken on day 1 (pretreatment baseline) and then on days 8, 14, and 28, in the morning between 8 and 10 a.m., before intake of study medication.

Each of these decisions should be made with science, methodology, and feasibility in mind. The measurement has to be scientifically sound, adequate to meeting the objective of the study, and feasible in the practical circumstances of the study. When this last requirement is ignored or underestimated by the researchers (as often happens), a poor outcome is very likely.

Step 2. From measurement to end-point (individual subject level). An *end-point* (also referred to as outcome variable) is a summary variable which combines all relevant measurements for an individual subject. Many end-points could be considered for the chosen measurement (DBP taken on days 1 [baseline], 8, 14, and 28). A few of the many possible options follow:

- Option #1: DBP difference from day 1 to day 28
- Option #2: time to DBP <85 mmHg
- Option #3: time to >5 mmHg reduction in DBP
- Option #4: mean (or median) of DBP values obtained at days 8, 14, and 28
- Option #5: lowest (or highest) DBP value over days 8, 14, and 28
- Option #6: responder/not responder (where, e.g., responder=subject with DBP <95 mmHg on day 28)

Again, the choice of the end-point is driven by many considerations, of which especially important are the objective of the study and the distribution of the end-point.

The choice of the number and timing of measurements is crucial. On one side, it is important to ensure that all measurements are indeed useful for the chosen end-point: for example, if the chosen option were number 1 (difference in DBP from baseline to day 28), then measurements on days 8 and 14 would have been useless. Measurements not contributing to the end-points are detrimental to the success of the study, as they only add to its complexity. On the other side, the possible presence of missing data should be taken into account: in fact, the primary analysis of a study is generally based on the intention to treat population, and this means that all patients being treated are to be included in the statistical analysis. An easy approach for assessing all patients is that of using the last observation carried forward: if, for example, the value at day 28 is missing, it could be replaced by the one observed at day 14; if no other observation is made before day 28, in case this is missing, it would be impossible to assess the patient in the intention to treat analysis.

In addition, there may be situations where the frequency of measurements must be increased. For option number 2 (time to DBP <85 mmHg), it would have probably been useful to plan more frequent measurements.

Let us assume that in our example the researchers chose option number 1 (disregarding the issue of missing data, for simplicity).

Step 3. From end-point to group indicator (treatment group level). We now move from the individual subject to the group of all subjects receiving a given treatment. A *group indicator* is a quantity which summarizes the data on the selected end-point for all subjects constituting each treatment group. In our example, where DBP difference from day 1 to day 28 was selected as the end-point, we could use the mean or the median of the DBP differences (depending on the distribution of such differences) as the group indicator. For our example, we choose the mean as the group indicator, assuming that the distribution of the DBP differences is symmetrical.

Step 4. From group indicator to signal (treatment group level). The *signal*, the final step of the process, is a summary quantity defining the overall effect of the experimental treatment at a group level and in comparative terms. Typically, the signal is expressed as either a difference or a ratio between group indicator A and group indicator B; occasionally, more complex signals are chosen, which may also involve more than two treatment groups (e.g., in dose-finding studies). In our example, we complete our journey by selecting the difference between treatment means of DBP changes from day 1 to day 28, as the signal for the primary objective of the trial.

As mentioned above, the whole process must be repeated for each of the objectives included in the protocol, primary as well as secondary. It must be emphasized that the conclusions of a clinical trial must be based on the predefined primary objective(s). Results from all other objectives, referred to as secondary or exploratory, will help to strengthen or weaken the conclusions based on the primary objective(s) and to qualify them with ancillary information but will never reverse them. Also, results from secondary objectives can be useful to generate new hypotheses to be tested in future trials.

Ideally, only one primary end-point (and corresponding signal) is selected to serve one primary objective for a given clinical trial. However, given the cost, duration, and complexity of a clinical trial, researchers are often tempted to include more than one primary objective and/or more than one end-point/signal for a primary objective, often with good reasons. Multiple primary end-points/signals come at a price: (1) larger sample size, due to the complex statistical problem of multiple comparisons, and (2) more difficult conclusions, as multiple primary end-points can give conflicting results.

Researchers can be more liberal with regard to the number of secondary end-points to be included in a study. However, it is still dangerous to include too many secondary end-points, as the complexity of the study and the volume of the data to be collected and checked for accuracy (or “cleaned”) will increase very quickly as the number of end-points increases, and the study will soon become unmanageable. The risk is that the study will “implode” because of excessive complexity. Such a frustrating outcome is far from infrequent and is typically caused by an excessive number and complexity of secondary end-points.

The primary end-point/signal must have external relevance and internal validity. *External relevance* is the ability to achieve the practical goals of the study, such as

regulatory approval, health economic justification, differentiation from current treatment, etc. *Internal validity* is the ability to draw valid conclusions on the causal relationship between treatment and the desired effect; it is accomplished by appropriate design and appropriate statistical analysis.

Surrogate and composite end-points are special types of end-points often used in clinical trials. *Surrogate end-points* [6] are instrumental or laboratory measurements used to substitute for clinical outcomes. Examples of surrogate end-points are diastolic blood pressure as surrogate for cardiovascular accidents (myocardial infarction, stroke, etc.) or the blood level (*titer*) of a specific antibody as surrogate for a vaccine's ability to protect against a given infection. The advantage of a surrogate end-point is that it allows smaller and shorter trials compared to those needed for the corresponding clinical end-point. This is especially important for rare events such as a rare infection prevented by a vaccine, for which clinical outcome trials are practically undoable. *Composite end-points* combine in one score the outcome of multiple individual end-points; typical examples are quality of life questionnaires. The advantage of a composite end-point is that it overcomes the issue of multiple comparisons.

The big hurdle for both surrogate and composite end-points is that they must undergo proper validation, a long and complex process, before being used in a clinical trial. Unfortunately, validation is often suboptimal, thus undermining the validity of the trial results and conclusions.

Defining the Study Sample

The sample is the group of subjects on which the study is performed. The choice of the sample requires qualitative and quantitative considerations ([1], Chap. 6). Among the qualitative aspects of the sample selection, crucial is the need to ensure that the sample is representative of the population to which one wants to extend the conclusions of the study. In Phase I, in general, representativeness is not required: trials are typically conducted in healthy volunteers, although, as mentioned at the beginning of this chapter, there are increasingly frequent exceptions, where Phase I trials are conducted in patients. The criteria qualifying a volunteer as *healthy* are far from obvious: if a long battery of clinical and laboratory tests are conducted and results within the normal range are required for every single test, almost nobody would be enrolled in the study. Phase II studies are typically conducted in patients with the disease in question, clearly more representative of the true target population than healthy volunteers. However, selection criteria in the initial stage of Phase II (Phase IIA) are typically strict, with exclusion of the most serious or atypical forms of the disease, as well as of most concomitant conditions and use of many concomitant medications; thus, again, representativeness with respect of the true population is limited, and results are likely to be better than what would be seen in real life. It is in the Phase IIB definitive dose and schedule finding trials and in Phase III that the sample must be as representative as possible of the true population. Clearly, complete representativeness will never be accomplished because, no matter

how large a Phase III trial, it will always be conducted in a small number of countries and institutions, with inevitable bias in socioeconomic status, racial mix, nutritional habits, etc. It is essential not to have too restrictive inclusion and exclusion criteria, i.e., allow entry to the *average* patient. For example, if we are conducting a Phase III study in chronic obstructive pulmonary disease (COPD), it would be wrong to deny entry to patients with cardiovascular conditions, as these are very common in COPD patients.

The quantitative aspect of the sample selection is equally crucial: how large should the size of the sample be? The sample must be large enough to allow the detection of the treatment effect, separating it from the natural variability of the phenomenon, with an acceptable degree of certainty. But how does one determine this? The decision on the sample size of a study is considered by many an exclusively statistical matter. This is not the case at all: there are of course formulas used to calculate the sample, which may change depending on the end-point, the signal, and the study design; however, the most difficult aspects of the sample size determination are the decisions on the assumptions behind the formulas, which require a close collaboration between the physician (or biologist), the statistician, and the expert in operational matters. Briefly, decisions on the following eight key assumptions are necessary for the sample size calculation (note: for each, it is assumed that all conditions other than the one being discussed are equal):

1. The design of the study and the kind of comparison to be investigated: for example, parallel group designs require more subjects than crossover designs, and non-inferiority/equivalence studies require more subjects than superiority studies (see below).
2. The magnitude of acceptable risk of type I and II errors: the smaller the risk we are willing to accept of obtaining a false-positive result (type I error, i.e., there is no true treatment difference, but the test erroneously detects a difference) and a false-negative result (type II error, i.e., there is a true treatment difference, but the test erroneously does not detect it), the greater the sample size. One can reduce the level of the type I error at the expense of the level of the type II error and vice versa, while maintaining approximately the same sample size, but if we want to reduce both types of errors at the same time, the sample size will need to be increased.
3. The magnitude of the signal (threshold of clinical relevance for superiority trials and margin of clinical irrelevance for non-inferiority/equivalence trials, see below): the smaller the difference between treatments we are prepared to accept as clinically relevant (or irrelevant), the greater the number of subjects we need.
4. The number of primary end-points and signals: the more primary end-points and signals we have in our protocol, the greater the sample size, as we need to adjust it upwards to account for multiple comparisons. Multiple treatment arms typically (although not necessarily) contribute to multiple signals.
5. The type and variability of the primary end-point(s): the greater the variability (intrinsic or induced by the measurement process), the more subjects are required

- to detect a given threshold of clinical relevance or to prove equivalence/non-inferiority given a pre-defined irrelevance margin.
6. The type of hypothesis: we will need more subjects for a bidirectional hypothesis (i.e., the study hypothesis is that A and B are different, and this difference can be in either directions) than for a unidirectional hypothesis (i.e., the hypothesis under study admits a difference only in one direction).
 7. The type of statistical test: for example, in general, parametric tests require fewer subjects than corresponding non-parametric tests.
 8. The expected rate of premature discontinuations: the more the discontinuations not contributing to the primary end-point(s), the larger the sample size.

Defining the Study Treatments

In the planning of a clinical trial, one should carefully define the treatments, both those that are the object of the experiment, referred to as study treatments, and those that are not, referred to as concomitant treatments ([1], Chap. 7). The study treatments include experimental and control treatments:

- *The experimental treatment* is the main object of the study. In general, only one experimental treatment is investigated, but there are situations where it is legitimate to test more than one in the same study (e.g., different combinations with other treatments or different doses). Experimental treatments can be new pharmacological preventive or therapeutic agents, but also surgical procedures, psychological/behavioral treatments, and even logistical/organizational solutions (e.g., the use of normal hospital wards for myocardial infarction patients replacing intensive care).
- *The control treatment* should be the standard of care against which the experimental treatment is assessed by comparison. If the medical community or the regulatory authority does not recognize a standard of care with proven positive benefit–risk ratio, the control treatment should be a placebo or no treatment (in cases where the use of placebo is not considered viable, e.g., intravenous procedure in young children). A *placebo* is an inactive treatment, identical to the experimental treatment in every aspect except for the presumed active substance. If a recognized standard of care does exist, then the control treatment should be the recognized active treatment. However, there are many intermediate situations in which there is no agreement as to whether or not a standard of care exists, for example, because common practice is based on old or unreliable data and/or there are multiple accepted best practices. In these situations, some complex practical and ethical dilemmas must be addressed, concerning whether or not placebo should be used and what standard should be picked as the best comparator. It is not uncommon that both placebo and an active comparator are required by a regulatory authority for definitive dose-finding and pivotal Phase III trials and that more than one active comparator is chosen in postmarketing Phase IV profiling trials.

- *The concomitant treatments* are drugs or other forms of treatment that are allowed during the study but are not the object of the experiment. Concomitant treatments at times represent useful end-points, for example, the amount of rescue bronchodilator taken each day in asthma trials or the time to intake of a pain killer following tooth extraction in trials testing an analgesic/anti-inflammatory agent. When the interaction between an experimental and a concomitant treatment is an objective of the trial, the latter should also be considered experimental.

For each type of treatment, the researcher must be very detailed in the protocol in describing not only the type of treatments but also their mode of administration (route, frequency, time, special instructions) and the method of blinding (see below). These choices are of critical importance as they directly influence both the conduct and the analysis of the study.

A critical dilemma for investigators concerns the decision of how many study treatments to investigate. On the one side, multiple study treatments may make the study more interesting and scientifically valuable. On the other side, multiple comparisons will require a sample size increase, more complicated drug supply management (blinding, packaging, shipment) and study conduct, statistical analysis, and interpretation of results. Unfortunately, no easy solution can be offered as to the number of treatments to be included in a trial. There are experimental designs that facilitate multiple study treatments, such as factorial and dose escalation designs and special designs to assess dose-response relationship (see below). Studies evaluating combinations of different treatments (with or without different dose levels) can also have multiple study treatments. Vice versa, large confirmatory Phase III trials are rarely successful with more than three study treatments.

Other difficult choices concern concomitant treatments: should we be liberal or strict in allowing concomitant treatments? Many investigators are afraid that concomitant treatments may interfere with the measurements and confound the results. This may well be the case. However, if a concomitant treatment is broadly used by patients in real-life situation (e.g., inhaled corticosteroids are used by almost all asthma patients), there is little practical value in sanitizing results by eliminating such treatments from the study. In general, it may be acceptable to be relatively conservative with concomitant treatments in Phases I and IIA (but not too much), whereas in Phases IIB (definitive dose-finding studies) and III, it is necessary to reflect real life as much as possible by being quite liberal with concomitant treatments.

Superiority Versus Non-inferiority

The comparison between treatments can be performed with two different objectives: (1) demonstrate the superiority of the new treatment over the standard one (or placebo), and (2) demonstrate the equivalence or, more frequently, the non-inferiority of the new treatment compared to the standard one.

Clinical trials with the former objective are called *superiority studies*; those with the latter objective are called *equivalence* or *non-inferiority studies* ([1], Chap. 11). The difference between equivalence and non-inferiority is that in equivalence studies, the aim is to demonstrate that the new treatment is neither inferior nor superior to the standard one, while in non-inferiority studies, the aim is only to demonstrate that the new treatment is not inferior to the standard one (if it is better, it is considered still not inferior).

Equivalence/non-inferiority studies are performed when:

- It is sufficient to demonstrate that the new treatment is similar to the standard one in terms of efficacy, because the new treatment has other advantages over the standard, for example, a better safety/tolerability profile, an easier schedule or route of administration, or a lower cost.
- It is an advantage to have several therapeutic options, based on a different active principle and/or a different mechanism of action, even if their efficacy and safety are on average about the same; indeed, the individual patient may respond better to one treatment than to another, may be allergic to a particular treatment but not to the other, may develop tolerance to one specific compound, and so on.

Equivalence studies play an important role in the development of so-called generics, or identical copies of marketed drugs no longer protected by a patent. To register the new generic drug, one needs to demonstrate that key pharmacokinetic and/or pharmacodynamic variables of the new treatment are equivalent, i.e., neither superior nor inferior, to the standard one.

The choice between the objective of demonstrating superiority and that of demonstrating equivalence/non-inferiority has a major impact on study planning, definition of the clinical threshold, sample size calculation, and statistical analysis. A common mistake is to plan and analyze an equivalence/non-inferiority study as if it were a superiority study. Instead, different methods must be used.

When planning a superiority study, the investigator must select a priori a threshold of clinical relevance (superiority margin), i.e., the smallest difference between treatments, judged as clinically meaningful. On the other hand, in an equivalence/non-inferiority study, the investigator must select a threshold of clinical irrelevance (equivalence or non-inferiority margin), i.e., the largest difference between treatments, judged as clinically irrelevant. A guidance document under the patronage of the EMA Committee for Human Medicinal Products (CHMP) on the choice of the equivalence/non-inferiority margin is available at http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003636.pdf.

In superiority studies, the null hypothesis, which we seek to reject in the traditional statistical testing, is that there is no difference between treatments. Vice versa, in equivalence/non-inferiority studies, the null hypothesis is that the treatments are different. In other words, in equivalence/non-inferiority studies, the system of hypotheses is inverted compared to superiority studies.

In superiority studies, the statistical test is used for decision-making. If the test is statistically significant, we can conclude that the difference observed between

treatments is unlikely due to chance, while if the test is not statistically significant, we can conclude that the difference is likely generated by chance.

The analysis of equivalence/non-inferiority studies must be based on confidence intervals. Assuming that we use the mean as the group indicator, and the difference between means as the signal, we must calculate the 95% confidence interval on the observed mean difference between the treatments (note that the 95% level for the confidence interval is set conventionally, just like the 5% level for the statistical test). Equivalence between the treatments is demonstrated if the two-sided confidence interval is entirely included within the equivalence margin. To grasp the meaning of this, it helps to recall that the two-sided *confidence interval* at the 95% level on the mean treatment difference is defined as the set of values of the estimated mean treatment difference which includes the true value of the mean treatment difference with a probability equal to 95%. Therefore, when the 95% confidence interval on the mean treatment difference is entirely included within the equivalence margin, there is a high probability (in fact equal to 95%) that the true value of the mean treatment difference is a clinically irrelevant difference between the treatments. Likewise, non-inferiority of one treatment vs. another is demonstrated if the one-sided 97.5% confidence interval of the difference between the two treatments is entirely below (or above) the non-inferiority threshold. As mentioned earlier, the equivalence/non-inferiority study generally requires a greater number of subjects compared to the corresponding superiority study with the same design, primary end-point, and experimental conditions. In fact, all other conditions being the same, the treatment differences on which the sample size calculation is based are typically smaller in an equivalence/non-inferiority study than in a superiority study. In addition, while in a superiority study we bet on treatment differences bigger than the threshold of clinical relevance, in an equivalence/non-inferiority study, we bet on treatment differences smaller than the equivalence margin: this reduces power of the study and therefore increases the sample size.

In superiority studies, the better the quality of the study, the greater the likelihood of detecting a difference between the study treatments, when it exists. Therefore, it is to the advantage of the researchers to plan and conduct the study in the best possible way. In equivalence studies, since the poorer the quality of the study, the lower the likelihood of detecting differences, if any, the researchers have no incentive to conduct the study in the best possible way. In other words, quality is even more important in equivalence and non-inferiority studies than in superiority studies.

The two treatments under comparison could be equivalent or one could be non-inferior to the other simply because both are ineffective. This is the main reason why in equivalence/non-inferiority studies, regulatory authorities recommend including a comparison with placebo whenever ethically acceptable, to confirm that the presumed active compounds separate from placebo, i.e., are indeed active (see guideline ICH E12). With a placebo arm included in the study, the equivalence/non-inferiority study has its own internal validity, i.e., it allows one to draw valid comparative conclusions. However, often the comparison to an active control is conducted because it is unethical to use the placebo. Theoretically, when there is no placebo group in the study, it is possible to use the placebo groups of the studies of

the active control as an indirect reference. The equivalence/non-inferiority study must be as similar as possible to these placebo-controlled superiority studies, with respect of study design and conduct (treatment duration, end-points, characteristics of the population, etc.). In this way, if the equivalence/non-inferiority study is properly performed, one should theoretically obtain for the active control results similar to those obtained in the previous superiority studies against placebo, and, under such conditions, one should be able to judge whether or not the treatments under comparison are both efficacious. Unfortunately, this reasoning is theoretical and often distant from reality. It is common in real life that the clinical studies, in which the efficacy of the active compounds has been tested, have different protocols and different results, so that the issue of which one to choose arises. Then, the comparison between the results of the reference placebo-controlled studies and those of the active-controlled equivalence/non-inferiority study has all the weakness of a comparison with a historical control, which ultimately makes it impossible to guarantee bias-free comparisons.

Experimental Designs

Definitions and Basic Concepts

The experimental design is the logical structure of an experiment. In the section above, the experimental and sub-experimental factors have been defined. By means of the experimental design, the researcher controls the experimental factors, typically the study treatments being compared and some of the most important sub-experimental factors, typically key protective and/or risk factors. There are two main objectives of the experimental design: (1) minimize the systematic error or bias, systematically favoring (or damaging) one of the treatment groups being compared, and (2) minimize the random error and consequently reduce the variability.

Bias is minimized mainly by means of randomization, blinding, and *a priori* definition of procedures and methods, as described below. By means of the experimental design, we try to deconstruct the total variability in pieces that are due to known factors (experimental and sub-experimental factors). The remaining unexplained part of variability, i.e., the part that cannot be attributed to any known factor, is referred to as chance or as *residual variability*. Residual variability is used for carrying out the statistical tests and for computing the confidence intervals on the estimate of the treatment effect. The smaller the residual variability, i.e., the variability attributable to chance, the bigger the power of the statistical test, and the greater the precision of the estimates.

A good experiment must allow comparisons that are *bias-free*, which means that systematic errors are absent or negligible and *precise*, which means that the residual variability is small. The different designs must be evaluated for these two main characteristics. In addition, in choosing the study design, the researcher must always keep simplicity in mind: simplicity of study conduct, data analysis, and interpretation of results. The studies which are too complex are not feasible, and often complexity is the cause of study failure.

Before-After Comparisons in a Single Treatment Group

The simplest form of experimental design is based on the before–after comparison in a single group of subjects, i.e., without a separate control group ([1], Chap. 8). *Before* and *after* refers to the beginning and the end of treatment, respectively.

This study design is indeed simple and close to the way the physicians are used to making decisions. However, there are numerous sources of bias in this design that make the *before* setting not comparable with the *after* setting. These are as follows:

- Temporal variations of the disease
- Temporal variations of personnel, equipment, and context
- Statistical regression to the mean, a phenomenon by which a variable having an extreme value (i.e., much greater or much smaller than the mean of its distribution in the population) in the first measurement will tend to be closer to the mean in subsequent measurements in the absence of any treatment effect [7, 8]
- Learning effect
- Psychological effect, caused by the awareness of being treated

These different types of bias undermine the reliability of conclusions. In most cases, with this design, bias favors the *after* over the *before*, thus simulating a treatment effect when such effect does not exist or amplifying it when it does exist. There are some exceptions, especially when serious diseases with predictable time course are studied; yet, in general, the before–after comparison in a single group of subjects is a severely biased design, which should be avoided.

Among the kinds of bias reported above, the *regression to the mean* is probably the least obvious. Regression to the mean stands literally for “turning back to the mean.” In clinical trials, this phenomenon occurs every time a group of subjects is selected based on *extreme* values of a variable, and that same variable is measured again in the same subjects at a later point in time. The mean of the values obtained in the second measurement will likely tend to be less extreme compared to the mean of the values obtained in the first measurement and, therefore, will be closer to the population mean. This probabilistic phenomenon will always occur, even in the absence of any treatment effect. Therefore, in a simple before–after study, if the variable used for the selection of patients is also used as an end-point, the effect of treatment will be confounded with the regression to the mean effect, and it will be very difficult to separate one from the other. If the researcher performing such a study ignores the possible effect of the regression to the mean, and attributes the observed improvement to the treatment, he/she will interpret the results in a biased way.

Antidotes Against Bias: Randomization, Blinding, and a Priori Definition of Analysis

The only way to avoid these problems is that of using study designs with one or more concurrent comparative groups. Three key procedures are used to minimize bias in experimental studies: randomization (against selection bias), blinding

(against assessment bias), and a priori definition of the statistical analysis, i.e., before the results are known (against the analysis bias) ([1], Chap. 3).

Randomization is the assignment of subjects to treatments (or sequence of treatments) with predefined probability and by chance. The basic point is that the assignment of an individual subject cannot be predicted based on previous assignments. Randomization is not haphazard assignment. In fact, with a haphazard assignment of subjects to treatments, there would be no predefined probability, and, most likely, subconscious patterns would influence the assignment. Randomization is also not systematic assignment (e.g., patients enrolled on odd days are assigned to A, on even days to B); in fact, by using such a method, there would be no chance assignment.

Randomization minimizes selection bias for known and unknown factors. It has to be taken into account that “no selection bias” does not necessarily mean “no imbalance” for key prognostic factors (e.g., age), especially in small trials. A baseline imbalance can occur also when using randomization to allocate subjects to treatments and can be problematic, for example, it may cause unequal regression toward the mean between the two groups being compared. Special forms of randomization (see below) may reduce the likelihood of large imbalances in small trials.

The other important role of randomization is that it legitimizes the traditional (frequentist) approach to statistical inference. In fact, the foundation of the frequentist approach is the assumption that the sample is extracted randomly from the population. As discussed earlier in this chapter, this does not happen in real-life clinical trials. The sample of patients enrolled in a trial is never a random representation of the overall population who will receive the treatment. Randomization reintroduces the random element through the assignment of patients to the treatments.

In the planning stage of a randomized clinical trial, the randomization list is generated according to predefined rules. For each randomization number in the list, a code containing a sequential numerical code is generated and placed on the pack containing that patient’s treatment. At this point, the randomization process can be directly executed by the investigator, by following the order of assignment of the pack codes (first pack code, i.e., the code with the lowest numerical code, must be assigned to the first eligible patient, second pack code to the second patient, and so on). The logistics of randomization can be very complex and is beyond the scope of this chapter.

There are numerous methods of random allocation of subjects to treatments. We will briefly cover the following: simple randomization, randomization in blocks, stratified randomization, adaptive randomization, and cluster randomization.

In the *simple randomization*, each subject has the same probability of receiving each of the study treatments or sequence of treatments. When the sample of a study is large, simple randomization will most likely assign almost the same number of subjects to each treatment group, through the effect of chance alone. The situation can be completely different in small studies. In such studies, to avoid relevant inequalities in the sizes of the treatment groups, the so-called *randomization in blocks* is used. The assignment occurs in subgroups, called blocks. Each block must have a number of subjects equal to the number of treatments or to a multiple of this number. Furthermore, within each block, each treatment must appear a predefined number of times. It should be noted that this randomization method obtains

treatment groups of similar size not only at the end of enrolment but also throughout the whole enrolment process.

Stratified randomization takes into account one or more prognostic (protective or risk) factors. It allows for the selected prognostic factor(s) to be evenly distributed among the treatment groups. The stratified randomization requires that each preselected factor be subdivided in exhaustive and mutually exclusive classes. For gender, for example, this is easily done by considering the two classes of males and females. The classes are called strata. When taking into account multiple prognostic factors, the strata originate by combining the classes of all factors. An independent randomization list is generated for each stratum, and a subject is assigned to a treatment according to the randomization list of the stratum to which he/she belongs.

In the *adaptive randomization methods*, the allocation of patients to treatments is based on information collected during the study. This information can be related to a protective/risk factor, with the goal of minimizing the imbalance between groups with respect to such a factor or to the accumulating results for a preestablished endpoint, generally the primary one: in this case, the assignment of a new patient is based on a probabilistic rule which favors the group showing the best result, at the time the new patient is ready to be randomized.

In *cluster randomization*, the unit of randomization is not the individual study participant but the cluster. A cluster is a group of study participants with a common geography, for example, subjects attending the same physician or hospital or living in the same village or city block [9]. This type of randomization is generally used in large studies, where the main focus is not the individual patient but the community, for example, when the objective is to evaluate the impact of a vaccine on the community, including non-vaccinated subjects (so-called herd effect) or the impact of a new standard of care on health outcomes. Another reason for using the cluster randomization is when there is a significant risk of contamination in the study, i.e., when some aspects of one intervention may be adopted by individuals that were randomized to another intervention, for example, in a clinical trial evaluating two different treatment strategies, patients waiting to be visited may discuss among themselves the respective strategies and decide to adopt the strategy to which they were not randomized.

Blinding (or masking) is the process by which two or more study treatments are made indistinguishable from one another. Blinding protects against various forms of bias, most important of which is the assessment bias.

The ideal situation would be that the study treatments differ with respect to the presumed active component but are otherwise identical in weight, shape, size, color, taste, viscosity, and any other feature that allows identifying the treatment. This would be a perfect double-blind, where all study staff and patients are blinded. However, in practice, often one has to accept a lower level of blinding, for example:

- Observer-blind: the patients and the study staff assessing the patients are blinded, whereas the staff administering the treatments are not.
- Single-blind: only patients are blinded.
- Open-label: no one is blinded.

The lower the level of blinding, the higher the risk of bias.

The *randomized, double-blind clinical trial* with concomitant control groups is the type of study that is most likely to achieve bias-free results, minimizing the impact of errors systematically favoring or penalizing one treatment over another.

Non-randomized and non-blinded studies generally cannot achieve a similar degree of methodological strength. However, one should not be dogmatic: a comparison before–after in a single group can be the best way to start the clinical development of a compound intended to treat a cancer with rapid and predictable outcome, especially for ethical reasons. An open-label randomized design can be stronger than a double-blind study, if the latter results in poor compliance to study medication by patients, for example, because the mechanism for blinding the treatments is too complex. The experienced clinical researcher will try to get as close as possible to the standard of the randomized, double-blind design. However, he/she will also give due consideration to the practical, logistic, technical, and economic aspects in making the final decision, keeping always in mind the value of simplicity. Finally, he/she will make a transparent report on the methods followed and on the reasons for the choices made at the time of presenting the results.

Parallel Group and Crossover Designs

There are two main categories of comparative study designs for clinical trials ([1], Chap. 10):

1. The parallel group designs in which there are as many groups as treatments, all groups are treated simultaneously, and every subject receives only one of the study treatments (or a combination tested as a single treatment).
2. The crossover designs in which each subject receives more than one study treatment in sequence but only one of the possible sequences of study treatments.

Parallel Group Designs

The *completely randomized parallel group design* is the simplest. Let us indicate the experimental factor, i.e., the treatment with T , and assume it has k levels, i.e., T_1, \dots, T_k . The levels can be different compounds or different doses of the same compound. Each level T_i of T is replicated on n_i subjects. The subjects are assigned in a random way at the different levels of T . The design matrix is shown in Table 4.1.

Table 4.1 The parallel group design matrix

T_1	T_2	...	T_k
Y_{11}	Y_{21}		Y_{k1}
...
Y_{1n_1}	Y_{2n_2}		Y_{kn_k}

In this design, it is possible to estimate only the treatment effect. Accordingly, the total variability is divided into two components: the part explained by the treatment and the part unexplained by the treatment, the latter totally attributed to chance.

The most important advantage of this study design is its simplicity, concerning both the study conduct and the statistical analysis. Its biggest disadvantages are as follows:

1. The variability of the end-points within each group is the biggest among all the experimental designs; therefore, all other aspects being equal, the statistical tests have less power, and the treatment estimates are less precise.
2. By chance, the groups under comparison may be imbalanced at baseline with respect to important sub-experimental factors (e.g., twice as many female subjects in one group). Baseline imbalances can be to some extent *adjusted* by statistical procedures; however, major baseline imbalances for important prognostic/risk factors render the groups not comparable.

It should be noted that, if the study is large enough, both disadvantages mentioned above are contained to acceptable levels and the advantages prevail. Thus, this design is often used for pivotal Phase III clinical trials.

Two methods can be used to reduce variability without increasing the sample size. These are as follows:

1. Group the subjects with respect to common characteristics by generating so-called strata or blocks.
2. Replicate the measurements on each subject.

In the *stratified parallel group design*, the researchers will select few (typically one or two) particularly important sub-experimental factors with well-known prognostic value on the end-point for which they want to avoid any relevant baseline imbalance. The levels of the considered sub-experimental factor(s) are categorized in classes (*strata*). Let us assume we choose age as the prognostic factor for which we want to ensure balance at baseline, which we then categorize in four strata: children (6–11 years of age), adolescents (12–17), non-elderly adults (18–64), and elderly adults (65 and above). Let us indicate the treatments with T and the strata with S ; the four strata are: S_1 , S_2 , S_3 , and S_4 . Each level T_i of T and stratum S_j of S is replicated on n_{ij} subjects. The subjects are randomly assigned to the different treatments, separately and independently within each individual stratum. As a consequence, by design, the strata are balanced between treatments. The design matrix of the stratified parallel group design is shown in Table 4.2.

In this design, it is possible to estimate the following effects:

- Main treatment effect, i.e., treatment effect without considering the stratification factor.
- Main effect of the stratification factor (in our case, age group), i.e., without considering the treatment.

Table 4.2 The design matrix of the stratified parallel group design

	T_1	T_2	...	T_k
S_1 children	Y_{111}	Y_{211}		Y_{k11}

	$Y_{11}n_{11}$	$Y_{21}n_{21}$		$Y_{k1}n_{k1}$
S_2 adolescents	Y_{121}	Y_{221}		Y_{k21}

	$Y_{12}n_{12}$	$Y_{22}n_{22}$		$Y_{k2}n_{k2}$
S_3 non-elderly adults	Y_{131}	Y_{231}		Y_{k31}

	$Y_{13}n_{13}$	$Y_{23}n_{23}$		$Y_{k3}n_{k3}$
S_4 elderly adults	Y_{141}	Y_{241}		Y_{k41}

	$Y_{14}n_{14}$	$Y_{24}n_{24}$		$Y_{k4}n_{k4}$

- Interaction between the two effects: there is an interaction between the treatment and the stratification factor when the effect of the treatment on the response changes across the different levels of the stratification factor and, likewise, the effect of the stratification factor changes across the different levels of the treatment factor.

Accordingly, in this type of design, the total variability is divided into four parts: the part explained by the treatment, the part explained by the sub-experimental factor(s), the part explained by the interaction between the treatment and the sub-experimental factor(s), and the residual variability attributed to chance (each computed by *averaging* the estimates of the variability calculated within each stratum). If the factor used for the stratification is a real prognostic factor, the residual variability of the stratified design is smaller than the residual variability of the completely randomized design. Therefore, the former provides more powerful tests and more precise estimates of the treatment effect than the latter. However, the stratified design is more complex than the completely randomized design, and this aspect should be carefully considered when choosing between the two designs.

Another design based on grouping the subjects with respect to common characteristics is the *randomized block design*. In this kind of design, as many subjects as the number of study treatments or a multiple of this number are “grouped” based on predefined prognostic factors. These groups of subjects are called “blocks.” The subjects within each “block” are randomized to the study treatments (randomization in blocks). The number of blocks to be randomized depends on the total sample size. If only two treatments are to be compared, the blocks have size of 2 or a multiple of 2. The case with a block of 2 is referred to as the *matched-paired design*, which is the variant of the randomized block design most often used in clinical trials. Often the randomized block design is used in clinical trials when the time of enrollment is one of the factors that should be controlled for. Time can be a known prognostic factor (e.g., in asthma, Reynaud syndrome) or just a sub-experimental factor with unknown prognostic value (e.g., in a study in which high turnover of

personnel is expected). In any case, with the randomized block design, the temporal changes are balanced between the treatment groups at regular intervals: the smaller the block, the shorter the intervals.

Crossover Designs

The *crossover design* is based on the concept that every subject is used as his/her own control. As already said, this implies that each subject receives more than one treatment ([1], Chap. 10).

We shall start with the so-called *two-by-two crossover design*, characterized by the use of two treatments in two periods. Suppose we have two treatments A and B. A is administered to the subjects of one group as first treatment (period 1), followed by B (period 2). Vice versa, B is administered as first treatment to the subjects of the other group (period 1) and then followed by A (period 2). Each of the two groups, AB and BA, is called *sequence*. In this design, the subjects are randomized to the sequences, not to the treatments. The design matrix of a *balanced crossover design* (i.e., a crossover design with the same sample size in each period and each sequence) is shown in Table 4.3.

The generic response Y_{ijr} is identified by three indices: i (sequence), j (period), and r (subject).

In a crossover design, it is possible to estimate the following effects:

- Treatment effect.
- *Period effect*, which is the effect of time, for example, spontaneous progression or improvement of the disease, seasonal or cyclic changes of the disease.
- Interaction between treatment and period.
- *Carry-over effect*. The carry-over is the continuation of a treatment effect from one period into the following period; a carry-over effect is a problem and can be detected only when it is unequal between treatments (e.g., the continuation of the effect of A is longer or greater than the continuation of the effect of B in the following period).
- *Sequence effect*, which is the effect of the entire sequence of study treatments on the end-point. It can be estimated by treating the crossover design as a parallel group design.
- *Subject effect*, which is due to the peculiar characteristics of each individual. It can be estimated by considering the repeated measures on each given subject. In very simple terms, if the subject effect is strong, all values measured on the same subject will be similar to one another.

Table 4.3 The crossover design matrix

		Sequence 1 (AB)	Sequence 2 (BA)
Period	1.	A: $Y_{111}, Y_{112}, \dots, Y_{11n}$	B: $Y_{211}, Y_{212}, \dots, Y_{21n}$
	2.	B: $Y_{121}, Y_{122}, \dots, Y_{12n}$	A: $Y_{221}, Y_{222}, \dots, Y_{22n}$

Actually, in the two-by-two crossover design, generally only the treatment, the period, and the carry-over effects are considered.

In the crossover design, the subject and the sequence effects have a very limited interest per se. However, quantifying these effects is useful to reduce the residual variability.

Presence of a significant carry-over effect is detrimental for the interpretation of the treatment effect. To attenuate, and possibly eliminate, the carry-over effect, often the so-called washout period is included between the two treatment periods, i.e., an additional period where the patients receive no treatment. However, also the use of the washout period cannot guarantee absence of the carry-over effect.

The statistical analysis typically starts with the test of this effect. If this is statistically significant, the solution generally applied is that of taking into account only the observations from the first period and discarding the ones from the second one. The study is then analyzed as if it were a parallel group design. Unfortunately, in most cases, the sample size is insufficient for a parallel group design; thus, in practice, a significant carry-over effect results in a failed study. If no statistically significant carry-over effect is detected, all data are considered in the analysis, and therefore both the period and the treatment effects are estimated. It should be noted that the test for the carry-over effect is often underpowered, thus unequal carry-over may go undetected.

The statistical test for the treatment effect and the one for the period effect are based on the within-subject component of the total variability, while the test for the carry-over effect uses the between-subject component of the total variability.

The observations on different patients are independent; the ones on the same patient are not, i.e., are correlated. The fundamental reason to use the crossover design instead of a parallel group design is that measurements taken on the same subject for more than one study treatment are expected to be correlated and therefore to result in a smaller total variability. This, in turn, results in a smaller sample size or more precise estimates of the effect for a given sample size. It should be noted, however, that this is true only when the measurements on the same subject are highly correlated and this is not a given (the measurements on the same subject are correlated by definition, but this correlation may be low).

In summary, if the measurements on the same subject are highly correlated, the crossover design generates a test on the treatment effect more powerful, i.e., requiring a smaller sample, than the test for the parallel group design, based on between-subject comparisons.

The most important advantages of the crossover designs are as follows:

- The concept that every subject is used as his/her own control is close to the common way of making judgments.
- If the observations on the same subject are highly correlated, the sample size is reduced compared to the matching parallel group design.

These advantages must be balanced against the following challenges:

- The crossover design is more complex for the logistical aspects than the parallel group design.

- The treatment effects must be fully reversible by the time the next treatment starts. Hence, crossover designs are not suitable for curative or disease-modifying treatments.
- The duration of the treatments must be relatively short; otherwise, the overall duration of follow-up in an individual patient will be untenable (washout periods must be added as well!).
- The statistical analysis requires more assumptions compared to the parallel group design and cannot cope well with dropouts.
- An unequal carry-over effect will generally invalidate the study.

Generally, the crossover design makes use of a simple randomization. However, the stratified and the randomized block crossover designs, which use the corresponding methods of randomization, do exist. The statistical analysis becomes more complicated.

In theory, the *complete crossover design* (where all possible sequences are used and each subject receives one sequence containing all of the treatments under study) is applicable to any number of treatments. However, if the treatments are more than three, the experiment becomes very complex. For example, if the treatments are 4, there are 24 possible sequences. Therefore, generally, with more than three treatments, only incomplete versions of the crossover design are used.

Two variants of *incomplete crossover designs* are possible. One is based on the use of a selection of complete sequences, for example, with 4 treatments, only 6 of the 24 possible sequences are used. The other is based on the use of incomplete sequences, i.e., the subjects do not receive all the treatments under study.

If there is reasonable certainty that the period effect is irrelevant, there is no need to guarantee balance among the sequences that have been included in the study. If instead there is no reasonable certainty that the period effect is irrelevant, the researcher must assure balance among the sequences by means of a special form of crossover called *Latin square design*. The main feature of the Latin square design is that any treatment appears only once in each row (representing the sequence) and only once in each column (representing the period). With three treatments, referred to as A, B, and C, there are two possible Latin square designs, as illustrated in Table 4.4.

In order to use the Latin square design, the sample size must be a multiple of the number of treatments in each sequence (in this case three).

Table 4.4 The Latin square design matrix

		Period
		1 2 3
Sequence	1	A B C
	2	B C A
	3	C A B

In the incomplete crossover design characterized by incomplete sequences, more treatments than periods are included in the design. For example, a design with three treatments and two periods can be obtained by removing one column from the Latin square designs shown above. This design maintains some level of balance: each treatment appears in the same number of sequences, and each treatment appears once in each period.

Variants of Parallel Group and Crossover Designs

Variants of the more frequently used designs exist, which are useful in special situations ([1], Chap. 11). Because of space limitations, we will mention just a few examples. In Phase I, the controlled *dose-escalation designs* are frequently used. These designs, in which each patient receives only one dose level, allow the evaluation of higher doses, only once sufficient evidence on the safety of the lower doses has been obtained.

Sometimes, for the first assessment of the dose-response curve of a new compound, the *dose-titration design* is used, in which increasing doses (if well tolerated) are administered to each patient, both in the active and in the control group, and the entire dose-response curves are compared between groups.

In the “*N of 1*” design, two or more treatments are repeatedly administered to a single patient: this approach is particularly useful in the study of symptomatic treatments of rare diseases or rare variants, for which the common approaches cannot be applied, simply because it is impossible to find the necessary number of patients. The restrictions are the same of those of any crossover design.

In the *simultaneous treatment design*, different treatments are simultaneously administered to the same patient. Such designs are generally used in ophthalmology and dermatology. All of the study treatments must have only a local effect (in terms of both efficacy and safety). These designs are analyzed as randomized block designs.

The *factorial designs* can be useful for studying two or more treatments simultaneously, when there is interest in the individual effects as well as in the combined ones.

Some therapeutic areas, such as oncology, have ethical problems of such magnitude that the trial designs must address these concerns first and foremost. In these situations, the *multistage designs* without control group are frequently used in Phase II of the clinical development.

Generally, the use of more sophisticated designs produces the undesired effect of increasing the complexity of the study, both at a practical and operational level and at a conceptual and methodological level. For example, the use of within-patient comparisons requires that each patient accepts a burden of visits and procedures which is often quite heavy. From a methodological point of view, these comparisons require that the researchers accept a considerable increase in the number of assumptions, which may be more or less verifiable. To justify the use of these strategies, these inconveniences must be balanced by relevant gains in terms of precision/efficiency and accuracy of the estimates.

Innovative Approaches to Drug Development

As mentioned at the beginning of this chapter, the classical clinical development of new pharmacological compounds is traditionally based on a sequence of studies, where the researcher has to wait for the end of the previous trial to design the next. In Phase II, generally, several studies are conducted in order to assess the clinical pharmacology and to evaluate different doses and/or dose schedules of an experimental drug. The Phase II results are then used for deciding whether subsequent Phase III trials should be conducted and at what dose(s) and schedule(s). In the conventional approach, Phase III typically consists of at least two independent randomized clinical studies with a predefined, fixed design. The decision by regulatory authorities on whether or not to grant marketing authorization is mostly made on the data collected in these final studies, assessed individually.

In most therapeutic areas, the costs and times associated with clinical development are prohibitive, and this, in turn, has ultimately a tremendous impact on the time that the patients should wait before new therapies can reach them.

It is obvious that a lot of effort is devoted to shortening the time and reducing the cost of the clinical development programs.

To this end, a variety of “flexible” approaches has been proposed to render clinical development more efficiently. The pioneer of such attempts has been the sequential approach, followed in more recent years by other promising approaches, including the adaptive methodology, the targeted or enrichment approach, and the pragmatic large randomized clinical study.

Table 4.5 provides a summary of the most common types of these innovative approaches.

In the *sequential* and the *adaptive approaches*, learning from accumulating data is used to plan or modify aspects of the same study, while it is ongoing, in order to optimize some of its characteristics, for example, minimize the sample size, optimize the probability of being exposed to the most promising treatment, reduce the overall study duration, etc. The adaptive designs have been particularly popular in

Table 4.5 Innovative approaches for clinical studies

Type of design	Description	Applicability, advantages, and limitations
1. Sequential designs	Based on a sequence of interim analyses, i.e., analyses that are performed while the study is ongoing. The interim analyses can be done for decision-making or for administrative reasons. In the former case, the aim is of deciding whether to interrupt the study for one of three reasons: (1) safety concerns, (2) early evidence of efficacy, and (3) early lack of efficacy (also referred as “futility”). In the latter case, the interim analyses are done for purposes that are external to the study itself, for example, for the planning of other studies	Can be applied when the treatment duration for each patient is short as compared to enrollment time Might worsen the precision of treatment effect estimates by increasing variability

Table 4.5 (continued)

Type of design	Description	Applicability, advantages, and limitations
1.1. Pure sequential designs	A new analysis is performed every time a new patient or a number of patients equal to the number of study treatments (one for each treatment) reaches the primary end-point [10]. The objective of each new sequential analysis is to decide whether to continue the study or not, based on predefined criteria. In the oncology and cardiovascular fields, these types of design are generally used with mortality as the outcome	
1.2. Group sequential designs	A prespecified number of interim analyses is performed on groups of patients enrolled sequentially [11, 12, 13]. The objective is the same as for the pure sequential designs. These designs are frequently applied in oncology	
2. Adaptive designs	Allow changes in key trial characteristics during the conduct of a study in response to information accumulating during the study itself, without introducing bias in treatment comparison and, if a frequentist approach is used, maintaining the overall level of type I and II errors under control. All information available at the time of performing one or more preplanned interim analyses is used for planning the subsequent steps of the study(ies) [14–16]. This approach may be particularly useful in Phases I and II of drug development	Applicable in those situations where the following criteria are met: Enrollment is relatively slow The efficacy end-point can be evaluated rapidly The data can be collected and analyzed quickly Some neurological (e.g., migraine), respiratory (e.g., asthma), and oncological (with its increasing availability of biomarkers) indications fulfill these criteria The use of an interactive voice response system (IVRS) for randomization is a prerequisite in any adaptive design Might be complicated on a logistical ground, for example, for drug supply
2.1. Flexible sample size reestimation	Allows sample size reestimation based on the results of one or more interim analyses. May or may not require unblinding of randomization code [17–20]	When unblinding is required, there may be an increase in overall sample size at study level

(continued)

Table 4.5 (continued)

Type of design	Description	Applicability, advantages, and limitations
2.2. Response-adaptive randomization	Allows modification of the randomization schedule with the aim of assigning more patients to the most promising study treatment(s), i.e., at each new patient entry, the probabilities of treatment allocation are recomputed based on study results obtained till that time [21, 22]. Often these designs use a Bayesian approach [23]	The approach is interesting because of the efficiency (expected sample size and trial duration) gain and the ethical advantage of assigning fewer patients to treatment arms with inferior outcomes Despite the use of stringent eligibility criteria, there may be a drift in patient characteristics over time
2.3 Adaptive dose-finding and dose-ranging	There are numerous uses of this type of design. In Phase I, the continual reassessment method is used in model-guided, adaptive dose-escalation designs [24]. It allows continual reassessment of the dose-response relationship based on the cumulative data collected on an ongoing basis [25]. A comparison of several types of adaptive dose-ranging studies with emphasis on Phase II has been carried out by a PhARMA Working Group [26]. The common feature of these approaches is that decisions on how to allocate future patients to one of the different dose levels are based on data observed up to the decision time; the decisions may include dropping dose levels that are “losers” or including new ones. Often these designs use a Bayesian approach [27, 28]	The choice of the starting dose level and dose range is a common problem, particularly in Phase I, but this is not unique to adaptive dose-ranging studies Complexity is a drawback of this approach
2.4. Phase I/II and Phase II/III seamless trials	Combines in a single study objectives that are traditionally addressed in separate trials. The methods for combining Phases II and III are based on adaptive two-stage designs, where stage 1 plays the role of the Phase II study and stage 2 plays the role of the Phase III study. In the first stage, the patients are randomized to one of several experimental treatments (generally different doses of the same treatment) and a control, and, at a predefined point, an interim analysis is performed to decide whether to continue the development of the experimental treatment and at what dose(s). The second stage is conducted in accordance with a protocol adapted at the time of the interim analysis, in terms of doses to be compared, sample size, and other study characteristics. At the end of the study, data from both stages are combined for the final analysis [29–33]	Substantial resources can be saved, and overall drug development time can be shortened; however, these gains should be assessed against the disadvantage that the logistical aspects of the study may become very complex

Table 4.5 (continued)

Type of design	Description	Applicability, advantages, and limitations
2.5. Other adaptive approaches	These include, but are not limited to, the adaptive treatment switching [34] that allows a patient to be switched from one treatment to another if there is evidence of lack of efficacy or safety issues emerge; the adaptive hypothesis design [35] that allows changing the hypothesis being tested after one or more interim analyses; the multiple adaptive design [36] that allows different changes of the study design	
3. Targeted or enrichment designs	All patients are screened for molecular alterations, and only the subpopulation, who either expresses or does not express a specific mutation or molecular alteration, is enrolled in the clinical trial [37]	Early evidence of benefit is required. Mostly used in oncology. Validation of the biomarkers is currently affected by several challenges, such as multitude of assessment methods, reliability in terms of sensitivity and specificity, and reproducibility of test/assay
3.1. Basket trials	Allow testing of one single treatment on patients with multiple diseases sharing the same drug target. Such studies have emerged in oncology with the aim of testing the hypothesis that a therapy aimed at a specific molecular target may be effective independently of tumor histology, as long as the molecular target is present. Later, basket trials have been extended to other therapeutic areas. The target can be a single mutation in a variety of cancer types or a molecular alteration responsible for different diseases. Each basket is a subgroup that may correspond to a specific disease, a specific combination of diseases and targets, or even more complex combinations. The efficiency of this strategy can be improved by assessing the heterogeneity of the basket's response at an interim analysis and aggregating the baskets that are proved to be homogeneous in the second stage [38–40]	Complexity may be a drawback In addition, the ideal design options may not be aligned to the different questions being asked, and this is a general problem when attempting to answer multiple questions in a single study
3.2. Umbrella or platform trials	Allow testing different treatments on a single disease, by building an experimental platform that continues to exist after the evaluation of a particular treatment or set of treatments. In a platform trial, all patients, even those assigned to treatments no longer under investigation, help in understanding and adjusting for the effects of confounding factors [41, 42]	Substantial resources can be saved by the use of the same trial infrastructure to evaluate multiple therapies, but a very good cooperation and coordination among different stakeholders (industry, academia, public sponsors) is essential

(continued)

Table 4.5 (continued)

Type of design	Description	Applicability, advantages, and limitations
4. Pragmatic approach	Allows verifying whether an intervention is effective in real-world conditions. The intervention may be a treatment but is often a service delivery or a policy implementation	The results are generalizable because the study setting is close to real-life conditions
4.1. Large simple trials	The pragmatic clinical studies are generally very large ($\geq 10,000$ patients), apply the parallel group design, adopt the cluster randomization, are based on simple protocols that include few nonrestrictive eligibility criteria, require to collect only the data that are immediately relevant to the primary end-point, and do not require a high degree of data monitoring [9]	Large simple trials are the gold standard for informing decision-makers on the benefit-risks of new interventions at population level. These studies tend to generate great variability and, therefore, require huge sample sizes
4.2. Stepped wedge cluster randomized trials	Offer a robust method of evaluation of an intervention delivered at the level of the cluster. Several clusters are included in the study. After an initial period in which no cluster is exposed to the intervention, this is rolled out at regular intervals (steps), with one or more clusters switching from control to intervention based on a randomized scheme, until all clusters have crossed over to the intervention. When designing such a study, the total number of clusters, the number of clusters to be randomized at each step, and the number and length of steps should be determined using statistical considerations. The design requires the fitting of complex models of statistical analysis, including adjustment for the time effect [43, 44]	This design is an alternative to the parallel cluster study: it is more efficient when the intra-cluster correlation is expected to be high and the clusters have a large size. It is preferable to the parallel cluster randomized study when there is already some evidence in support of the intervention and, therefore, there is resistance to the use of the parallel design, where only half of the clusters receive the intervention. More clusters are exposed to the intervention in the final stage of the study, and this implies that, in situations of an underlying temporal trend, the intervention effect may be confounded with the time effect

early drug development, in rare diseases, and in oncology, but the need for introducing more flexibility in clinical development is present in all therapeutic areas.

Actually, the rate of negative Phase III clinical studies has been very high [45] in many therapeutic areas, including common diseases, such as Alzheimer's disease [46], stroke [47], and various types of cancers.

The high rate of study failure in Phase III is to a large extent determined by a poor choice of the dose in Phase II [48]. Simulation and model-based dose estimation approaches and dose-exposure-response characterization may be very useful to improve the quality of dose selection: these methods are preferable to the traditional statistical pairwise comparisons to select the dose(s) for Phase III.

The adaptive designs described above may also be very useful to guide dose selection, because, for a given sample size, these designs allow to explore a bigger number of doses than the fixed designs, and to collect more data providing meaningful information on the dose-response curve (i.e., those in the steep part of the curve).

Another reason for the high rate of Phase III study failure is that, in many diseases, a unique treatment that is effective in most patients does not exist. In many areas, a successful therapeutic approach requires interventions that affect multiple targets with a combination of drugs or treatments that target specific subgroups of patients defined by genetic, proteomic, or other types of biomarkers. This fragmentation implies that frequent diseases are composed of a multitude of rare sub-diseases, each the target of a different treatment.

This need for therapeutic “precision” is the basis for the so-called targeted or enrichment designs. Two basic kinds of design have been developed in this area, the platform or umbrella trial and the basket trial, both aimed at facilitating patient enrollment. In *platform trials*, the patients with one disease (e.g., a cancer originating in one organ) are assessed for the presence of a series of biomarkers and are allocated to different treatment arms based on the results of this assessment. In *umbrella trials*, patients affected by the same molecular alteration, even if this is manifested in different diseases, are allocated to the same treatment arm(s).

At the opposite end of the study design spectrum, there is the so-called pragmatic clinical study paradigm: whereas the traditional randomized clinical trials are designed to maximize internal validity and often require expensive infrastructure to allow compliance with complex protocols, the pragmatic clinical studies are designed to maximize generalizability and external validity. Often the simple parallel cluster randomized designs are applied, which are very simple [9] and, therefore, often referred to as *large simple studies*. The stepped wedge cluster randomized study is also a pragmatic study design, which tries to reconcile the constraints of real life with the need for a rigorous evaluation of the interventions delivered at the level of the cluster.

Different study designs are appropriate in different situations, and some of the abovementioned approaches can be combined, for example, an adaptive strategy might be used in platform or umbrella trials.

New technologies (e.g., for data capture and study management), approaches (e.g., simulation), and regulatory options are evolving, all with the goal of reducing the overall time and costs of clinical development. The design principles and constructs described here drive the requirements for clinical research information systems (described in Chap. 8) and have implications for all aspects of clinical research planning, conduct, and analysis.

References

1. Bacchieri A, Della Cioppa G. Fundamentals of clinical research. Bridging medicine, statistics and operations. Milan: Springer; 2007.
2. Hill RG, Rang HP, editors. Drug discovery and development. 2nd ed. Churchill Livingstone: Elsevier; 2012.

3. DiMasi J, Hansen R, Gabrowski H. The price of innovation: new estimates of drug development cost. *J Health Econ.* 2003;22:151–8.
4. Lilienfeld AM, Lilienfeld DE. Foundations of epidemiology. 2nd ed. New York: Oxford University Press; 1980.
5. https://ec.europa.eu/health/sites/health/files/files/eudralex/vol-1/reg_2014_536/reg_2014_536_en.pdf. L158/12. Clinical Trials Regulation (EU) No 536/2014, p. 12.
6. Pretince R. Surrogate end-points in clinical trials: definition and operational criteria. *Stat Med.* 1989;8:431–40.
7. Bland JM, Altman DG. Regression toward the mean. *BMJ.* 1994;308:1499.
8. Bland JM, Altman DG. Some examples of regression toward the mean. *BMJ.* 1994;309:780.
9. <http://www.rethinkingclinicaltrials.org/>. Living textbook of pragmatic clinical trials.
10. Armitage P. Sequential medical trials. Blackwell Scientific Publications. Oxford: London; 1975.
11. Pocock SJ. Group sequential methods in the design and analysis of clinical trials. *Biometrika.* 1977;64(2):191–9.
12. O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. *Biometrics.* 1979;35(3):549–56.
13. Demets DL, Lan KG. Interim analysis: the alpha spending approach. *Stat Med.* 1994;13(13–14):1341–52.
14. Chow SC, Chang M. Adaptive design methods in clinical trials – a review. *Orphanet J Rare Dis.* 2008;3:11. <https://doi.org/10.1186/1750-1172-3-11>.
15. Meurer WJ, Lewis RJ, Berry DA. Adaptive clinical trials: a partial remedy for the therapeutic misconception? *JAMA.* 2012;307(22):2377–8.
16. Bauer P, Kohne K. Evaluation of experiments with adaptive interim analyses. *Biometrics.* 1994;50:1029–41.
17. Jennison C, Tumbull BW. Mid-course sample size modification in clinical trials based on the observed treatment effect. *Stat Med.* 2003;22:971–93.
18. Proschan M, Liu Q, Hunsberger S. Practical mid-course sample size modification in clinical trials. *Control Clin Trials.* 2003;24:4–15.
19. Shun Z. Sample size re-estimation in clinical trials. *Drug Inf J.* 2001;35:1409–22.
20. Gould AL. Sample size re-estimation: recent developments and practical considerations. *Stat Med.* 2001;20:2625–43.
21. Lin J, Lin LA, Sankoh S. A general overview of adaptive randomization design for clinical trials. *J Biom Biostat.* 2016;7:2. <https://doi.org/10.4172/2155-6180.1000294>.
22. Hu F, Rosenberger WF. The theory of response-adaptive randomization in clinical trials. Hoboken: Wiley. 2006
23. Thall PF, Wathen JK. Practical Bayesian adaptive randomization in clinical trials. *Eur J Cancer.* 2007;43:859–66.
24. Iasonos A, O'Quigley J. Adaptive dose-finding studies: a review of model-guided phase I clinical trials. *J Clin Oncol.* 2014;32(23):2505–11.
25. O'Quigley J, Pepe M, Fisher L. Continual reassessment method: a practical design for phase I clinical trials in cancer. *Biometrics.* 1990;46(1):33–48.
26. White paper of the PhARMA Working Group on adaptive dose-ranging studies. <https://www.pharma.org/search?incmode=keywordsearch&keyword=%202026.%20White%20paper%20of%20the%20PhARMA%20Working%20Group%20on%20adaptive%>.
27. Gaydos B, Krams M, Perevozskaya I, et al. Adaptive dose-response studies. *Drug Inf J.* 2006;40:451–61.
28. Bauer P, Rohmel J. An adaptive method for establishing a dose-response relationship. *Stat Med.* 1995;14:1595–607.
29. Maca J, Bhattacharya S, Dragalin V, et al. Adaptive seamless phase II/III designs. Background, operational aspects and examples. *Drug Inf J.* 2006;40:463–73.
30. Liu Q, Pledger GW. Phase 2 and 3 combination designs to accelerate drug development. *J Am Stat Assoc.* 2005;100:493–502.
31. Era 21Liu Q, Proschan MA, Pledger GW. A unified theory of two-stage adaptive designs. *J Am Stat Soc.* 2002;97:1034–41.

32. Era 22Bauer P, Kieser M. Combining different phases in the development of medical treatments within a single trial. *Stat Med.* 1999;18:1833–48.
33. Don GA. A varying-stage adaptive phase II/III clinical trial design. *Stat Med.* 2014;33:1272–87.
34. Branson M, Whitehead J. Estimating a treatment effect in survival studies in which patients switch treatment. *Stat Med.* 2002;21(17):2449–63.
35. Hommel G. Adaptive modifications of hypotheses after an interim analysis. *Biom J.* 2001;43:581–9.
36. Muller HH, Schafer H. A general statistical principle for changing a design any time during the course of a trial. *Stat Med.* 2004;23:2497–508.
37. Biankin AV, Piantadosi S, Hollingsworth SJ. Patient-centric trials for therapeutic development in precision oncology. *Nature.* 2015;526:361–70.
38. Chen C, Li X, Yuan S, Antonijevic Z, Kalamegham R, Beckman RA. Statistical design and considerations of a phase III basket trial for simultaneous investigation of multiple tumor types in one study. *Stat Biopharm Res.* 2016;8(3):248–57.
39. Cunanan KM, Gonen M, Shen R, Hyman DM, Riely GI, Begg CB, Iasonos A. Basket trials in oncology: a trade-off between complexity and efficiency. *J Clin Oncol.* 2017;35(3):271–3.
40. Cunanan KM, Iasonos A, Shen R, Hyman D, Begg CB, Gonen M. An efficient basket trial design. *Stat Med.* 2017;36(10):1568–79.
41. Berry SM, Connor JT, Lewis RJ. The platform trial: an efficient strategy for evaluating multiple treatments. *JAMA.* 2015;313(16):1619–20.
42. Saville BR, Berry SM. Efficiencies of platform clinical trials: a vision of the future. *Clin Trials.* 2016;13(3):358–66.
43. Hussey MA, Hughes JP. Design and analysis of stepped wedge cluster randomized trials. *Contemp Clin Trials.* 2007;28:182–91.
44. Hemming K, et al. The stepped wedge cluster randomized trial: rationale, design, analysis, and reporting. *BMJ.* 2015;h391:351.
45. Kola I, Landis J. Can the pharmaceutical industry reduce attrition rates? *Nat Rev Drug Discov.* 2004;3:711–6.
46. Cummings IL, Morstorf T, Zhong K. Alzheimer’s disease drug-development pipeline: few candidates, frequent failure. *Alzheimers Res Ther.* 2014;6(4):37.
47. Minnerup J, Wersching H, Schilling M, Schabitz WR. Analysis of early phase and subsequent phase III stroke studies of neuroprotectants outcomes and predictor for success. *Exp Transl Stroke Med.* 2014;6(1):2.
48. Sacks LV, et al. Scientific and regulatory reasons for delay and denial of FDA approval of initial applications for new drugs, 2000–2012. *JAMA.* 2014;311:378–84.



Public Policy Issues in Clinical Research Informatics

5

Jeffery R. L. Smith

Abstract

Recently, a national imperative to “develop better cures faster” has been a rally cry for clinical research, as stakeholders work to apply advances in data storage, computation, and methodology toward the clinical research enterprise. This work is, at its core, the domain of Clinical Research Informatics (CRI), and the intersection of public policy and CRI will be the focus of this chapter. The goal of this chapter is to provide a foundation for clinical research policies impacting the domain of CRI and to describe the emerging landscape of public policies likely to impact CRI for some time to come.

Keywords

Public policy · Common Rule · Regulatory science · Compliance · Privacy · Consent · Data sharing · HIPAA

Acronyms

AHRQ	Agency for Healthcare Research and Quality
All of Us	NIH All of Us Research Program
AMIA	American Medical Informatics Association
CDC	Centers for Disease Control and Prevention
CDRH	FDA Center for Devices and Radiological Health
CFR	Code of Federal Regulations
Common Rule	Federal Policy for the Protection of Human Subjects (45 CFR Part 46)

J. R. L. Smith, MPP (✉)

American Medical Informatics Association, Bethesda, MD, USA

e-mail: jsmith@amia.org

FD&C Act	Food, Drug, and Cosmetic Act of 1938
FDA	Food and Drug Administration
HeLa	Immortalized cell line of Henrietta Lacks
HIPAA	Health Insurance Portability and Accountability Act of 1996
IRB	Investigational Review Board
MDUFA	Medical Device User Fee Act
MIDD	Model-informed drug development
NIH	National Institutes of Health
NLM	National Library of Medicine
NPRM	Notice of proposed rulemaking
OIRA	Office of Information and Regulatory Affairs
OMB	White House Office of Management and Budget
ONC	Office of the National Coordinator for Health Information Technology
PCORI	Patient-Centered Outcomes Research Institute
PDUFA	Prescription Drug User Fee Act
PHI	Protected health information
PHSA	Public Health Services Act of 1944
PMI	Precision Medicine Initiative
PPACA	Patient Protection and Affordable Care Act of 2010
PreCert	FDA software precertification pilot program
RWD	Real-world data
RWE	Real-world evidence
SaMD	Software-as-a-Medical Device
SIMD	Software-inside-a-Medical Device

Introduction and the Role of Public Policy in CRI

Public policy can be generally defined as a system of laws, regulatory measures, courses of action, and funding priorities concerning a given topic promulgated by a governmental entity or its representatives.¹ Public policy reflects the norms and values of a society and is influenced by ideology, business concerns, and special interests. At its core, public policy creates distinct sets of winners and losers. Most of the public policy we will focus attempts to maximize public benefit while mitigating the effects for those negatively impacted by rules and regulations. If CRI is meant “to optimize the design and conduct of clinical research and the analysis, interpretation, and dissemination of the information generated,” from clinical research,² then health

¹ Kilpatrick, D. Definitions of Public Policy and the Law. Available at: <https://mainweb-v.musc.edu/vawprevention/policy/definition.shtml>.

² Embi PJ, Payne PR. Clinical research informatics: challenges, opportunities and definition for an emerging domain. J Am Med Inform Assoc. 2009;16(3):316–27.

and clinical research-related public policy can act as a driver to encourage, limit, or influence the use of informatics in clinical research.

Foundations of Clinical Research Policy

Over the last three decades, the role of public policy related to clinical research has increased dramatically in size and scope. The introduction of computers and computational methods to clinical research has added new levels of opportunity, as well as new kinds of risk to individual privacy, autonomy, and safety. Given that the focus of clinical research is directed toward human subjects, or performed to improve the human condition, the ethical and moral dimensions of clinical research public policy are pronounced.

Clinical research in the United States is governed by myriad laws and regulations and developed by numerous bodies at the federal, state, and local (institutional) level. This section will provide a brief history and description of landmark legislation, discuss how federal agencies implement legislation to enact laws through regulation, and reveal how recent policy development treats CRI as both an innovative tool for clinical research and an innovation unto itself.

Foundational Federal Legislation

Two landmark statutes for clinical research include the Food, Drug, and Cosmetic Act of 1938 and the Public Health Service Act of 1944. These statutes provide the underpinnings for many of the familiar federal agencies and programs that are well-known today. As will be demonstrated, these statutes are consequences of specific points in time, dynamic to changing cultural, social, and technological norms. The legislation we preview at the end of this section will discuss policies as influential as the landmark statutes, but more relatable to the current environment in which CRI has emerged and is evolving.

Food, Drug, and Cosmetic Act of 1938

An oft remarked quote inside Washington, D.C., stems from the French writer Victor Hugo: “You can resist an invading army; you cannot resist an idea whose time has come.” This notion of “an idea whose time has come” nicely captures the need for consensus in policymaking. It also evokes a sense of strategy and calculation or, perhaps, a sense of luck.

Even if the need for a new piece of legislation is apparent to some, our system of government requires that the underlying or motivating view be held by many. More than that, our system of government requires a sense of urgency and concern for inaction, and it requires the navigation of constantly changing political landscapes. Generally, these are critical components for the kinds of public policy that impact large swaths of the American public.

One such policy is the Food, Drug, and Cosmetic (FD&C) Act, which was a series of laws passed by Congress in 1938 that gave rise to the modern Food and Drug Administration (FDA). The FDA nicely summarizes the history of their authorizing statute:

The first comprehensive federal consumer protection law was the 1906 Food and Drugs Act, which prohibited misbranded and adulterated food and drugs in interstate commerce. Arguably the pinnacle of Progressive Era legislation, the act nevertheless had shortcomings—gaps in commodities it covered plus many products it left untouched—and many hazardous consumer items remained on the market legally.

The political will to effect a change came in the early 1930s, spurred on by growing national outrage over some egregious examples of consumer products that poisoned, maimed, and killed many people.

The tipping point came in 1937, when an untested pharmaceutical killed scores of patients, including many children, as soon as it went on the market. The enactment of the 1938 Food, Drug, and Cosmetic Act tightened controls over drugs and food, included new consumer protection against unlawful cosmetics and medical devices, and enhanced the government's ability to enforce the law. This law, as amended, is still in force today.³

The “growing national outrage” was the death of over 100 patients due to a sulfanilamide medication where diethylene glycol was used to dissolve the drug and make a liquid. While the public outcry from the sulfanilamide elixir disaster is credited with providing an element of urgency to amend existing policy, it was a federal report documenting the incident and the government’s response that laid bare the need for reform. A 1937 *New York Times* article explains the report’s findings that “before the elixir was put on the market it was tested for flavor but not for its effect on human life” and that “the existing Food and Drugs Act does not require that new drugs be tested before they are placed on sale.”⁴ It continues to quote the report: “Since the Federal Food and Drugs Act contains no provision against dangerous drugs, seizures had to be based on a charge that the word ‘elixir’ implies an alcoholic solution, whereas this product was a diethylene glycol solution. Had the product been called a ‘solution’ rather than an ‘elixir,’ no charge of violating the law could have been brought.” The article also highlighted how investigators had to sift through 20,000 sales slips in one of the distribution centers to understand where the elixir was sold and shipped. Note: The importance of data collection and need for supportive data processing tools to identify threats, leverage public health response, and inform policy solutions was evident almost 100 years ago. Not a surprise, as IT has grown, modern informatics tools have impacted the field (see Chap. 20 – Pharmacovigilance).

The FD&C Act has been amended many times since its passage, and various sections have been expanded or built upon. The modern FDA is seen the world over as a source of trusted innovation because it regulates drugs for safety and effectiveness. As we will discuss later, this has created incentives for both the FDA and manufacturers of drug and device to advance technology-assisted information management to support manufacturing processes, safety, and evaluations of effectiveness.

³<https://www.fda.gov/AboutFDA/Transparency/Basics/ucm214416.htm>.

⁴“‘Death Drug’ Hunt Covered 15 States,” *New York Times*, Nov. 26, 1937 <https://nyti.ms/2EzSmoO>.

Public Health Services Act of 1944

Whereas the FD&C Act can be described as sweeping legislation, developed in response to national threats related to food, drugs, and cosmetics, the Public Health Services Act (PHSA) of 1944 is important because it serves as an encompassing catalogue of laws related to public health and welfare. Both of these laws have expanded overtime, but where the FD&C Act is a book of many chapters that gives the FDA its charge, the PHSA is a set of encyclopedias that give charge to dozens of federal agencies and offices.

Agencies such as the National Institutes of Health,⁵ Agency for Healthcare Research and Quality,⁶ Office of the National Coordinator for Health Information Technology,⁷ and Substance Abuse and Mental Health Services Administration⁸ are all created through the PHSA and subsequent amendments. Some of the most well-known amendments include the Health Insurance Portability and Accountability Act (HIPAA) of 1996, the Patient Protection and Affordable Care Act (PPACA) of 2010, and various sections of the 21st Century Cures Act of 2016 amended both the PHSA and the FD&C Act. Collectively, the PHSA agencies and policies greatly impact research priorities and processes through funding and regulation. Billions of dollars are distributed through the NIH, CDC, AHRQ, and other PHSA agencies, and important policies relating to patient privacy, participant autonomy, and sharing of clinical trials data are heavily influenced by informatics.

Core Regulations and Guidance for CRI

Legislation is used to create federal agencies and to charge them with specific functions. For example, the 1956 statute that authorized the creation of the National Library of Medicine (NLM) states, “In order to assist the advancement of medical and related sciences and to aid the dissemination and exchange of scientific and other information important to the progress of medicine and to the public health, there is established the National Library of Medicine.”⁹ One of the NLM’s main functions is to promote the use of computers and telecommunications by health professionals for the purpose of improving access to biomedical information for healthcare delivery and medical research,¹⁰ and NLM has established various educational, research, and service programs over the years to carry out this charge.

Another example of legislation leading to formation of a new agency comes from Subchapter 28 of the PHSA, which established within the Department of Health and Human Services an Office of the National Coordinator for Health Information Technology (ONC) in 2009. Briefly, ONC is charged “with the development of a

⁵Title 42, Chapter 6A, Subchapter III.

⁶Title 42, Chapter 6A, Subchapter VII.

⁷Title 42, Chapter 6A, Subchapter XXVIII.

⁸Title 42, Chapter 6A, Subchapter III-A.

⁹Title 42, Chapter 6A, Subchapter III, Part D, Subpart 1 § 286(a).

¹⁰Ibid. § 286(b)(7).

nationwide health information technology infrastructure that allows for the electronic use and exchange of information...,” and functions or characteristics of that infrastructure are described in statute.¹¹ In addition to charging the new Office with development of a nationwide health IT infrastructure, the statute also charged ONC with identifying priority uses cases and standards related to the incentive programs for the meaningful use of certified EHR technology through development of a voluntary certification program. While the statute described the purpose of the certification program, it didn’t specify which standards to use or which use cases (e.g., computerized provider order entry) to prioritize. These tasks were left to ONC and other HHS agencies to determine how best to carry out the legislation. The translation of statutory language into specific programs and activities, known as implementation, is usually done through regulation.

From the seeds of legislation blooms dozens, sometimes hundreds, of rules and regulations. These are developed by federal agencies and offices and catalogued through the Code of Federal Regulations (CFR). Regulations are proposed and finalized daily, and these updates are communicated through the *Federal Register* (available at: <https://www.federalregister.gov/>). The Administrative Procedures Act of 1946 outlines the process for developing regulations, which are managed by the White House Office of Management and Budget’s (OMB) Office of Information and Regulatory Affairs (OIRA).¹² Some of the most influential and important regulations for CRI are discussed below.

Common Rule

Just as the modern FDA was born from preventable tragedy and the public will to better protect human health, so too are the policies governing our modern clinical research enterprise. A 40-year clinical study meant to understand the natural progression of untreated syphilis in rural African-American men and the ubiquitous use of cancer cells taken from tissues without consent in the 1950s are two of the most important catalysts for change in American clinical research.

Known as the Tuskegee Syphilis Study, a total of 600 impoverished African-American males were enrolled in a study conducted by the US Public Health Service and Tuskegee University under the guise of receiving free healthcare from the United States. The study lasted 40 years, from 1932 to 1972, and involved 399 men who had previously contracted syphilis before the study and 201 men who did not have the disease. The men were told the study would last 6 months, and they were given free meals, medical care, and burial insurance for participation. Those that had syphilis were never told they had the disease, and they were never treated with penicillin even after the antibiotic had become the standard of care by 1947. After a whistle-blower ended the study in 1972, only 74 of the test subjects were alive. Of the original 399 men, 28 had died of syphilis, 100 were dead of related complications, 40 of their wives had been infected, and 19 of their children were born with congenital syphilis.

¹¹Title 42, Chapter 6A, Subchapter XXVII, Part A §300jj–11(b). Office of the National Coordinator for Health Information Technology.

¹²Public Law 79–404, 60 Stat. 237, enacted June 11, 1946.

Revelations of the experiment led Congress in 1974 to pass the National Research Act, which created a national commission, a federal advisory council, and codified the role of institutional review boards (IRBs) in grant applications for human subjects research. The National Advisory Council for the Protection of Subjects of Biomedical and Behavioral Research was established within the Department of Health, Welfare, and Education to advise, consult with, and make recommendations to the Secretary concerning all matters pertaining to the protection of human subjects of biomedical and behavioral research. Separately, the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research was established and tasked with considering four issues for analysis, including (1) the boundaries between biomedical and behavioral research and the accepted and routine practice of medicine, (2) the role of assessment of risk-benefit criteria in the determination of the appropriateness of research involving human subjects, (3) appropriate guidelines for the selection of human subjects for participation in such research, and (4) the nature and definition of informed consent in various research settings.

The Commission published its report, issued in 1979, expressing three “basic ethical principles” relevant to researching involving human subjects: the principles of respect of persons, beneficence, and justice. Respect for persons portends that individuals should be treated as autonomous agents and that persons with diminished autonomy are entitled to protection. The principle of beneficence was described in the Belmont Report as an obligation to (1) do no harm and (2) maximize possible benefits and minimize possible harms. In the context of research, the report notes the obligations of beneficence affect both individual investigators and society at large – calling on both to “recognize the longer term benefits and risks that may result from the improvement of knowledge and from the development of novel medical, psychotherapeutic, and social procedures.” Last, the report describes the principle of justice as a necessary consideration to determine the proper distribution of burdens and benefits. The Belmont Report also included three applications of these principles to be (1) informed consent, (2) assessment of risk and benefits, and (3) selection of subjects. Together, these principles and applications form the basis for the Federal Policy for the Protections for Human Subjects, also known as the Common Rule.

The Common Rule is part of a set of regulations with origins dating to 1974 and adopted in its modern form by HHS and 15 other federal departments and agencies in 1991. For all intents and purposes, the Common Rule is the clinical researcher’s bible, dictating the contours of how research involving human subjects must be carried out.

Key definitions included as part of the Common Rule include *research* and *human subject*. Research “means a systematic investigation, including research development, testing and evaluation, designed to develop or contribute to generalizable knowledge.”¹³ The operative phrase in the first definition of research is “generalizable knowledge.” The practical implication of this definition means that

¹³<https://www.hhs.gov/ohrp/regulations-and-policy/regulations/45-cfr-46/index.html#46.101>.

retrospective review of EHR data to determine which hip implant performs better over time is not subject to the Common Rule if it is part of a hospital's internal quality improvement. However, if the findings of this review are published in a peer-reviewed journal, making it "generalizable knowledge," it would be subject to the Common Rule.

Another important definition is *human subject*, which "means a living individual about whom an investigator (whether professional or student) conducting research obtains (1) data through intervention or interaction with the individual, or (2) identifiable private information."¹⁴ The definition of human subject is also important because the parameters of this definition can have profound impacts in a world where unidentified genomic and other "-omic" data can be reidentified through emerging analytic results.

Beginning in 2011, HHS signaled its intention to revise the Common Rule. Substantive updates had not occurred in more than 10 years, and HHS heard from several stakeholders, many of whom included the CRI community that advancements in computing power, digital storage, and other methodological improvements were changing the nature of clinical research. Another primary motivation for revising the Common Rule was an acknowledgment of widespread use of Henrietta Lacks' immortalized cell line without her, or her family's, consent.

The HeLa case chronicled in "The Immortal Life of Henrietta Lacks"¹⁵ created controversy as yet another example where the research enterprise failed to protect autonomy and justice for research participants. The revelations of widespread use of the HeLa cell line lead many public officials to wonder if more systems and controls would be needed to ensure that such a case would never reoccur. In fact, the HeLa case was central to proposed revisions to the Common Rule, introduced in 2015.

Common Rule Revisions

In 2015, HHS proposed revisions to the Common Rule through a notice of proposed rulemaking, or NPRM. The preamble of the revised Common Rule NPRM stated that the volume and landscape of research had changed dramatically, citing expansions in the number and type of clinical trials and observation studies, increased use of sophisticated analytic techniques, and growing use of EHR data.¹⁶ The NPRM's most impactful, and most controversial, revision was largely in acknowledgment of the HeLa case and changing cultural norms related to consent.

The NPRM made the claim that allowing secondary research with biospecimens collected without consent places publicly funded research in an increasingly untenable position in the eyes of the public. It went on to propose that the Common Rule should apply to the obtaining, use, study, or analysis of biospecimens, regardless of identifiability.¹⁷ The premise of this proposal was also based on the presumption that

¹⁴Ibid.

¹⁵Skloot, R. *The Immortal Life of Henrietta Lacks*. (2011). Broadway Books.

¹⁶80 Fed. Reg. 173. Pages 53933-54061. September 8, 2015.

¹⁷Ibid.

biospecimens could, at some point, become readily identifiable as a result of increasingly sophisticated technology.¹⁸

The implications of requiring consent for all biospecimens have tremendous implications for CRI. Implementing the systems and controls necessary to comply with such a revision would be extensive, including the need for new software and processes to (1) capture and store consent for each biospecimen collected, which will have implications for data entry/intake and database management, (2) track versioning of consent forms, (3) enable querying of consent status, and (4) enable patients to change or withdraw consent preferences. This kind of technical system and set of controls would also need to fit within existing informatics infrastructure and policies.

The clinical research community was split on the proposed approach to revise the definition of human subjects to capture consent on all biospecimens regardless of identifiability. The American Medical Informatics Association (AMIA) provided tepid support, urging officials to take various measures to mitigate burdens associated with the proposal. Others, including the Presidential Commission for the Study of Bioethical Issues and HHS's own Advisory Committee, flatly rejected the proposal, cautioning it would stall certain kinds of research using de-identified biospecimens that pose no risk to human subjects and are unlikely to impact participants' autonomy interests.¹⁹

Ultimately, HHS did not finalize the rule as proposed. In the preamble of the final rule, officials acknowledged that "the majority of commenters who addressed this expansion [to include all biospecimens] opposed it for a variety of reasons," raising "sufficient questions" about the premise of autonomy raised in the NPRM. Clearly, this idea's time has yet to come.²⁰

The final rule contained many provisions to improve the availability of data for secondary research while strengthening protections for research participants. The final rule also opted not to move forward with several provisions that would have incurred undue burden to those involved with conducting research. Specifically, the final rule:

- Made important changes to consent by requiring the most important information regarding a study to be explained clearly and concisely and in a way that a "reasonable person" could understand²¹
- Permitted researchers to seek broad consent, which is meant to improve the availability of biospecimen- and patient-reported data (including real-time data from mobile applications and devices) for secondary research²²

¹⁸Ibid.

¹⁹Council on Governmental Relations. Analysis of Public Comments on the Common Rule NPRM. May 2016. Available at: <http://www.cogr.edu/sites/default/files/Analysis%20of%20Common%20Rule%20Comments.pdf>.

²⁰82 Fed. Reg. 12, Pages 7149–7274. January 19, 2017.

²¹§_____.116(a), 0.116(b) & 0.116(c) discussion beginning 82 Fed. Reg. 12, page 7210.

²²§_____.116(d) discussion beginning 82 Fed. Reg. 12, page 7216.

- Enabled more secondary research of EHR data by exempting certain low-risk studies conducted by HIPAA covered entities²³
- Clarified that certain public health surveillance activities are explicitly outside the scope of the Common Rule, so that the spread of disease can be more easily monitored²⁴
- Eliminated the need for continuing review for many studies, reducing administrative burden²⁵
- Provided a new option meant to help screening of potential participants, so patients who qualify for new treatments are more likely to learn about them²⁶

The expected compliance date for the revised Common Rule is in 2019. While imperfect, the revised Common Rule exemplified the kind of transparent, deliberate, and constructive process sought by stakeholders, and it will have lasting impact as more stakeholders become familiar with its new provisions. IT and informatics will be an enabler to more efficient compliance, but so too will informatics require policy to evolve. As technology and methods for generating, collecting, analyzing, and applying data to clinical research advance, it is likely the Common Rule will need to undergo periodic review. Good public policy is extensible and has processes for review and amendment. This is even more important in the domain of technology policy, given how rapid best practices change.

Food and Drugs Regulation and Guidance

While the FDA has conforming provisions related to the Common Rule,²⁷ numerous FDA regulations have implications for CRI, mostly as an administrative tool to ensure drugs and devices are safe and effective. Title 21 CFR Part 11 set forth the criteria by which FDA considers “electronic records, electronic signatures, and handwritten signatures executed to electronic records to be trustworthy, reliable, and generally equivalent to paper records and handwritten signatures executed on paper.”²⁸ Unique to the FDA are the use of guidance documents. These guidance documents are meant to interpret legislation for industry and other stakeholders without spelling out required activities via regulations. Relevant guidance usually focuses on ways to facilitate approval of a drug or device and relies more on “shoulds” than “shall.”

Guidance documents can be complex to understand and navigate. However, they are vital to the work of clinical research and have increasing relevance to CRI. The FDA has developed a guidance resource page (available at: <https://www.fda.gov/RegulatoryInformation/Guidances/>) for industry to learn about their guidance process and search active guidance documents.

²³ § _____.104(d)(4) discussion beginning 82 Fed. Reg. 12, page 7191.

²⁴ § _____.102(l)(2) discussion beginning 82 Fed. Reg. 12, page 7175.

²⁵ § _____.109(f) discussion beginning 82 Fed. Reg. 12, page 7205.

²⁶ § _____.116(g) discussion beginning 82 Fed. Reg. 12, page 7227.

²⁷ see Title 21 CFR Part 50 and 56.

²⁸ Title 21 CFR §11.1(a).

HIPAA Privacy Rule and Research

The Health Insurance Portability and Accountability Act of 1996 (HIPAA) was developed to modernize how health information is managed and used for care, billing, and insurance. A specific mandate from HIPAA included the development of rules for the security and privacy of individually identifiable health information. These rules applied to covered entities, which are health plans, healthcare clearinghouses, and healthcare providers that transmit health information electronically in connection with certain defined HIPAA transactions, such as claims or eligibility inquiries. Generally, researchers are not required to comply with HIPAA unless they are employees of a covered entity, such as a hospital or insurer. However, researchers who are not themselves covered entities or work for covered entities, yet use data supplied by covered entities, may be indirectly affected by the privacy rule.

According to the HIPAA privacy rule, protected health information (PHI) can only be used or disclosed in certain circumstances or under certain conditions without specific patient consent. Most well-known of these circumstances are treatment, payment, or operations, or TPO.

Research is not among the reasons that PHI may be used or disclosed without consent. However, the HIPAA privacy rule establishes the conditions under which PHI may be used or disclosed for research purposes. Research is defined in the HIPAA privacy rule as “a systematic investigation, including research development, testing, and evaluation, designed to develop or contribute to generalizable knowledge.”²⁹

The privacy rule allows covered entities to use and disclose protected health information for research with individual authorization or without individual authorization under limited circumstances. The privacy rule permits a covered entity to use or disclose PHI for research under the following circumstances and conditions:

- If the subject of the PHI has granted specific written permission through an authorization
- For reviews preparatory to research with representations obtained from the researcher
- For research solely on decedents’ information with certain representations and, if requested, documentation obtained from the researcher
- If the covered entity receives appropriate documentation that an IRB or a Privacy Board has granted a waiver of the authorization requirement
- If the covered entity obtains documentation of an IRB or Privacy Board’s alteration of the authorization requirement as well as the altered authorization from the individual
- If the PHI has been de-identified in accordance with the standards set by the privacy rule at section 164.514(a)–(c) (in which case, the health information is no longer PHI)
- If the information is released in the form of a limited data set, with certain identifiers removed and with a data use agreement between the researcher and the covered entity

²⁹ See 45 CFR 164.501.

It is worth noting that HIPAA uses exemptions and waivers to accommodate various uses of PHI. Similar to the Common Rule, such an approach requires dedicated professionals to interpret the applicability of various uses cases to HIPAA, and it has created uncertainty among clinical practitioners and researchers about what may or may not be subject to HIPAA. The NIH has published a HIPAA Privacy Rule Booklet for Research online (available at: https://privacyruleandresearch.nih.gov/pr_02.asp) containing numerous helpful links and resources.

Now that the basic underpinnings of clinical research policy have been reviewed, an examination of more recent legislation and agency activities will reveal how federal policy drives and constrains the use of informatics in clinical research.

Regulatory Science and the Role of Informatics

The federal government has been a driving force for the use of informatics in clinical research by being both a consumer and regulator of informatics tools. Decisions over how to use, and how to require the use of, such tools for research will continue to play a major role in the evolution of CRI. This section will highlight several of these strategic efforts and important trends, especially those at the FDA and NIH.

Regulatory Science as a Driver of Informatics at the FDA

In 2007, a report from the FDA Science Board’s Subcommittee on Science and Technology found that “FDA’s inability to keep up with scientific advances means that American lives are at risk. While the world of drug discovery and development has undergone revolutionary change—shifting from cellular to molecular and gene-based approaches — FDA’s evaluation methods have remained largely unchanged over the last half-century.”³⁰ This finding lead to the development of several documents outlining strategies for how the FDA could better harness recent and emerging breakthroughs in research and information technology.

A 2010 report, “Advancing Regulatory Science for Public Health: A Framework for FDA’s Regulatory Science Initiative,” said the FDA “must play an increasingly integral role as an agency not just dedicated to ensuring safe and effective products, but also to promote public health and participate more actively in the scientific research enterprise directed towards new treatments and interventions.”³¹ The report noted the need to “modernize our evaluation and approval processes to ensure that innovative products reach the patients who need them, when they need them.”³² This

³⁰ FDA Science Board, FDA Science and Mission at Risk, Report of the Subcommittee on Science and Technology, November 2007. https://www.fda.gov/ohrms/dockets/ac/07/briefing/2007-4329b_02_01_FDA%20Report%20on%20Science%20and%20Technology.pdf.

³¹ Food and Drug Administration. “Advancing Regulatory Science for Public Health: A Framework for FDA’s Regulatory Science Initiative,” October 2010. Available at <https://www.fda.gov/downloads/ScienceResearch/SpecialTopics/RegulatoryScience/UCM228444.pdf>.

³² Ibid.

framework introduced the concept of regulatory science and provided rationale for the need to invest in such work on behalf of American public health.

FDA defined its view of regulatory science as, “the science of developing new tools, standards and approaches to assess the safety, efficacy, quality and performance of FDA-regulated products.”³³ Section IV of the FDA’s framework focused on “Enhancing Safety and Health Through Informatics.”

FDA houses the largest known repository of clinical data — unique, high-quality data on the safety, efficacy and performance of drugs, biologics and devices, both before and after approval...But we lack the right infrastructure, tools and resources to organize and analyze these large data sets across the multiple studies and data streams. In other words, we have a valuable library full of information, but no indices or tools for translation.³⁴

The report noted that an increased investment in regulatory science would allow the FDA to leverage existing historical data as well as the new data coming into FDA every day to provide “unprecedented insight into the mechanisms that govern [therapies’] successes or failures.” The FDA identified various areas for advancements, including real-time monitoring of safety data using healthcare data and data mining and scientific computing to:

- Develop and implement active post-market safety surveillance system that queries health system databases to identify and evaluate drug safety
 - Employ advanced informatics, modeling and data mining to better detect and analyze safety signals
 - Apply computer-simulated modeling to risk assessment and risk communication strategies that identify and evaluate threats to patient safety; develop methods for quantitative risk-benefit assessments
 - Enhance IT infrastructure to support the scientific computing required for meta-analyses and computer models for risk assessment
 - Apply clinical trial simulation modeling and adaptive and Bayesian clinical trial design methods to facilitate development of novel products
 - Apply human genomic science to the analysis, development, and evaluation of novel diagnostics, therapeutics, and vaccines
1. The 2010 framework was expanded into a “Strategic Plan for Advancing Regulatory Science at the FDA” in 2001.³⁵ This plan identified eight priority areas of regulatory science where new or enhanced engagement is essential to the continued success of FDA’s public health and regulatory mission. Section 5 of the plan articulated the FDA’s intentions to develop agency informatics capabilities by enhancing IT infrastructure development and data mining, applying

³³ Ibid.

³⁴ Ibid.

³⁵ Food and Drug Administration. “Advancing Regulatory Science at FDA: A Strategic Plan,” August 2011. Available at: <https://www.fda.gov/downloads/ScienceResearch/SpecialTopics/RegulatoryScience/UCM268225.pdf>.

simulation models for product life cycles and risk assessments, and analyzing large-scale clinical and preclinical data sets.

- From 2010 to 2018, FDA has endeavored to refine and enhance their use of informatics to better regulate drugs and devices. For example, in late 2016, FDA released its “Regulatory Science Priorities for Fiscal Year 2017,” identifying the top ten regulatory science needs for the Food and Drug Administration’s Center for Devices and Radiological Health (CDRH) in fiscal year 2017.³⁶ These priorities serve as a guide for making funding decisions to ensure that the CDRH’s research is focused on issues that are relevant and critical to the regulatory science of medical devices.

This document argued that increased funding was necessary to develop the infrastructure; statistical or analytical tools and models; information retrieval and processing for Big Data, relevant to enhancing safety; and performance and quality of medical devices. It also previewed an emerging buzz word: Digital Health. Noting that medical devices are increasingly connected to other devices, internal networks, the Internet, and portable media, the report called for more research on ways to regulate the safety, effectiveness, and cybersecurity of medical device and software.³⁷

In January 2018, FDA issued its 2018 Strategic Policy Roadmap pledging to continue its work in regulatory science, opting to call it FDA’s Regulatory Toolbox.³⁸ FDA intends to embrace advances like predictive toxicology methods and computational modeling across its different product centers, and the FDA pledged to make new investments in the FDA’s high-performance, scientific computing. Pointedly, the 2018 Policy Roadmap says that the Agency’s own policies and approaches must “keep pace with the sophistication of the products that we are being asked to regulate, and the opportunities enabled by improvements in science.”³⁹

Real-World Evidence

The FDA is funded through two primary mechanisms: traditional appropriations and user fees, paid for by regulated industry. What began as user fees from medical devices and pharmaceutical manufacturers has been expanded to include biosimilar biologic products and generic drugs. User fees amount for roughly \$1 billion per year from industry. Periodically, the FDA renegotiates the terms of the user fees with industry to produce a “commitment letter.” These commitment letters then inform legislative language to reauthorize the FDA to collect user fees.

³⁶Food and Drug Administration. “Regulatory Science Priorities for Fiscal Year 2017,” September 2016. Available at: <https://www.fda.gov/downloads/MedicalDevices/ScienceandResearch/UCM521503.pdf>.

³⁷Ibid.

³⁸Food and Drug Administration. “Health Innovations, Safer Families: FDA’s 2018 Strategic Policy Roadmap,” January 2018. Available at: <https://www.fda.gov/downloads/AboutFDA/ReportsManualsForms/Reports/UCM592001.pdf>.

³⁹Ibid.

Commitment letters outline performance goals, articulating how FDA will implement its strategic priorities for how it intends to spend user fees for review of new drugs and manufacturing processes. The most recent use fee agreements, known as UFAs, were reauthorized in 2017 by Congress, which enabled the FDA to proceed with several important initiatives that operationalize many of the regulatory science aspirations espoused in the earlier documents.

For example, both the prescription drug (PDUFA) and medical device (MDUFA) user fees articulated plans to leverage real-world evidence (RWE) for premarket reviews and decision-making. The PDUFA commitment letter said FDA will leverage user fees to launch pilot studies and develop draft guidance on how RWE can contribute to safety and effectiveness efforts,⁴⁰ and minutes from a May 16, 2016, MDUFA negotiation meeting with industry sought \$30 million to “contribute to the implementation of a system that improves the quality of RWE and linkages among data sources” in the premarket setting.⁴¹

The PDUFA letter also discussed ways FDA would facilitate the development of exposure-based, biological, and statistical models derived from preclinical and clinical data sources, referred to as “model-informed drug development,” or MIDD.⁴² User fees will also provide support for post-marketing drug safety evaluation, through expansion of the Sentinel System and integration into the FDA pharmacovigilance activities and timely communication of post-marketing safety findings related to human drugs.⁴³

While industry has greeted these developments with optimism, the ability to leverage RWE and operationalize MIDD is still in early stages. In 2016, FDA issued draft guidance on the use of EHR data in clinical investigations.⁴⁴ Comments submitted to the FDA by AMIA said the draft guidance could serve as a valuable signal to industry, and other stakeholders, on how to orient technical functionalities and organizational policies to leverage EHR data for FDA-regulated research. However, AMIA also questioned the state of readiness for most EHRs to provide research quality data, especially for prospective randomized controlled trials. “With more than 96 percent of U.S. hospitals and 83 percent of U.S. office-based physicians using EHRs to deliver clinical care, we have an unprecedented opportunity to utilize digitized healthcare data for supplemental uses, such as clinical investigations,” AMIA said in comments. “However, we strongly caution the FDA from assuming

⁴⁰Food and Drug Administration. “PDUFA Reauthorization Performance Goals and Procedures Fiscal Years 2018 Through 2022,” June 2017. Available at: <https://www.fda.gov/downloads/ForIndustry/UserFees/PrescriptionDrugUserFee/UCM511438.pdf>.

⁴¹Food and Drug Administration. “Industry MDUFA IV Reauthorization Meeting.” May 16, 2016. Available at: <https://www.fda.gov/downloads/ForIndustry/UserFees/MedicalDeviceUserFee/UCM507305.pdf>.

⁴²Food and Drug Administration. “PDUFA Reauthorization Performance Goals and Procedures Fiscal Years 2018 Through 2022,” June 2017 (page 30).

⁴³Ibid. (page 35).

⁴⁴Food and Drug Administration. “Use of Electronic Health Record Data in Clinical Investigations Draft Guidance for Industry.” May 2016. Available at: <https://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM501068.pdf>.

EHRs are readily configurable for clinical investigations, even among more advanced institutions.”⁴⁵

Given the strategic importance of RWE across multiple FDA centers, it is likely that much more work and funding will be devoted to the concepts articulated by the user fees over the next 5 years and beyond.

Digital Health

Two important developments in medical devices have occurred in the last few years: (1) software used inside medical devices has become more pervasive as the Internet of Things has entered the medical space, and (2) software is being developed as the medical device. Known as Software-inside-a-Medical Device (SiMD) and Software-as-a-Medical Device (SaMD), these devices are blurring the lines between informatics as a tool to regulate and informatics as a tool to be regulated.

The May 2016 MDUFA commitment letter articulated the need for funding to ensure “consistent review of software, streamlining and aligning FDA review processes with software life cycles, continued engagement in international harmonization efforts related to software review, and other activities related to Digital Health.” In June 2017, FDA officials detailed, for the first time, how FDA hopes to implement new policy concepts for emerging technology that rely heavily on software and data.

Dubbed the “FDA Digital Health Innovation Plan,” Commissioner Gottlieb detailed in a 2017 blog how the agency hopes to foster “innovation at the intersection of medicine and digital health technology” while promoting “the development of safe and effective medical technologies that can help consumers improve their health.”⁴⁶ New regulatory guidance, firm-based premarket review, and improved post-market surveillance using real-world data are the hallmarks of this new strategy for emerging medical devices. This plan articulated the need to develop and disseminate new regulatory guidance to help innovators better understand when their products will be regulated by FDA and when they will not. Specifically, FDA said it intends to issue guidance on (1) products that contain multiple software functions, where some fall outside the scope of FDA regulation, but others do not and (2) technologies that present “low enough risks” that FDA does not intend to subject them to certain premarket regulatory requirements.

The plan also described a pilot program to test the use of a third-party certification process where lower-risk digital health products could be marketed without FDA premarket review and higher risk products could be marketed with a streamlined FDA premarket review.⁴⁷ FDA refers to this as a firm-based, rather than a

⁴⁵ American Medical Informatics Association. Letter to FDA Commissioner Dr. Robert Califf RE: “Use of Electronic Health Record Data in Clinical Investigations; Draft Guidance for Industry.” Available at: <https://www.amia.org/sites/default/files/AMIA-Response-to-FDA-Draft-Guidance-on-Using-EHR-Data-in-Clinical%20Investigations.pdf>.

⁴⁶ Food and Drug Administration. FDA Voice Blog, “Fostering Medical Innovation: A Plan for Digital Health Devices.” June 15, 2017. Available at: <https://blogs.fda.gov/fdavoice/index.php/2017/06/fostering-medical-innovation-a-plan-for-digital-health-devices/>.

⁴⁷ Ibid.

product-based, approach to premarket review, for medical devices manufacturers that are deemed to have consistent and reliable high-quality software design and testing (validation) and ongoing maintenance of its software products.

FDA officials added definition to this plan a month later in July 2017, in the form of a Digital Health Innovation and Action Plan.⁴⁸ As part of this plan, FDA officials announced the “Software Precertification (PreCert) Pilot Program,” to develop a tailored approach toward regulating software-based medical technologies.⁴⁹ The PreCert Pilot is in response to new medical device manufacturing processes that enable continuous modification and updates through Internet connectivity, which are requiring the FDA to rethink its approach to regulation. According to the Action Plan, “FDA intends to develop a precertification program that could replace the need for a premarket submission for certain products and allow for decreased submission content and/or faster review of the marketing submission for other products. The first step is a pilot program to develop a new approach toward regulating this technology – by looking first at the software developer or digital health technology developer, not the product.”

The PreCert Pilot is underway with Apple, Fitbit, Johnson & Johnson, Pear Therapeutics, Roche, Samsung, and Google’s Verily among the participating companies.⁵⁰ The results of the pilot are expected in 2019.

NIH as a Driver of Informatics Through Public Policy

Where the FDA is developing policy that will require internal informatics capacity to better assess emerging drugs and devices for safety and efficacy, the NIH is poised to drive demand for informatics capacity through public policy in its attempts to tackle long-standing issues related to research data sharing and reproducibility.

21st Century Cures Act

The 21st Century Cures Act of 2016 reflected a watershed moment for clinical research. Not only was the legislation passed in overwhelmingly bipartisan fashion, it infused the NIH budget with billions of dollars in funding for the development of a million-person cohort observational study and continued efforts to cure cancer, better understand the human brain, and develop new approaches to regenerative medicine. In a nutshell, the Cures Act is the most important piece of legislation for clinical research in a generation.

⁴⁸ Food and Drug Administration. “Digital Health Innovation and Action Plan,” July 2017. Available at: <https://www.fda.gov/downloads/MedicalDevices/DigitalHealth/UCM568735.pdf>.

⁴⁹ Food and Drug Administration. “Digital Health Software Precertification (PreCert) Program,” July 2017. Available at: <https://www.fda.gov/MedicalDevices/DigitalHealth/DigitalHealthPreCertProgram/Default.htm>.

⁵⁰ Food and Drug Administration. “Software Precertification Pilot Program Participants,” Sept. 2017. Available at: <https://www.fda.gov/MedicalDevices/DigitalHealth/DigitalHealthPreCertProgram/ucm577330.htm>.

The Cures Act codified and funded many important programs including the Precision Medicine Initiative, known as the All of Us Research program,⁵¹ and the Cancer Moonshot Initiative, known as the Beau Biden Cancer Moonshot.⁵² The sheer scope of the All of Us Research program will have a transformational impact on CRI. Numerous and complicated policy development has been initiated to implement this program and orchestrate the pan-NIH and intra-agency activities. For example, key aspects of the program will require that the million-person cohort donate their EHR data for research purposes and that research results be returned to participants. Policies to support these activities were crafted in 2015 as the PMI Privacy and Trust Principles⁵³ and the PMI Data Security Policy Principles and Framework.⁵⁴ In addition, Sync for Science⁵⁵ and Sync for Genes⁵⁶ are two pilots attempting to develop standards and protocols for this kind of data donation and sharing.

The Core Protocol version 1 of the All of Us Research program was published in August 2017, articulating how consent will be managed, data access policies implemented, and other aspects of the study carried out.⁵⁷ The program will rely on participant-provided information, EHRs, physical measurements, biospecimens, and passive mobile and digital health data to create a resource for research. The informatics components of this program are substantial. Formal business processes, process data collection, and quality assurance and improvement methods will be used to test and improve methods for patient recruitment, engagement, and retention. The program is committed to recruiting diverse and historically underrepresented populations in research and will undoubtedly include a variety of approaches to reach different geographical, racial, and sociodemographic populations. Further, the technology and communication approaches for sharing results with participants will impact our evidence base of how to conduct research efficiently and effectively. Further, the program will have to address privacy and security issues that are critical for examining a range of data – including genetic – on individuals in order to protect and preserve trust in research for patients, families, and communities. (See other chapters – recruitment, consumer, and future of CRI.)

⁵¹National Institutes of Health. All of Us Research Program. Available at: <https://allofus.nih.gov/>.

⁵²National Institutes of Health. Cancer Moonshot Initiative. Available at: <https://www.cancer.gov/research/key-initiatives/moonshot-cancer-initiative>.

⁵³White House. “Precision Medicine Initiative: Privacy and Trust Principles,” Nov. 9, 2015. Available at: <https://allofus.nih.gov/sites/default/files/privacy-trust-principles.pdf>.

⁵⁴White House. “Precision Medicine Initiative: Data Security Policy Principles and Framework,” May 25, 2016. Available at: <https://allofus.nih.gov/sites/default/files/security-principles-framework.pdf>.

⁵⁵<http://syncfor.science/>.

⁵⁶<http://www.sync4genes.org/>.

⁵⁷National Institutes of Health. All of Us Research Program Protocol Version 1. Aug. 2017. Available at: https://allofus.nih.gov/sites/default/files/allofus-initialprotocol-v1_0.pdf.

Data Sharing Policies

Over the last 2–3 years, the lack of data sharing in research has moved from an academic annoyance to the subject of congressional hearings and headline news across the country. Americans believe the time has come for sharing data and that publicly funded research should result in data that is a public resource.

In 2015, the unexpected death of Vice President Joseph Biden's son, Beau Biden, galvanized lawmakers in Washington, DC, and revealed a porous landscape of data sharing across the national research enterprise. The Biden family's experience with the healthcare system and research enterprise, mirroring the experience of thousands of American families, lead to a “cancer moonshot” – a multibillion dollar initiative to speed cancer research and make more treatments available to more patients, as well as to improve cancer detection and prevention.⁵⁸

The federal agencies and Congress quickly worked to develop a comprehensive strategy with funding.⁵⁹ Federal advisors from academia, industry, nonprofit, and public sectors formed a Blue Ribbon Panel and began work on developing policy ideas. Many of the challenges identified by the Panel were systemic; the way government funds research, the way academia rewards publication, the way journals encourage data deposition, and the way tenure is established all contribute to the current state. A specific Working Group was established as part of the Blue Ribbon Panel, which said the federal government had an opportunity and obligation to develop a national infrastructure for sharing cancer data.⁶⁰ Such an infrastructure would support the development of a Cancer Data Ecosystem meant to enable all participants across the cancer research and care continuum to contribute, access, combine, and analyze diverse data that will enable new discoveries and lead to lowering the burden of cancer across the country. The Working Group identified major barriers to this ecosystem to include a lack of searchable and interconnected data repositories with associated tools and services, lack of data standards and interoperability, and lack of harmonized consent and data use agreements. The group recommended development of software tools and services that would:

- Better enable clinical research activities such as the design of future prospective trials, clinical trial recruitment feasibility analysis, or provide a retrospective cohort as a comparator arm
- Facilitate patient-centeredness through dynamic consent, access to current information about specific conditions, clinical trials, research opportunities, and integration with the many cancer advocacy and disease-focused communities

⁵⁸ Scott, D., “Joe Biden calls for ‘moonshot’ to cure cancer,” Oct. 21, 2015. *STAT*. Available at: <https://www.statnews.com/2015/10/21/joe-biden-calls-for-moonshot-to-cure-cancer/>.

⁵⁹ National Cancer Institute. Cancer Moonshot Blue Ribbon Panel Report, October 2016, available at <https://www.cancer.gov/research/key-initiatives/moonshot-cancer-initiative/blue-ribbon-panel#ui-id-3>.

⁶⁰ National Cancer Institute. Cancer Moonshot Blue Ribbon Panel Report, Enhanced Data Sharing Working Group Recommendation: The Cancer Data Ecosystem <https://www.cancer.gov/research/key-initiatives/moonshot-cancer-initiative/blue-ribbon-panel/enhanced-data-sharing-working-group-report.pdf>.

- Improve clinical decision support tools to leverage knowledge base and data repositories will be integrated into clinical workflows and clinical information systems, enabling healthcare providers and patients to engage in shared decision-making for treatment prioritization for individual patients

Some of the report's recommendations have translated into policy work at PCORI and the NIH, and it has spurred activity in the private sector. For instance, PCORI developed a draft Data Sharing Policy late in 2016,⁶¹ meant to enable improved use of data generated by PCORI-funded projects. In November of 2016, the NIH issued a request for information on data management and sharing strategies in order to consider (1) how digital scientific data generated from NIH-funded research should be managed and, to the fullest extent possible, made publicly available and (2) how to set standards for citing shared data and software.⁶² This effort to improve data sharing spread to the private sector when, in July 2017, the International Committee of Medical Journal Editors required data sharing statements and data sharing plans as part of new conditions for publication of certain studies.⁶³ Clearly, data sharing and data access to research results will remain priority policy issues for years to come, and CRI will be an essential enabler.

Still, the systemic issues identified in the Blue Ribbon report will require ongoing work, much of which can be done as part of implementing the 21st Century Cures Act, including a Research Policy Board as well as a Working Group on Research Rigor and Reproducibility. In addition, Cures requires reports on the research workforce and compliance with ClinicalTrials.gov be developed before 2020.

Emerging Policy Trends in CRI

The last several years have seen a resurgence of clinical research policies and programs. Indeed, the amount of funding and support for clinical and biomedical research – even in these austere times – is significant. With more funding comes more accountability and higher expectations for innovation, and CRI is primed to deliver on both.

There is a growing appreciation for the need to coordinate national research infrastructure and resources, and programs such as All of Us are positioned to drive increased demand for CRI tools and methods for the foreseeable future. Further,

⁶¹ PCORI Policy for Data Access and Data Sharing, Draft for Public Comment. October 2016. Available at: <https://www.pcori.org/sites/default/files/PCORI-Data-Access-Data-Sharing-DRAFT-for-Public-Comment-October-2016.pdf>.

⁶² NIH Request for Information (RFI): Strategies for NIH Data Management, Sharing, and Citation. Nov. 14, 2016. Available at: <https://grants.nih.gov/grants/guide/notice-files/NOT-OD-17-015.html>.

⁶³ Taichman D. Data sharing statements for clinical trials: a requirement of the international committee of medical journal editors. Ann Intern Med. <https://doi.org/10.7326/M17-1028>.

clinical data generated across hospital and physician offices through EHRs present the research enterprise with unprecedented opportunities to increase our knowledge of health and disease.

Meanwhile, the adequacy of federal research policy is an ongoing conversation. A new regulatory framework was issued by the National Academies of Science, Engineering, and Medicine in 2016.⁶⁴ The 280-page report paints a disquieting picture of a stressed federal-academic partnership, concluding “The regulatory regime (comprising laws, regulations, rules, policies, guidances, and requirements) governing federally funded academic research should be critically reexamined and recalibrated.”

Policy is not made in a vacuum. Capitalizing on the numerous and extraordinary opportunities to improve development and delivery of new interventions will depend heavily on the application of CRI. It is vital that students of CRI understand and engage with the policymaking process.

⁶⁴National Academies of Sciences, Engineering, and Medicine. *Optimizing the Nation's investment in academic research: a new regulatory framework for the 21st century*. Washington, DC: The National Academies Press; 2016. <https://doi.org/10.17226/21824>.



Informatics Approaches to Participant Recruitment

6

Chunhua Weng and Peter J. Embi

Abstract

Clinical research is essential to the advancement of medical science and is a priority for academic health centers, research funding agencies, and industries working to develop and deploy new treatments. In addition, the growing rate of biomedical discoveries makes conducting high-quality and efficient clinical research increasingly important. Participant recruitment continues to represent a major bottleneck in the successful conduct of human studies. Barriers to clinical research enrollment include patient factors and physician factors, as well as recruitment challenges added by patient privacy regulations such as the Health Insurance Portability and Accountability Act (HIPAA) in the USA. Another major deterrent to enrollment is the challenge of identifying eligible patients, which has traditionally been a labor-intensive procedure. In this chapter, we review the informatics interventions for improving the efficiency and accuracy of eligibility determination and trial recruitment that have been used in the past and that are maturing as the underlying technologies improve, and we summarize the common sociotechnical challenges that need continuous dedicated work in the future.

Keywords

Internet-based patient matching systems · Research recruitment workflows
Informatics interventions in clinical research recruitment · Computerized clinical trial · EHR-based recruitment

C. Weng, PhD (✉)

Department of Biomedical Informatics, Columbia University, New York, NY, USA

e-mail: chunhua@columbia.edu

P. J. Embi, MD, MS

Regenstrief Institute, Inc, and Indiana University School of Medicine, Indianapolis, IN, USA

e-mail: pembi@regenstrief.org

Clinical research is essential to the advancement of medical science and is a priority for academic health centers, research funding agencies, and industries working to develop and deploy new treatments [1, 2]. In addition, the growing rate of biomedical discoveries makes conducting high-quality and efficient clinical research increasingly important. Participant recruitment continues to represent a major bottleneck in the successful conduct of human studies. Failure to meet recruitment goals can impede the development and evaluation of new therapies and can increase costs to the healthcare system. According to recent data, a clinical trial averages \$124 million and takes more than a decade to complete per drug candidate [3], with half of this time spent on patient, site, and investigator recruitment [4]. It has also been noted that 86% of all clinical trials are delayed in patient recruitment for 1–6 months and that 13% are delayed by more than 6 months [5, 6]. Indeed, inefficient recruitment processes threaten the success of clinical research and can have a range of effects including delayed study completion, trial failure, weakened results, introduction of bias, increased costs, slowing of scientific progress, and limiting the availability of beneficial therapies.

Barriers to clinical research enrollment include patient factors [7] (please reference Robert Califf's work here https://www.researchamerica.org/sites/default/files/July2017ClinicalResearchSurveyPressReleaseDeck_0.pdf) and physician factors [8], as well as recruitment challenges added by patient privacy regulations such as the Health Insurance Portability and Accountability Act (HIPAA) in the USA. Another major deterrent to enrollment is the challenge of identifying eligible patients, which has traditionally been a labor-intensive procedure. Studies have shown that 60–95% of the eligible patients often go unidentified [9, 10] and consequently miss the opportunity to participate in research studies. To overcome research recruitment challenges, informatics approaches have been developed and have demonstrated their potential to improve clinical research recruitment efficiency. In this chapter, we review the informatics interventions for improving the efficiency and accuracy of eligibility determination and trial recruitment that have been used in the past and that are maturing as the underlying technologies improve, and we summarize the common sociotechnical challenges that need continuous dedicated work in the future.

Typical Clinical Research Recruitment Workflows

Over the past 20 years, many efforts have been made to address the challenges involved in clinical trial recruitment and have been applied to major stakeholders in the recruitment process: investigators, patients, and healthcare providers. Many efforts to improve the awareness of clinical trials among physicians, patients, and the public have been pursued, ranging from distribution of paper and electronic flyers by trial centers to direct-to-consumer advertising and to the use of government and privately sponsored websites (Fig. 6.1). In addition, patients can now be matched to trials and trials to patients by information-based computer programs using computer-based protocol systems, electronic health records, web-based trial

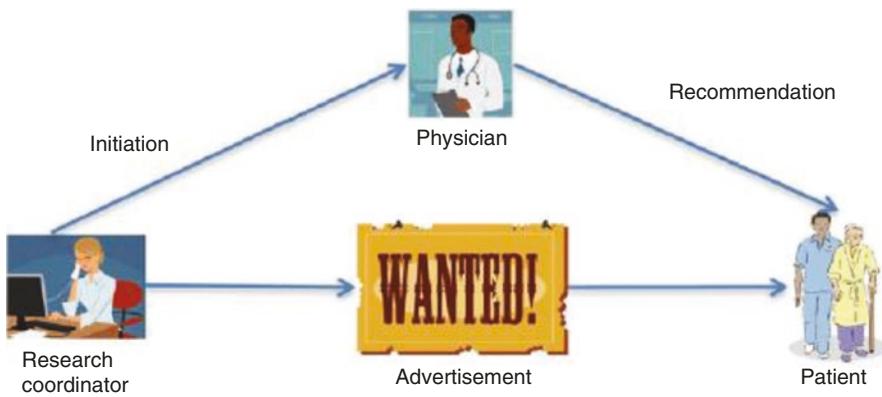


Fig. 6.1 Traditional researcher-initiated trial recruitment workflow



Fig. 6.2 Patient-initiated research recruitment workflow

matching tools, clinical data repositories or warehouses, or clinical registries [11–16] (Figs. 6.2, 6.3, and 6.4). Figures 6.2, 6.3, and 6.4 show three common recruitment workflows initiated by investigators, physicians, and patients, respectively. Accepting Dr. Robert Califf's assertion that "clinical research sites are the underappreciated component of the Clinical Research System," [17] then by extension clinical research coordinators are central to all of these three workflows. The simplest among the three is the workflow initiated by patients involving web-based trial matching (Fig. 6.2), which provides direct links between patients and research coordinators but also presents challenges such as discrepant health literacy levels among patients, heterogeneous data representations provided by different patients, and data incompleteness. Subsequently, the results are not fine-grained recommendations and need manual filtering. The workflow initiated by physicians (Fig. 6.3) has medium efficiency and complexity. The challenge for this recruitment mode lies in providing appropriate incentive for physicians to help with clinical research recruitment in their tight patient care schedules. The workflow using the clinical data warehouse (Fig. 6.4) is the most complicated among the three because it involves requests and queries initiated by investigators, permissions by care providers, and consent by patients. However, the highest positive predictive accuracy for trial screening is achieved by leveraging the data repositories. The following discussion will be focused on the three information-based recruitment workflows in the chronological order of their occurrence (Figs. 6.2, 6.3, and 6.4).

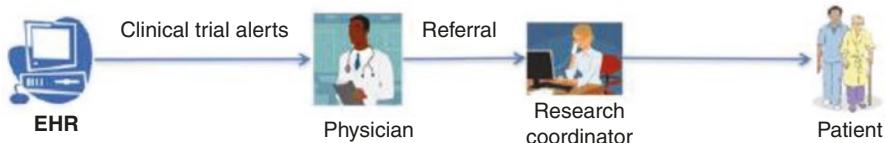


Fig. 6.3 Healthcare provider-initiated research recruitment workflow

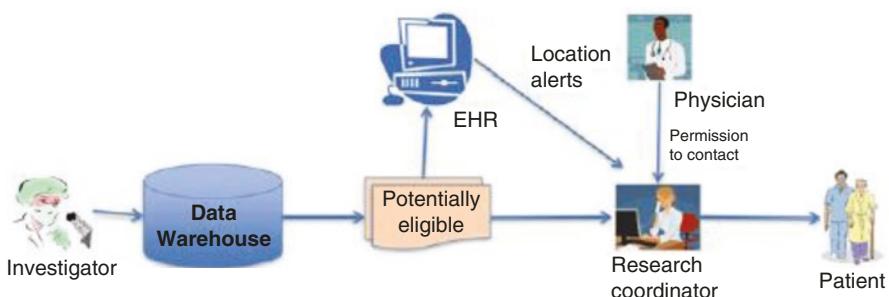


Fig. 6.4 Data warehouse-aided research screening and recruitment model

Informatics Interventions in Clinical Research Recruitment

Computerized Clinical Trial Decision Support

As early as the late 1980s, researchers have been seeking computational solutions to improving clinical research recruitment. Since protocol is at the heart of every clinical trial [18], earlier work largely concentrated on providing decision support to investigators through computer-based clinical research protocol systems [9, 11, 15, 19–21]. T-Helper was the earliest ontology-based eligibility screening decision support system [20] that offered patient-specific and situation-specific advice concerning new protocols for which patients might be eligible. Later, Tu et al. developed a comprehensive and generic problem solver [15] for eligibility decision support using Protégé [22]. Gennari et al. extended Tu et al.'s work and developed the EligWriter to support knowledge acquisition of eligibility criteria and to assist with patient screening [19]. Ohno-Machado et al. addressed uncertainty issues in eligibility determination and divided knowledge representations for eligibility criteria into three levels [23]: (1) the classification level, where medical concepts are modeled; (2) the belief network level, where uncertainty related to missing values are modeled; and (3) the control level that represents procedural knowledge and stores information regarding the connections between the other two levels, predefined information retrieval priorities, and protocol-specific information [23]. Other approaches include decision trees [11, 21], Bayesian networks [24, 25], and web-based interactive designs [9]. DS-TRIEL [11] used a handheld computer to match eligibility criteria represented to patient data entered by human experts using a

decision tree. OncoDoc [21] was a guideline-based eligibility screening system for breast cancer trials in which users could browse eligibility criteria represented as decision trees in the context of patient information. Cooper et al. used Bayesian networks to select a superset of patients with certain hard-coded characteristics from a clinical data repository [25]. Fink et al. developed an expert system for minimizing the total screening cost needed to determine patient eligibility [9].

In the 1990s, Musen et al. tested the T-helper system, designed to help community-based HIV/AIDS practitioners manage patients and adhere to clinical trial protocols. Their investigations revealed that many patients eligible for ongoing trials were overlooked [26, 27]. In their 1995 manuscript, Carlson et al. concluded, “The true value of a computer-based eligibility screening system such as ours will thus be recognized only when such systems are linked to integrated, computer-based medical-record systems” [27]. In a move toward that end, Butte et al. made use of a locally developed automated paging system to alert a trial’s coordinator when a potentially eligible patient’s data were entered into a database upon presentation to an emergency department [28, 29]. This approach was effective at increasing referrals for certain trials in that particular setting [30]. In another approach, Afrin et al. combined the use of paging and email systems linked to a healthcare system’s laboratory database to identify patients who might be eligible for an ongoing trial and then to notify the patient’s physician [31]. The system complied with privacy regulations and was successful in signaling the patient’s physician most of the time. However, most physicians did not follow up on the alerts, likely owing to the fact that the alert took place outside the context of the patient encounter and relied on the physician initiating contact with the patient after the visit had concluded, events that might be expected to reduce effectiveness.

Internet-Based Patient Matching Systems

Before the broad adoption of computer-based medical records systems as hoped for by Carlson et al., another technology revolution emerged that introduced new opportunities for improving clinical research recruitment: the Internet. With the penetration of the Internet starting in the mid-1990s, clinical research opportunities have been presented to more and more patients through online health information. Patient-enabling tools have emerged to help patients find relevant clinical research trials. Physician Data Query (PDQ) is a comprehensive trial registry database created by the National Cancer Institute (NCI) for patients to search for trials using stage, disease, and patient demographics [32]; however, PDQ does not support trial screening based on lab tests or detailed patient information. The search results often have low specificity and need further filtering. Ohno-Machado et al. developed an XML-based eligibility criteria database to support trial filtering for patients [12]. This system, known as the caMatch project, is a more recent Internet-based, patient-centric clinical trial eligibility matching application conceived by patient advocates [33] with a focus on developing common data elements for eligibility criteria rather than on automatic mass screening. It requires patients to build online personal health

records to be matched to structured eligibility criteria [34]. Niland also developed the Agreement on Standardized Protocol Inclusion Requirements for Eligibility (ASPIRE) to help cancer patients search for highly relevant clinical trials online [35]. Trialx (<http://www.trialx.com>) is another new web-based tool matching patients to trials using semantic web technologies [36]. In the past few years, Harris extended a local research registry for engaging the patient community for research participation into a national registry (ResearchMatch.org) to link patients, investigators, and clinical trials for the USA [37]. Both Trialx and ResearchMatch provide location-aware trial recommendation to patients over the Internet. As the semantic web technologies and the next generation of Web mature, more and more Internet-based research recruitment and patient education opportunities will undoubtedly emerge. With the wide implementation of patient portals connected to many EHR systems, patient portals promise to be an effective approach for recruitment [38], because patient information can be matched with clinical research eligibility criteria for automated eligibility determination and clinical study recommendation. This creates opportunities for novel informatics methodology development for matching research opportunities with patients.

Electronic Health Records-Based Recruitment Support

So far, the above interventions largely rely on matching structured entry of limited patient data elements to structured protocol eligibility criteria. While they are appropriate for providing patient-specific recommendations, some of them may not be practical for large-scale mass screening due to the lack of patient details for high-accuracy trial matching and the laborious, error-prone patient data entry process. In recent years, the adoption of electronic health records (EHRs) in both hospitals and private practice has been rising steadily, with 50% of US hospitals currently using EHR systems [3]. EHR systems contain rich patient information and are a promising resource for mass screening for clinical research by physicians. However, relatively few physicians contribute to research recruitment due to various barriers, including the lack of time and technical limitations of existing systems. To make participating in the recruitment process easier for non-researcher clinicians, Embi et al. pioneered methods to generate EHR-based clinical trial alerts (CTAs). These point-of-care alerts build on and repurpose clinical decision support tools to alert clinicians when they encounter a patient who might qualify for an ongoing trial, and they enable a physician to quickly and unobtrusively connect a patient with a study coordinator, all while being HIPAA compliant [39]. The CTA intervention has now been associated in multiple studies with significant increases both in the number of physicians generating referrals and enrollments and in the rates of referrals and enrollments themselves. Indeed, during Embi et al.'s initial CTA intervention study applied to a study of type 2 diabetes mellitus, the CTA intervention was associated with significant increases in the number of physicians generating referrals (5 before and 42 after; $P = 0.001$) and enrollments (5 before and 11 after; $P = 0.03$), a tenfold increase in those physicians' referral

rate (5.7/month before and 59.5/month after; rate ratio, 10.44; 95% confidence interval, 7.98–13.68; $P = 0.001$), and a doubling of their enrollment rate (2.9/month before and 6.0/month after; rate ratio, 2.06; 95% confidence interval, 1.22–3.46; $P = 0.007$). Moreover, a follow-up survey of physicians' perceptions of this informatics intervention [40] indicated that most physicians felt that the approach to point-of-care trial recruitment was easy to use and that they would like to see it used again. The CTA approach has subsequently been tested in other venues, further demonstrating improvements to recruitment rates [41–43].

Data Repository-Based Clinical Trial Recruitment Support

Another promising intervention for mass screening is the use of data repositories or data warehouses. In fact, automation of participant identification by leveraging large data repositories dates back to the early 1990s. With the increasing adoption of EHRs worldwide, many institutions have been able to aggregate data collected from EHRs into clinical data warehouses to support intelligent data analysis for administration and research. Kamal et al. developed a web-based prototype using an information warehouse to identify eligible patients for clinical trials [44]. Thadani et al. demonstrated that electronic screening for clinical trial recruitment using a Columbia University Clinical Data Warehouse reduced the manual review effort for the large randomized trial ACCORD by 80% [45]. Compared with EHRs, data warehouses are often optimized for efficient cross-patient queries and can be linked to computer-based clinical research decision support systems, such as alerts systems, to facilitate recruitment workflow. Furthermore, Weng et al. compared the effectiveness of a diabetes registry and a clinical data warehouse for improving recruitment for the diabetes trial TECOS [46]. Clinical registries are created for clinicians with disease-specific information; they are easy to use and contain information of simplicity and better quality. For example, not all diabetic patients identified using the clinical data warehouse have regular A1C measurement; therefore, applying A1C eligibility criteria on these patients with incomplete data to determine their eligibility is difficult. The diabetic patients identified using the diabetes registry, on the other hand, often do have regular A1C measurements due to the requirements of establishing clinical registries to improve quality monitoring of chronic diseases like diabetes. However, the results showed that the registry generated so many false-positive recommendations that the research team could not complete the review of the recommended patients. The data warehouse, though, generated an accurate, short patient list that helped the researcher become the top recruiter in the USA for this study. Weng et al. concluded that a clinical data warehouse in general contains the most comprehensive patient, physician, and organization information for applying complex exclusion criteria and can achieve higher positive predictive accuracy for electronic trial screening. The only disadvantage is that its use mandates approvals from the institutional review board (IRB) and sophisticated database query skills, which are barriers for clinical researchers or physicians wishing to use it directly for trial recruitment.

Sociotechnical Challenges

The availability of electronic patient information by itself does not entail an easy solution. There are regulatory, procedural, and technical challenges. Regulatory barriers for using electronic trial screening primarily come from HIPAA. HIPAA forbids nonconsensual release of patient information to a third party not involved with treatment, payment, or other routine operations associated with the provision of healthcare to the patient; therefore, concerns regarding privacy represent a growing barrier to electronic screening for clinical trials accrual [47]. In addition, technical barriers, including heterogeneous data representations and poor data quality (e.g., incompleteness, inconsistency, and fragmentation), pose the primary challenges for EHR-based patient eligibility identification [48, 49]. Moreover, differences in EHR implementation represent another roadblock with respect to the reuse of computer-based eligibility queries across different institutions. Parker and Embley developed a system to automatically generate medical logical modules in Arden syntax for clinical trials eligibility criteria [50]; however, queries represented in Arden syntax have the “curly braces problem” because the syntactic construct included in curly braces has to be changed for each site specifically [51], which could entail considerable knowledge engineering costs. In addition, poor data quality, unclear information sources, and incomplete data elements all contribute to making eligibility determination difficult [52]. Inconsistent data representations (both terminology and information model) are significant barriers to reliable patient eligibility determination. Weng et al. found significant inconsistency between structured and unstructured data in EHRs [53, 54], which posed great challenges for reusing clinical data for recruitment. Data incompleteness is another serious problem. Criteria such as “life expectancy greater than 3 months” or “women who are breast feeding” are often unavailable in EHRs. As Kahn has observed [55], EHR systems configured to support routine care do well identifying patients using only demographics and lab tests but do poorly with diagnostic tests and questionnaires [55]. Moreover, oftentimes patients are subsequently found ineligible at detailed screening because of treatment regimens or other factors that are exclusion factors in the protocol. Heterogeneous semantic representation is perhaps the greatest technical challenge. While EHRs or data warehouses all typically contain continuous variables, time-series tracings, and text, these rich data are not stored in a consistent manner for decision support, such as identifying eligible patients for clinical trials. For example, one EHR implementation might enter “abdominal rebound pain” as a specific nominal variable with value “YES,” and another might provide only the option of entering “abdominal pain” as free text or store a value on a visual analogue scale from 1 to 10. Hence, Chute asserts that eligibility determination using electronic patient information is essentially a problem of phenotype retrieval, whose big challenge is the semantic boundary that characterizes the differences between two descriptions of an object by different linguistic representations [56]. A challenge for the implementation of EHRs or data warehousing for clinical research recruitment is the semantic gulf between clinical data and clinical trial eligibility criteria. No single formalism is capable of representing the variety of eligibility

rules and clinical statements that we can find in clinical databases [57]. More research is needed to identify: (1) common manual tasks and strategies involved to craft EHR-based data queries for complex eligibility rules; (2) the broad spectrum of complexities in eligibility rules; (3) the breadth, depth, and variety of clinical data; and (4) the coverage of current terminologies in the concepts of eligibility criteria. As there is a significant distinction between high-level classifications (such as the ICDs) from detailed nomenclatures (such as SNOMEDCT) [58], in order to bridge the semantic gap between eligibility concepts and clinical manifestations in EHRs, we need to address the divergence and granularity discrepancies across different data encoding standards in our proposed research.

Also, a data-centric approach is indispensable to any e-clinical solution, but no existing approach has appeared to have the robust data connectivity required for data-driven clinical trials' mass screening. Thorough coverage of existing knowledge representation for eligibility criteria can be found in Weng et al.'s literature review [59]. Natural language processing (NLP) is a high-throughput technology that formalizes the grammar rules of a language in algorithms, then extracts data and terms from free text documents, and converts them into an encoded representation. Medical language processing (MLP) is NLP in the medical domain [60]. MLP has demonstrated its broad uses for a variety of applications, such as extracting knowledge from medical literature [61, 62], indexing radiology reports in clinical information systems [63–65], and abstracting or summarizing patient characteristics [66]. One of the widely used tools is MetaMap Transfer (MMTx) [67], which is available to biomedical researchers in a generic, configurable environment. It maps arbitrary text to concepts in the UMLS Metathesaurus [68]. Chapman demonstrated in her studies that MLP is superior to ICD-9 in detecting cases and syndromes from chief complaint reports [69, 70]; this finding was also confirmed by Li et al. in a study comparing discharge summaries and ICD-9 codes for recruitment uses [54]. The most mature MLP system is MedLEE [71]. In numerous evaluations carried out by independent users, MedLEE performed well [72]. To date, MedLEE is one of the most comprehensive operational NLP systems formally shown to be as accurate as physicians in interpreting narrative patient reports in medical records. EHR systems contain much narrative clinical data. The cost and effort associated with human classification of such data is not a scalable or sustainable undertaking in modern research infrastructure [58]. For this reason, it is well-recognized that we need NLP such as MedLEE to structure clinical data for trial recruitment.

Conclusion and Future Work

Ongoing attempts to use electronic patient information for patient eligibility determination underscore a great need for a long-range research plan to design and evaluate different methods to surmount the social, (https://www.researchamerica.org/sites/default/files/July2017ClinicalResearchSurveyPressReleaseDeck_0.pdf) organizational, and technical challenges facing clinical trial recruitment, the key

components of the plan being (1) to improve the data accuracy and completeness for EHR systems; (2) to design better data presentation techniques for EHR systems to enable patient-centered, problem-oriented data presentation; (3) to reduce ambiguities and to increase the computability of clinical research eligibility criteria; (4) to develop automatic methods for aligning the semantics between eligibility criteria and clinical data in EHRs; and (5) to integrate clinical research and patient care workflows to support clinical and translational research. The culmination of EHR-based recruitment efforts demonstrates that effort should be made to facilitate collaboration and workflow support between clinical research and patient care, which unfortunately still represent two distinct, disconnected processes and which divide professional communities and organizational personnel and regulations. Inadequate interoperability of workflow processes and electronic systems between clinical research and patient care can lead to costly, redundant tests and visits and to dangerous drug-drug interactions. In 2009, Conway and Clancy suggested that “use of requisite research will be most efficient and relevant if generated as a by-product of care delivery” [73]. A meaningful fusion of clinical care and research workflows promises to avoid conflicts, to improve safety and efficiency for clinical research [3], and to make EHR-based research more efficient and productive.

The ongoing All of Us program has an ambitious goal of recruiting one million diverse patients across the USA to collect comprehensive data from them in support of precision medicine. This program employs comprehensive community and patient engagement methods to recruit the public through various channels, including social media, clinics, churches, supermarkets, libraries, and so on. Anyone can sign up online or consent at a clinic to participate in the study. For patients recruitment through clinics or health provider organizations, their electronic health records data can flow to the recruitment system seamlessly, which is a big step ahead of many prior clinical trial recruitment effort. Efforts to return research results to participants and invite participants to join as research partners to contribute research questions are new features of this study that differentiates it from conventional clinical studies, which are also responsive to the new culture of patient-centered research.

In addition to EHR-based recruitment that is led by clinical research teams, parallel efforts that are more patient-centered have also been growing rapidly in recent years. For example, Apple launched an open-source framework called ResearchKit that promises to reach a large group of iPhone users to facilitate rapid recruitment and robust data collection, although this approach still needs to deal with informed consent and population biases challenges. In the future, as we collect more data electronically and in a standardized way, and as we increase our ability to collect and continuously track patient eligibility over a patient’s lifetime, we can think more about watching patients that may not originally be eligible for a study at one point but may become eligible based over their lifetime or course of health or disease state. Better data collection and standards can allow more precise targeting of patients for recruitment. With the rising of citizen science, we foresee to see more patient-driven research design and recruitment systems to be in the norm in the future.

References

1. Nathan DG, Wilson JD. Clinical research and the NIH – a report card. *N Engl J Med.* 2003;349(19):1860–5.
2. Campbell EG, Weissman JS, Moy E, Blumenthal D. Status of clinical research in academic health centers: views from the research leadership. *JAMA.* 2001;286(7):800–6.
3. Mowry M, Constantinou D. Electronic health records: a magic pill? *Appl Clin Trials.* 2007; 2(1). <http://appliedclinicaltrialsonline.findpharma.com/appliedclinicaltrials/article/articleDetail.jsp?id=401622>.
4. Canavan C, Grossman S, Kush R, Walker J. Integrating recruitment into eHealth patient records. *Appl Clin Trials.* 2006.
5. Sinackevich N, Tassignon J-P. Speeding the critical path. *Appl Clin Trials.* 2004;31:241–54.
6. Sullivan J. Subject recruitment and retention: barriers to success. *Appl Clin Trials.* 2004.
7. Schain W. Barriers to clinical trials, part 2: knowledge and attitudes of potential participants. *Cancer.* 1994;74:2666–71.
8. Mansour E. Barriers to clinical trials, part 3: knowledge and attitudes of health care providers. *Cancer.* 1994;74:2672–5.
9. Fink E, Kokku PK, Nikiforou S, Hall LO, Goldgof DB, Krischer JP. Selection of patients for clinical trials: an interactive web-based system. *Artif Intell Med.* 2004;31(3):241–54.
10. Carlson R, Tu S, Lane N, Lai T, Kemper C, Musen M, Shortliffe E. Computer-based screening of patients with HIV/AIDS for clinical-trial eligibility. *Online J Curr Clin Trials.* 1995. Doc No 179.
11. Breitfeld PP, Weisbord M, Overhage JM, Sledge G Jr, Tierney WM. Pilot study of a point-of-use decision support tool for cancer clinical trials eligibility. *J Am Med Inform Assoc.* 1999;6(6):466–77.
12. Ash N, Ogunyemi O, Zeng Q, Ohno-Machado L. Finding appropriate clinical trials: evaluating encoded eligibility criteria with incomplete data. *Proc AMIA Symp.* 2001:27–31.
 13. Papaconstantinou C, Theocarous G, Mahadevan S. An expert system for assigning patients into clinical trials based on Bayesian networks. *J Med Syst.* 1998;22(3):189–202.
14. Thompson DS, Oberteuffer R, Dorman T. Sepsis alert and diagnostic system: integrating clinical systems to enhance study coordinator efficiency. *Comput Inform Nurs.* 2003;21(1):22–6; quiz 27–8.
15. Tu SW, Kemper CA, Lane NM, Carlson RW, Musen MA. A methodology for determining patients' eligibility for clinical trials. *Methods Inf Med.* 1993;32(4):317–25.
16. Ohno-Machado L, Wang SJ, Mar P, Boxwala AA. Decision support for clinical trial eligibility determination in breast cancer. *Proc AMIA Symp.* 1999:340–4.
17. Calif R. Clinical research sites – the underappreciated component of the clinical research system. *JAMA.* 2009;302(18):2025–7.
18. Kush B. The protocol is at the heart of every clinical trial. 2007. <http://www.ngpharma.com/pastissue/article.asp?art=25518&issue=143>. Accessed Aug 2011.
19. Gennari J, Sklar D, Silva J. Cross-tool communication: from protocol authoring to eligibility determination. In: Proceedings of the AMIA'01 symposium, Washington, DC; 2001. p. 199–203.
20. Musen MA, Carlson RW, Fagan LM, Deresinski SC. T-HELPER: automated support for community-based clinical research. In: 16th annual symposium on computer applications in medical care, Washington, DC; 1992.
21. Seroussi B, Bouaud J. Using OncoDoc as a computer-based eligibility screening system to improve accrual onto breast cancer clinical trials. *Artif Intell Med.* 2003;29(1): 153–67.
22. Protege. 2007. <http://protege.stanford.edu/>. Accessed Aug 2011.
23. Ohno-Machado L, Parra E, Henry SB, Tu SW, Musen MA. AIDS2: a decision-support tool for decreasing physicians' uncertainty regarding patient eligibility for HIV treatment protocols. In: Proceedings of 17th annual symposium on computer applications in medical care, Washington, DC; 1993. p. 429–33.

24. Aronis J, Cooper G, Kayaalp M, Buchanan B. Identifying patient subgroups with simple Bayes. Proc AMIA Symp. 1999;658–62.
25. Cooper G, Buchanan B, Kayaalp M, Saul M, Vries J. Using computer modeling to help identify patient subgroups in clinical data repositories. Proc AMIA Symp. 1998;180–4.
26. Musen MA, Carlson RW, Fagan LM, Deresinski SC, Shortliffe EH. T-HELPER: automated support for community-based clinical research. Proc Annu Symp Comput Appl Med Care. 1992;719–23.
27. Carlson RW, Tu SW, Lane NM, Lai TL, Kemper CA, Musen MA, Shortliffe EH. Computer-based screening of patients with HIV/AIDS for clinical-trial eligibility. Online J Curr Clin Trials. 1995;Doc No 179:[3347 words; 3332 paragraphs].
28. Weiner DL, Butte AJ, Hibberd PL, Fleisher GR. Computerized recruiting for clinical trials in real time. Ann Emerg Med. 2003;41(2):242–6.
29. Butte AJ, Weinstein DA, Kohane IS. Enrolling patients into clinical trials faster using RealTime Recruiting. Proc AMIA Symp. 2000;111–5.
30. U.S. Health Insurance Portability and Accountability Act of 1996. <http://www.cms.gov/HIPAAGenInfo/Downloads/HIPAALaw.pdf>. Accessed Aug 2011.
31. Afrin LB, Oates JC, Boyd CK, Daniels MS. Leveraging of open EMR architecture for clinical trial accrual. Proc AMIA Symp. 2003;2003:16–20.
32. Physician Data Query (PDQ). 2007. <http://www.cancer.gov/cancertopics/pdq/cancerdatabase>. Accessed Aug 2011.
33. Assuring a health dimension for the National Information Infrastructure: a concept paper by the National Committee on Vital Health Statistics. Presented to the US Department of Health and Human Services Data Council, Washington, DC; 1998.
34. Cohen E, et al. caMATCH: a patient matching tool for clinical trials, caBIG annual meeting, Washington, DC; 2005.
35. Niland J. Integration of Clinical Research and EHR: eligibility coding standards, podium presentation to the 2010 AMIA Clinical Research Informatics Summit meeting, San Francisco; http://crisummit2010.ami.org/files/symposium2008/S14_Niland.pdf, Accessed on 13 Dec 2011.
36. Trialx. 2010. <http://www.trialx.com>. Accessed Aug 2011.
37. Harris PA, Lane L, Biaggioni I. Clinical research subject recruitment: the volunteer for Vanderbilt research program www.vanderbilthhealth.com/clinicaltrials/13133. J Am Med Inform Assoc. 2005;12(6):608–13.
38. Samuels MH, et al. Effectiveness and cost of recruiting healthy volunteers for clinical research studies using an electronic patient portal: a randomized study. J Clin Transl Sci [Internet]. 2017;1(6):366–72. 2018/04/23. Cambridge University Press.
39. Embi PJ, Jain A, Clark J, Bizjack S, Hornung R, Harris CM. Effect of a clinical trial alert system on physician participation in trial recruitment. Arch Intern Med. 2005;165:2272–7.
40. Embi PJ, Jain A, Harris CM. Physicians' perceptions of an electronic health record-based clinical trial alert approach to subject recruitment: a survey. BMC Med Inform Decis Mak. 2008;8:13.
41. Embi PJ, Lieberman MI, Ricciardi TN. Early development of a clinical trial alert system in an EHR used in small practices: toward generalizability. AMIA Spring Congress. Phoenix; 2006.
42. Rollman BL, Fischer GS, Zhu F, Belnap BH. Comparison of electronic physician prompts versus waitroom case-finding on clinical trial enrollment. J Gen Intern Med. 2008;23(4):447–50.
43. Grundmeier RW, Swietlik M, Bell LM. Research subject enrollment by primary care pediatricians using an electronic health record. AMIA Annu Symp Proc. 2007;2007:289–93.
44. Kamal J, Pasuparthi K, Rogers P, Buskirk J, Mekhjian H. Using an information warehouse to screen patients for clinical trials: a prototype. Proc of AMIA. 2005:1004.
45. Thadani SR, Weng C, Bigger JT, Ennever JF, Wajngurt D. Electronic screening improves efficiency in clinical trial recruitment. J Am Med Inform Assoc. 2009;16(6):869–73.
46. Weng C, Bigger J, Busacca L, Wilcox A, Getaneh A. Comparing the effectiveness of a clinical data warehouse and a clinical registry for supporting clinical trial recruitment: a case study. Proc AMIA Annu Fall Symp. 2010:867–71.

47. Sung NS, Crowley WF Jr, Genel M, Salber P, Sandy L, Sherwood LM, Johnson SB, Catanese V, Tilson H, Getz K, Larson EL, Scheinberg D, Reece EA, Slavkin H, Dobs A, Grebb J, Martinez RA, Korn A, Rimoin D. Central challenges facing the national clinical research enterprise. *JAMA*. 2003;289(10):1278–87.
48. Van Spall HGC, Toren A, Kiss A, Fowler RA. Eligibility criteria of randomized controlled trials: a systematic sampling review. *JAMA*. 2007;297(11):1233–40.
49. Musen MA, Rohn JA, Fagan LM, Shortliffe EH. Knowledge engineering for a clinical trial advice system: uncovering errors in protocol specification. *Bull Cancer*. 1985;74:291–6.
50. Parker CG, Embley DW. Generating medical logic modules for clinical trial eligibility criteria. *AMIA Annu Symp Proc*. 2003;2003:964.
51. Jenders R, Sujansky W, Broverman C, Chadwick M. Towards improved knowledge sharing: assessment of the HL7 Reference Information Model to support medical logic module queries. *AMIA Annu Symp Proc*. 1997:308–12.
52. Lin J-H, Haug PJ. Data preparation framework for preprocessing clinical data in data mining. *AMIA Annu Symp Proc*. 2006;2006:489–93.
53. Carlo L, Chase H, Weng C. Reconciling structured and unstructured medical problems using UMLS. *Proc AMIA Fall Symp*. 2010:91–5.
54. Li L, Chase HS, Patel CO, Friedman C, Weng C. Comparing ICD9-encoded diagnoses and NLP-processed discharge summaries for clinical trials pre-screening: a case study. *AMIA Annu Symp Proc*. 2008;2008:404–8.
55. Kahn MG. Integrating electronic health records and clinical trials. 2007. <http://www.esibethesda.com/hcrrworkshops/clinicalResearch/pdf/MichaelKahnPaper.pdf>. Accessed Aug 2011.
56. Lewis JR. IBM computer usability satisfaction questionnaires: psychometric evaluation and instructions for use. *Int J Hum-Comput Interact*. 1995;7(1):57.
57. Ruberg S. A proposal and challenge for a new approach to integrated electronic solutions. *Appl Clin Trials*. 2002;2002:42–9.
58. Chute C. The horizontal and vertical nature of patient phenotype retrieval: new directions for clinical text processing. *Proc AMIA Symp*. 2002:165–9.
59. Weng C, Tu SW, Sim I, Richesson R. Formal representation of eligibility criteria: a literature review. *J Biomed Inform*. 2010;43(3):451–67.
60. Friedman C, Hripcsak G. Natural language processing and its future in medicine. *Acad Med*. 1999;74:890–5.
61. Friedman C, Chen L. Extracting phenotypic information from the literature via natural language. *Stud Health Technol Inform*. 2004;107:758–62.
62. Friedman C, Kra P, Yu H, Krauthammer M, Rzhetsky A. GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. *Bioinformatics*. 2001;17(Supl 1):74–82.
63. Mendonca E, Haas J, Shagina L, Larson E, Friedman C. Extracting information on pneumonia in infants using natural language processing of radiology reports. *J Biomed Inform*. 2005;38(4):314–21.
64. Friedman C, Hripcsak G, Shagina L, Liu H. Representing information in patient reports using natural language processing and the extensible markup language. *J Am Med Inform Assoc*. 1999;6(1):76–87.
65. Friedman C, Shagina L, Lussier Y, Hripcsak G. Automated encoding of clinical documents based on natural language processing. *J Am Med Inform Assoc*. 2004;11(5):392–402.
66. Baud R, Lovis C, Ruch P, Rassinoux A. Conceptual search in electronic patient record. *Medinfo*. 2001;84:156–60.
67. Yasnoff WA, Humphreys BL, Overhage JM, Detmer DE, Brennan PF, Morris RW, Middleton B, Bates DW, Fanning JP. A consensus action agenda for achieving the national health information infrastructure. *J Am Med Inform Assoc*. 2004;11(4):332–8.
68. Brailey DJ. The decade of health information technology: delivering consumer-centric and information-rich health care. Framework for strategic action. 2004. <http://www.hhs.gov/healthit/frameworkchapters.html>. Accessed 31 Jan 2005.

69. Fiszman M, Chapman W, Aronsky D, Evans R, Haug P. Automatic detection of acute bacterial pneumonia from chest X-ray reports. *J Am Med Inform Assoc.* 2000;7:593–604.
70. Fiszman M, Chapman W, Evans S, Haug P. Automatic identification of pneumonia related concepts on chest x-ray reports. *Proc AMIA Symp.* 1999:67–71.
71. Friedman C. Towards a comprehensive medical language processing system: methods and issues. *Proc AMIA Annu Fall Symp.* 1997:595–9.
72. Hripcsak G, Friedman C, Alderson P, DuMouchel W, Johnson S, Clayton P. Unlocking clinical data from narrative reports: a study of natural language processing. *Ann Intern Med.* 1995;122(9):681–8.
73. Conway PH, Commentary CC. Transformation of health care at the front line. *JAMA.* 2009;301(7):763–5. <https://doi.org/10.1001/jama.2009.103>.



The Evolving Role of Consumers

7

James E. Andrews, J. David Johnson,
and Christina Eldredge

Abstract

The culmination of the changes in healthcare, motivated in many ways by the rapid evolution of information and communication technologies in parallel with the shift toward increased patient decision-making and empowerment, has critical implications for clinical research, from recruitment and participation to, ultimately, successful outcomes. This chapter explores the developments impacting health consumers from various perspectives, with some focus on foundational issues in health communication and information behaviors as related to health consumerism. An overarching concern is the information environment within which health consumers are immersed, which is increasingly social, and underlying communication issues and emerging technologies contributing to the changing nature of patients' information world. Not surprisingly, we will see that core findings from communication and information behavior research have relevance for our current understanding and future models of the evolving role of the health consumer.

Keywords

Health consumerism · Consumer health information · Consumer health movement · Patient empowerment · Patient engagement · Public access technologies · Personalization of medicine

J. E. Andrews, PhD (✉) · C. Eldredge, MD

School of Information, College of Arts and Sciences, University of South Florida,
Tampa, FL, USA

e-mail: jimandrews@usf.edu; celdredge2@usf.edu

J. David Johnson, PhD

Department of Communication, University of Kentucky, Lexington, KY, USA
e-mail: jdj@uky.edu

The premise is that we are at a new phase of health and medical care, where more decisions are being made by individuals on their own behalf, rather than by physicians, and that, furthermore, these decisions are being informed by new tools based on statistics, data, and predictions... We will act on the basis of risk factors and predictive scores, rather than on conventional wisdom and doctors' recommendations. We will act in collaboration with others, drawing on collective experience with health and disease... these tools will create a new opportunity and a new responsibility for people to act – to make health decisions well before they become patients.

Thomas Goetz, cited by Swan [1], from The Decision Tree, <http://thedecisiontree.com/blog/2008/12/introducing-the-decision-tree>

Overview

The role of patients as consumers has been evolving for well over a generation. In the past few decades, patients have transformed from passive to active participants both in clinical care and clinical research. Generally, the goal has been greater patient empowerment, defined by the World Health Organization (WHO) as “a process by which people, organizations and communities gain mastery over their affairs” [2] or more practically as “self-reliance through individual choice (consumer perspective)” [3]. As such, the responsibility for health-related matters is passing to the individual, partly because of legal decisions that have entitled patients to fuller information access. Ever since the 1970s, patients have become more active participants in decisions affecting healthcare [4]. Given the centrality of patients to clinical research, the evolution toward greater involvement and empowerment poses challenges and issues that stand to impact how clinical research might be conducted in the future and its ultimate success. Trends in health consumerism have fueled this movement, as have new tools such as decision-support-related advances and increasingly effective access to authoritative information resources, social networking capabilities, and personal decision aids. Moreover, an implication of this is that consumerism and empowerment assume, or even require, some level of health, information, and digital literacy on the part of consumers. This poses challenges for developers, researchers, and healthcare providers, given evolving consumer needs and expectations from the healthcare system, other patients, and, indeed, their relationships to their own information.

This culmination of changes in healthcare, motivated in many ways by the rapid evolution of information and communication technologies (ICTs) in parallel with the shift toward increased patient decision-making and empowerment, has critical implications for clinical research, from recruitment and participation to, ultimately, successful outcomes, precision medicine, and more rapid discovery. As noted, there is a growing onus on individuals to develop literacy skills (health, information, and digital) in order for all to fully realize the potential. For those who are incapable (or unwilling) to move in that direction, new structures or pathways will need to be created. This chapter explores these and other developments first via a broad look at some foundational issues in health communication as related to health consumerism. We also discuss the information environment within which health consumers

are immersed and the changing nature of patients' information world. Not surprisingly, we will see that core findings from communication and behavior research have relevance for our current understanding and future studies of the evolving role of the consumer. We also describe some emerging models and tools that seem to hold promise for helping usher in the next generation of clinical research consumers.

Traditional Perspectives: Health Campaigns and Information Behaviors

We know a lot about how formal organizations (for instance, from the National Cancer Institute, the American Cancer Society, and others) conduct campaigns to change individual behaviors [6]. Before some of the more recent transformative advances in ICTs, these traditional paradigms were used to inform information systems and strategy development and generally how we have thought about patient health behavior, including in clinical research contexts. Public health communication campaigns represent "purposive attempts to inform, persuade, or motivate behavior changes in a relatively well-defined and large audience, generally for non-commercial benefits to the individuals and/or society, typically within a given time period, by means of organized communication activities involving the mass media and often complemented by interpersonal support" [5]. Increasingly, however, individual actions, particularly health-related information seeking, determine what messages individuals will be exposed to and how they will behave.

In our view, actors operate in *information fields* (covered in more detail below) where they recurrently process resources and information. This field operates much like a market where individuals make choices (often based on only incomplete information and often irrationally) that determine how they will act regarding their health. This contrasts directly with the view of more traditional health information campaigns that tend to view the world as rational, known, and which concentrate on controlling individuals to seek values of efficiency and effectiveness [5].

A focus on information seeking develops a true receiver's perspective and forces us to examine how an individual acts within an information field containing multiple information carriers. Some of these carriers may be actively trying to reach individuals, but many contain passive information awaiting access and use. While there may be some commonalities across information fields, individuals' information environments are becoming so fragmented due to individual contextualizing that assessing media effects (or campaign ones) is increasingly difficult [7]. There is a commonplace recognition now that mass media alone is unlikely to have the desired impacts and that they must be supplemented with interpersonal communication as well as within social networks [8], thus giving rise to the near ubiquity of ICTs supporting social interactions and sharing.

Campaigns may result in felt needs on the part of the individual, but the individual and his or her placement in a particular social context will determine how needs are acted upon. An accurate picture of the impact of communication on health

needs to contain elements of both perspectives. Yet, most of the work in this area tilts in the direction of understanding more formal campaigns, with increasingly sophisticated methods [9, 10]; however, for our purposes the primary focus will be on how individuals make sense of the information fields within which they act. This focus on receivers dovetails nicely with the renewed focus on the patient as consumer, as expert, and as one seeking empowerment.

Traditional health communicators learned that classic approaches are not very effective unless the needs of the audience and their reaction to messages are considered [11, 12]. Thus, it soon became apparent that while there were some notable successes, audiences could be remarkably resistant to campaigns, especially when they did not correspond to the views of their immediate social network [13–16]. Indeed, campaigns tend to reach those who are already interested and typically bypass those who are most in need of their messages [15]. In effect, campaigns ironically reach those who are already converted. While this might have a beneficial effect of further reinforcing beliefs, the audience members who are most in need of being reached are precisely those members who are least likely to attend to health professionals' messages [13].

One area where the limitations of public health campaigns are most clearly revealed is in the difficulty and considerable expense involved in recruiting people into clinical research studies. According to Allison [17], less ~3% of eligible cancer patients enroll in trials, and roughly one in five of NCI-sponsored trials fail to meet their necessary enrollment [18]. One area where the trial recruitment challenge is particularly salient is rare diseases, where there are relatively low numbers of affected individuals who may be geographically dispersed. Even with new technologies to better match patients with trials or other health information, privacy and credibility underlie and potentially impede these efforts [19], and researchers must consider whether they are getting representative samples given that those seeking trials might disproportionately represent certain demographics [20]. The extremely low accrual rates in clinical research show that even within subsets of the population who might be eligible to participate in particular trials, the traditional “one size fits all” approach to health campaigns is insufficient. Expectations have understandably risen on the part of consumers, who have access to more targeted or even personalized information to assist them with such decisions and whose support groups may reinforce their natural predispositions.

We will discuss this context as it relates to patient recruitment further below (and already discussed recruitment in Chap. 5) after we look at other foundational issues.

A More Social World of Health Consumers and Information

A compelling development in consumer health over the past 15 years or more has been the emergence of a dynamic social world fueled further by WWW-based social media applications. The interactions and relationships among people, the evolving healthcare environment, technology, and information resources are incredibly complex and continually in flux. The frequently cited Pew Internet report on the social

life of health information showed that large percentages of adults seek health information online [21]. While most (86%) of all adults still continue to seek information from traditional sources (i.e., health professionals), the social world is “robust,” with more than half of online health information seekers doing so for someone else and discussing such information with others [21]. Online support groups are also showing signs of fostering patient empowerment or management [22] and participation tools that may lead to more positive outcomes, especially for rare diseases [23].

An overview of the context of previous communication and behavioral research on health consumers, including those who are engaged in or might consider participating in clinical research, is important as we consider the technologies and approaches that currently populate the landscape of consumerism and engagement in relation to clinical research. First, in this section, we present in greater detail the notion of information fields where health consumers are embedded. We then explore interpersonal interactions among individuals in social networks and the complex relationships and dynamics this presents despite emerging technologies.

Information Fields

One conception of an information environment is that of the information field within which the individual is embedded [24]. An individual’s information field provides the more static context for their information seeking, containing resources, constraints, and carriers of information [4, 25]. It provides the starting point for information seeking [26] representing the typical arrangement of information stimuli to which an individual is regularly exposed [7] and the information resources they routinely use [27]. Individuals are embedded in physical and virtual worlds that involve recurring contacts with an interpersonal network of friends and/or family (and, increasingly, strangers). They are also regularly exposed to the same mediated communication channels (company news bulletins, local newspapers, television news, and so on). The information field in which an individual is located constrains the very possibility of selecting particular sources of information.

The concept of field has a long tradition in the social sciences tracing back to the seminal work of Lewin [28] with interesting recent variants such as the information horizons approach [27]. Potential fields for patients have become incredibly richer over the last decade, providing them resources that can dramatically change their relationships with clinicians and researchers, as well as with patient advocacy groups and other health-related agencies and organizations.

People can, if they so desire, arrange the elements of their information field to maximize their surveillance of health information, providing an initial contextualizing of their environment. Individuals who are more concerned with their health are likely to mold their information fields to include a richer mixture of health-related information sources. How they shape this field over time determines not only their knowledge of general health issues but also their incidental exposure to information that may stimulate them to more purposive information seeking. The nature of an individual’s interpersonal environment, or social fields, has important

consequences for information seeking and for health practices [4]. Its importance is increasing with rising consumerism, a focus on prevention, self/home care, and a greater focus on individual responsibility. In a sense, individuals are embedded in a field that acts on them, the more traditional view of health campaigns. However, they also make choices about the nature of their fields, the types of media they attend to, the friendships they form and the neighborhoods they live in, and the social media they participate in, which are often based on their information needs and preferences which are greatly facilitated by the Internet and explosion of choices among even traditional media such as cable television and online media.

Naturally, an information field can be modified to reflect changes in an individual's life, which at times are also directly related to changing information seeking demands such as a pressing health problem. When an individual becomes a cancer patient, for instance, his or her interpersonal network changes to include other cancer patients who are proximate during treatment. They also may be exposed to a greater array of mediated communication (e.g., pamphlets, videos, and more tailored electronic communication—to name a few) concerning the nature of their diseases, treatment options, or availability of relevant clinical research studies. As individuals become more focused in their information seeking, they change the nature of their information field to support the acquisition of information related to particular purposes [29]. In this sense, individuals act strategically to achieve their ends and in doing so construct local communication structures in a field that mirrors their interests [30].

In some ways, the total of a person's information fields has analogies to the notion of social capital in that it describes the resource an individual has to draw upon when confronting a problem. When individuals share the same information field, the [31]. This sense of shared context is central in the development of online communities and related tools that have been growing in number in recent years and that extend the reach of one's effective social network through information behavior involving the development of weak ties.

Interpersonal Communication in Social Networks

There have been a number of studies that demonstrate a clear link between individuals' positioning in social networks and their health [32, 33]. These show there are four basic dynamics involved:

1. Lack of adequate social network ties worsens health, increasing demands for medical services.
2. Social networks shape beliefs and access to lay consultation.
3. Disruptions in social networks trigger help seeking.
4. Social networks moderate (or amplify) other stressors.

An individual's effective network is constituted by friends, family members, and other close associates, while an extended network is composed of casual

acquaintances and friends of friends who, because they have different contacts than the focal individual, can provide them with unique information. Effective networks impart normative expectations to individuals, and these expectations are often linked to behavioral intentions and actions that can represent convergence of network members around symbolic meanings of support [34, 35]. These networks, in effect, constitute elaborate feedback processes through which individual behavior is regulated and maintained [34, 35].

Social networks are often viewed as the infrastructure of social support with social support seen as "...inextricably woven into communication behavior" [34, 35]. Generally, two crucial dimensions of support are isolated, informational and emotional, with informational support being associated with a feeling of mastery and control over one's environment and emotional support being crucial to feelings of personal coping, enhanced self-esteem, and needs for affiliation [4]. Individuals need the social support of their immediate social networks to deal effectively with the disease and with the maintenance of long-term health behaviors [36], but they also need authoritative professional guidance in the institution of proper treatment protocols, search and selection of clinical trials, and comprehension of the most recent research.

However, interlocking personal networks lack openness (the degree to which a group exchanges information with the environment) and may simply facilitate the sharing of ignorance among individuals. "The degree of individual integration in personal communication networks is negatively related to the potential for information exchange" [37]. The degree to which individuals expand their networks and are encouraged to do so by members of their effective network has important consequences for health-related information acquisition and subsequent actions.

The strength of weak ties is perhaps the best-known concept related to network analysis. It refers to our less developed relationships that are more limited in space, place, time, and depth of emotional bonds [38]. This concept has been intimately tied to the flow of information. Weak ties' notions are derived from the work of Granovetter [39] on how people acquire information related to potential jobs. It turns out that the most useful information came from individuals in a person's extended networks, casual acquaintances, and friends of friends. This information was the most useful precisely because it comes from our infrequent or weak contacts. Strong contacts are likely to be people with whom there is a constant sharing of the same information; as a result, individuals within these groupings have come to have the same information base. Information from outside this base gives unique perspectives that may be crucial to confronting a newly developed health problem.

Weak ties provide critical informational support because they transcend the limitation of our strong ties and because, as often happens in sickness, our strong ties can be disrupted or unavailable [34]. In online support groups, weak ties might benefit participants (or have potentially negative consequences), given the disinhibition effect often referred to in online communication, where people are known to say or do things they would not normally do within closer networks [22]. As in other weak tie contexts, disinhibition can foster a sense of closeness, empathy, and kindness and a certain level of bonding that may break the inertia of

the fields in which an individual has habitually been embedded and introduce them to new individuals or third parties.

The Role of Third Parties in Information Seeking

There are a number of ways that use of third parties (for instance, knowledge brokers) can complement clinical practice and, by extension, participation in research. First, individuals who want to be fully prepared before they visit the doctor often consult the Internet [40, 41]. Lowery and Anderson [42] suggest that prior information use may impact respondents' perception of physicians. Second, there appears to be an interesting split among Internet users, with as many as 60% of users reporting that while they look for information, they only rely on it if their doctors tell them to [21, 41]. While the WWW makes a wealth of information available for particular purposes, it is often difficult for the novitiate to weigh the credibility of the information, a critical service that a knowledge broker, such as a clinical professional or consumer health librarian, can provide. This suggests that a precursor to a better patient-doctor dialogue would be to increase the public's knowledge base and to provide alternative but also complementary, information sources by shaping clients' information fields. To achieve behavioral change regarding health promotion, a message must be repeated over a long period via multiple sources [43]. By shaping and influencing the external sources a patient will consult both before and after visits, clinical practices can simultaneously reduce their own burden for explaining (or defending) their approach and increase the likelihood of patient compliance. Naturally, it is easy to see this all has implications for clinical research accrual, retention, and overall satisfaction.

Although intermediaries (e.g., navigators) play an important role despite an increase of more Web-based consumer health information, increasing health literacy by encouraging autonomous information seekers also should be a goal of our healthcare system [44]. While it is well known that individuals often consult a variety of others before presenting themselves in clinical or research settings [4] outside of HMO and organizational contexts, there have been few systematic attempts to shape the nature of these prior consultations. If these prior information searches happen in a relatively uncontrolled, random, parallel manner, expectations (e.g., treatment options, expected outcomes, diagnosis, trial retention and completion) may be established that will be unfulfilled.

The emergence of the WWW as an omnibus source of information also has apparently changed the nature of opinion leadership, as well; both more authoritative (e.g., medical journals and literature) and more interpersonal (e.g., support or advocacy groups) sources are readily available and accessible online [45]. This is part of a broader trend that Shapiro [46] refers to as "disintermediation," or the capability of the Web to allow the general public to bypass experts in their quest for information, products, and services. The risk here, however, is that individuals can quickly become overloaded or confused in an undirected environment. In other words, while the goal may be to reduce uncertainty or help bridge a knowledge gap,

the effect can be increased uncertainty and, ultimately, decreased sense of efficacy for future searching. A focus on promoting health information literacy, then, would mean helping people gain the skills to access, to judge the credibility of, and to effectively utilize a wide range of health information.

Increasing use of secondary information disseminators, or brokers, is really a variant on classic notions of opinion leadership [14] and gatekeepers [47] and instantiates weak ties [48]. Opinion leadership suggests ideas flow from the media to opinion leaders to those *less active* segments of the population serving a relay function, as well as providing social support information to individuals [48], reinforcing messages by their social influence over them [18], and validating the authoritativeness of the information [49]. So, not only do opinion leaders serve to disseminate ideas but they also, because of the interpersonal nature of their ties, provide additional pressure to conform as well [48, 50]. Another trend in this area is the recognition of human gatekeepers, community-based individuals who can provide information to at-risk individuals and refer them to more authoritative sources for treatments [4]. Recognizing the powers of peer opinion leaders, many health institutions are establishing patient advocacy programs, for example, where cancer survivors can serve to guide new patients through their treatments. However, these highly intelligent seekers also may create unexpected problems for agencies since they may create different paths and approaches to dealing with treating a disease or motivating clinical research studies.

Formal Approaches to Support Patients: Self-Help and Advocacy

For a number of years, formal groups have continued to serve as opinion leaders and information seekers for individuals or support their everyday health information needs. Self-help groups are estimated to be in the hundreds of thousands across a wide variety of diseases with members numbering in the millions [22]. They also can provide critical information on the personal side of disease: *How will my spouse react? Am I in danger of losing my job? Will I get proper treatment in a clinical study?* In addition, these groups also can prepare someone psychologically for a more active or directed search for information once his or her immediate personal reactions are dealt with or as more knowledge is gained on a particular disease, clinical trial options, and so on. Driving this movement has been the belief that self-help groups have the potential to affect outcomes by supporting patients' general well-being and sense of personal empowerment [22], and the diversity of tools now available have the potential to further enable this.

The WWW has increased the impact of these groups and the functionality and tools available to individuals, with the additional twist that formal institutions or private companies often support these groups. One prominent and relatively recent example of a robust and multifaceted online support system (or health social network) is PatientsLikeMe (PLM) (www.patientslikeme.com). PLM is essentially an online support group that uses patient-reported outcomes, symptoms, and various treatment data to help individuals find and communicate with others with similar

health issues [51]. Its developers have noted that the essential question asked by patients participating in one of the several disease communities is: “Given my current situation, what is the best outcome I can expect to achieve and how do I get there?” [52]. Personal health records, graphical profiles, and various communication and networking tools help patients in their quest to answer this. Enhanced access to others willing to share experiences is obviously critical and would certainly have been nearly impossible prior to the information and communication technologies available today.

Another prominent and long-lasting self-help intervention is the Comprehensive Health Enhancement Support System (CHESS) which has focused on a variety of diseases with educational and group components, closed membership, fixed duration, and decision support [53]. Computer-mediated support group (CMSG) interventions such as CHESS have been shown in a recent meta-analysis to increase social support, to decrease depression, and to increase quality of life and self-efficacy, with their effects moderated by group size, the type of communication channel, and the duration of the intervention [54].

Although motives will vary from one group to the next, commonalities across these include diverse approaches for social support, information exchange, and patient data tracking and also finding and connecting patients to clinical trials. A few examples of these other sites with varied tools for patients are shown below:

Site	Description	URL
ClinicalTrials.gov	Publically available clinical trial database	https://www.clinicaltrials.gov/
Center Watch	Clinical trial information for patients and professionals	http://www.centerwatch.com/
Pfizer Link	Online community for their clinical trial participants	https://www.pfizerlink.com/
Antidote (formerly TrialReach):	Digital health company which matches patients with pertinent research studies	https://www.antidote.me/
Fox Trial Finder	Aids in recruiting clinical trial participants for Parkinson’s studies	https://foxtrialfinder.michaeljfox.org/
ISRCTN registry	Accepts all proposed, ongoing, and completed clinical research studies and works to “support researchers in providing lay summaries and research feedback”	https://www.isrctn.com/

The emergence of *advocacy groups* over (at least) the last half century comes from people with the same disease or afflictions who need to share efforts in facing similar challenges, to exchange knowledge that is recognized as different from that of health professionals, and to speak with a more unified voice to impact policy and promote research [55]. Advocacy groups have interests beyond serving and supporting the needs of their individual members; however, they may seek to change societal reactions to their members or insure that sufficient resources are devoted to the needs of their groups [56]. At times, these groups will have agendas that do not

necessarily coincide with an individual's needs. Advocacy groups need members to advance the group's agendas. For example, they often are especially interested in insuring that the latest information on treatment is made available to patients, sometimes pressing for the release of information on experimental treatments before they would traditionally be available. A recent Department of Health and Human Services (HHS) rule requires drug development transparency through increased patient access to information on experimental therapies and expanded access to these treatments. The rule requires investigators to submit information on expanded access to experimental therapies to [ClinicalTrial.gov](#) and is administrated by the FDA and NIH [57]. Advocacy groups are particularly active when there are no clear treatment options or when they are perceived to be ineffective [56]. Thus, at times individual and group interests coincide, and at times, of course, they do not.

Patient Researcher

Previously we noted that *PatientsLikeMe* is an early example of a certain kind of robust, patient-driven research platform that we may see more of in the future. The site essentially furthers the notion of the *patient researcher*, serves many of the same positive elements of social interaction described above (e.g., emotional, social, and informational support), and shows promise for positively impacting outcomes [23]. Yet unlike social groups of the past, sites such as PLM are much more dynamic. Specifically, thousands of patients' data are aggregated, so individuals can compare their own diagnoses, treatments, symptoms, etc. with many others in order to help them choose a more personal path toward a better outcome. This path is lined with social and emotional support, quantified and visualized self-tracking, and opportunities for other treatments or research participation. Also of interest is how such a technology has seemingly accelerated the kind of networks and patient interactions that have benefited consumers and how this acceleration has the potential to strengthen or better enable patient empowerment.

As models in health are changing and are more reflective of the consumer health movement involving personal empowerment, social networking, and enabling technologies, there has been a concomitant emergence of new challenges in research that these have fostered. Patients, the advocacy or related groups they form or join, and even research enterprises are all helping to move into the "obvious next phase of active patient participation in health social networks," the area of patient-inspired or patient-run research [1]. The promise of new research models may be great, but as with any shift or change, new challenges and issues will arise. Much of the traditional medical literature has focused on very real concerns about poor health literacy and the growing gaps in knowledge or awareness of large segments of the public [58, 59], yet most of the threats to clinical research focus on hyperseekers who constitute only a small proportion of the public. Still, the consumer movement assumes increasingly sophisticated individuals who can understand issues ranging from advanced cell biology to psychosocial adjustment to pain management.

Patients now have incredible options to operate in an information field that is personalized, quantifiable, linked to others, and with even more choices for resources. The citizen researcher of even the recent past needed more than Internet access; they needed to analyze and integrate information from sources ranging from those specifically for laypersons (e.g., [healthfinder.gov](#)) to extraordinarily sophisticated information and tools (e.g., the array of tools and resources available at the National Center for Biotechnology Information—NCBI).

New technologies create an increasingly fragmented and privatized information environment, as opposed to the more mass, public access technologies represented by television and radio [60]. In response to these trends, governmental agencies are adopting policies to promote information equity among various segments of our society [61], but some question whether access to information resources can ever truly be universal, in spite of the best intentions of our policy makers [62].

Clinical research requires access to patient data, and PLM and related online consumer networks encourage patients to share their own data, ultimately for aggregated analysis, so it can be sold to or otherwise accessed by research companies and agencies of various sorts [66]. Importantly, since the data is provided directly by the patient, the hurdles normally associated with clinical research can be partially removed [63]. The obvious questions that are arising about this model relate to how PLM and such sites can balance their own profit motive with the altruistic one stated in their “openness philosophy” (<http://www.patientslikeme.com/about/openness>), which is one that seeks to accelerate and democratize research. Such sites considerably speed the dissemination of research results to those who can benefit from them [64]. The individual patient’s desire to become a partner in research, to learn, to share, and ultimately to identify a positive outcome for a certain disease is leveraged in this democratization process. Frost notes that this model of sharing is continually under review by PLM to understand how this level of participation impacts decision-making and actions [51]. In one small study, there are telling questions and responses highlighted that show many patients communicate with others in the community to seek treatment recommendations [65]. Much of the advice given seemed to come from personal research or firsthand experiences. Such information sharing can be quite compelling to individuals in dire need for some answer, in particular since the information exchanges occur among patients with similar data profiles and medical concerns. This is an area that has not been explored deeply at this time, but one that requires a host of approaches to better understand.

For instance, there is relatively little known in this context regarding the impact of visualization of the data has on comprehension. Visual representation of information, especially risk, can be interpreted differently and with varying psychosocial effects, many unintended. Some patients might react to any increased risk for a disease or any adverse side effect very negatively, which could preclude taking appropriate preventive measures or lead to depression or other negative reactions. Moreover, even if patients are similar in certain data-supported ways, the desire for a resolution to one’s needs and concerns could lead to overly optimistic hopes for untested treatments, such as complementary or alternative medicines.

Evolving Models to Engage Consumers in Clinical Research

Empowering patients with accessibility to and ownership of their own medical data reverses the predominantly one-way dynamic of today's health care system [66].

In the previous sections, we used broad strokes to lay a foundation from traditional communication and health information research that may be useful for framing an understanding of the evolving role of consumers. Furthermore, our premise is that a major goal of the consumer health movement is the fostering of patient or consumer empowerment. In part, this means a continuing shift from traditional models of medicine and clinical research to ones where patients have a greater role in their own decision-making, from treatment options to involvement in clinical research, to actually initiating and conducting research themselves. The core issues relate to more than choice in and of itself, but rather choice for achieving more personalized medicine, for increasing safety in research and care, and for accomplishing other altruistic aims that may be supported by social networks that enable knowledge transfer, greater voice, and concerted action evoking the wisdom of crowds.

In this section, we offer a discussion of newer or emerging models and enabling technologies that we believe will help in the movement toward greater emphasis on consumer empowerment, patient engagement, and evolving consumer/patient relationships with information and technology.

Patient Engagement Models in Clinical Research

Recently, national research goals have shifted to include initiatives which promote patient engagement in clinical research. The literature notes the disconnect between patient and investigator research priorities, which may contribute problems such as the ongoing challenge of low clinical trial recruitment. In the patient-engaged research model, the patients or patient communities are active participants in the design, process, and analysis of clinical trial research. In addition, patients may engage in the design, recruitment, data collection, and dissemination of clinical trial results [67, 68].

Along with the movement to engage patients in research, new national priorities involve improving patient outcomes through patient-centered care and research. Positive clinical outcomes must be seen as important to not only the clinical researchers but also the patient, such as quality of life indicators in addition to laboratory values [69]. In 2010, the Patient-Centered Outcomes Research Trust Fund (PCORTF) was established by the Patient Protection and Affordable Care Act of 2010. PCORTF provides federal funding for the Patient-Centered Outcomes Research Institute (PCORI) [70]. According to PCORI, their mission is to aid patients, families, and their caregivers to "make informed healthcare decisions[...] by producing and promoting high-integrity, evidence-based information that comes from research guided by patients, caregivers, and the broader healthcare

community” [71]. A primary initiative of PCORI has been the establishment of the National Patient-Centered Clinical Research Network (PCORnet) which is a distributed research network linking health information from over 130 patient groups and health systems (approximately 100 million patients across the United States) as of 2015 [72]. PCORnet aims to leverage patient health data, by partnering with stakeholders across the United States including patients and patient advocacy groups in addition to research and clinical stakeholders, to improve health research efficiency and access to difficult-to-reach patient populations such as those patients with rare diseases. This extensive health research information network also has the potential to improve clinical research participant diversity [73]. Furthermore, patients are empowered with significant influence on the selection of research projects [72].

PCORI has also developed an *Engagement Rubric* designed to help guide how input from patients and stakeholders can be built into the research process throughout the study [74]. This is similar to other efforts or findings in the literature that show the experience-based expertise unique to patients (as well as certain non-patients in underrepresented demographics) is important to patient-centered approaches [75–77]. Several such considerations in improving research studies include the need to improve patient engagement in the research process, recruitment, and retention in projects and to produce research that is more relevant and accessible to consumers [78].

Direct-to-Consumer Testing and Data Collection for Research

In contrast to primarily investigator research data collection in clinical trials, there are several business models emerging that empower patients to directly engage in clinical testing, data collection, and research. These models promote consumer engagement in research through patient consent for de-identified data sharing. One example is direct-to-consumer marketing of genetic testing by companies such as *23andMe* and *AncestryDNA* (<https://www.23andme.com/>; [Ancestry.com](https://www.ancestry.com)). This type of genetic testing may also be referred to as “at-home genetic testing” [79]. The consumer pays a fee for their genetic profile incentivized by the ability to learn more about their genetic makeup and possible links to health disorders. Consumers do not need to have a healthcare provider or insurance company involved in the ordering process; however, some companies may have a healthcare provider or counselor available to discuss the results. With patient consent, health data collected can also be used by researchers to study the genetics of health conditions on a population level. According to the American Society of Human Genetics (ASHG), the benefits of this model include increased consumer access and consumer empowerment. Furthermore, increasing consumer access to genetic testings may result in more genetic data availability for health researchers and increased consumer awareness of genetic disorders [80]. The risks of this direct-to-consumer model include consumer misinterpretation of results, lack of access to genetic counseling, accuracy/validity of laboratory results, and privacy concerns [79, 80].

Crowdsourcing

In addition to the aforementioned PLM, there are other, large-scale initiatives that are built upon consumer engagement in research from both the private and government sectors. Specifically, the *All of Us Research Program*, which is part of the National Institutes of Health (NIH) Precision Medicine Initiative, is soliciting patient health data contributions from one million participants to support clinical research to improve the ability to deliver more personalized healthcare (*All of Us Research Program*, <https://www.joinallofus.org/en/program-overview>) [81]. The initiative is unique in that it will collect lifestyle and environment measurements in addition to biological and genetic data, which are often not entirely captured in traditional clinical datasets such as electronic health records [81]. The NIH also plans to leverage mobile health technology (see mobile health section below) to aid in the collection of activity and environmental data element measurements to study activity and environmental exposure effects on personal health [81]. However, with this new role of the patient as a partner in research comes the ethical/moral obligation of the patient to provide accurate and quality data.

The efforts of institutes such as PCORI and the NIH (*All of Us*) to engage health consumers in clinical research, especially through health data contributions, may improve clinical investigator access to health data, especially data on rare diseases and data not easily accessible from electronic health records, to support precision medicine research efforts. Precision medicine is defined by the Precision Medicine Initiative as, “an emerging approach for disease treatment and prevention that takes into account individual variability in genes, environment, and lifestyle for each person” [82, 83]. Crowdsourcing efforts, such as the one million patient data collection goal by the NIH All of Us Project, are required to collect the necessary diverse dataset to support this type of research.

Mobile Health (mHealth)

Mobile health technology can be used by clinical researchers to collect data points, such as lifestyle and environmental data, to analyze clinical trial outcomes. With the majority of the patient population having access to a mobile device such as a smartphone, mobile health research methods can facilitate research data collection. Such methods include the increasing use of wearable and sensor technology to collect health data [84, 85].

In addition, mobile health devices can also be used to record physiological measures such as the use of remote devices to measure pulse oximetry, blood glucose, heart rate, and blood pressure [86]. As noted, in this model of data collection, the consumer has a role in assuring the quality and quantity of data provided to the clinical researcher. However, with improving technology, some devices can transfer data automatically via wireless mobile connections which reduces “the burden” on consumers to participate in these types of trials.

Clinical Trial Involvement

Another attractive aspect of health-related social networks and crowdsourcing is the potential to overcome the discouraging barriers to patient recruitment into clinical trials [63] and other research projects. Projects like *All of Us* or *Army of Women* (armyofwomen.org) can greatly facilitate researcher access to willing populations for those who go through their elaborate approval process. There are a number of affective and practical reasons individuals do not, or cannot, be part of a clinical trial. Certainly, in traditional clinical research, access to the study site is an issue that is not easily overcome by many, particularly those in rural, underserved areas. Moreover, many patients understandably question how involvement in a study might impact his or her quality of life, even if they have strong feelings of altruism. Human nature suggests there might also be concerns of bias by physicians seeking to enroll patients into a trial, and knowing which trials are available has been a challenge even with such national efforts as ClinicalTrials.gov [17].

Social networking platforms present the potential for studying existing data as well as for mining these sites for likely study populations based on eligibility criteria or other factors [63]. While the nature of the participants in many of these sites may be that they are already willing partners seeking to find a path to a positive outcome for themselves and others like them, with reportedly 1/3 of traditional trial recruitment sites failing to recruit a single patient [17], online patient communities offer a far more promising outlook. For instance, Facebook advertising of trials has been shown to improve recruitment [87, 88], including hard to reach populations [89]. Critical to this potential revolution, however, is an understanding that such communities are not merely a gathering ground for X number of people with disease Y looking for a cure. Rather, these are increasingly savvy consumers who have empowered themselves with personal and collective knowledge and expertise, who are not likely to respond to every call for participants and who have been known to share information on ongoing trials in ways that can be very disruptive of traditional research. In other words, a shift in the research model will certainly need to be advanced but only with the consent of a more influential group. Potential collaborations among site developers, researchers, and patients could expedite research and advance the needs of all groups, for instance, through the use of patient registries on such sites (please see Chap. 14).

There are a number of factors that impact accrual to trials, retention, and satisfaction. Many of these can be enhanced or supported via consumer informatics tools or well-designed information systems and research studies. For instance, it is important to note that many individual patients place a high value on the information they receive during research trials and cite it as a key reason for choosing to participate [90–92]. Thus, bearing in mind health literacy issues, consumer involvement in research demands study designs that can be clearly, simply explained to participants without using jargon, with similarly accessible information updates throughout the research process [76, 93, 94]. Consumers also report that communication from researchers may be frequent at the start of studies but drops off dramatically, with many patients never even knowing the final study outcomes. Patients repeatedly cite the lack of follow-up information (including results) from researchers [96–98],

confirming related findings that patient engagement occurs a lot in the beginning and very little as the study progresses. Lastly, while altruism has been found to be a major motivator for participation in clinical research, people want to see *how* they might have made a difference [92, 96]. This can be achieved with even simple steps using email, texts, call-ins, annual “thank you” breakfasts, or social media updates, in order to help keep study participants in the loop, so to speak [95, 96].

Consumers Relationship with Their Own Information

The evolving role of consumers has also meant a more dynamic relationship with their own health data. As we have seen, national efforts toward more consumer engagement in health data collection, use, and management are seen in PCORI, *All of Us*, and other innovative models. Still, there is a noticeable lack of a complete health information network in the United States which may be driving the need for patients to manage their own data, e.g., Blue Button from the VA and Apple’s new health data application [97, 98]. Blue Button emerged as an initiative to address lack of Veteran access to their own medical records [98].

Additionally, empowered consumers can leverage information technologies to improve access to clinical trial results. [ClinicalTrials.gov](#) serves not only to increase patient awareness of available trials, but it also fulfills investigator obligations to share clinical results with participants, researchers, and communities. With the push for more health consumer access and control of their data nationally, new technology will emerge to fill this need. Therefore, it would be expected that large private technology companies would be interested in expanding their business models to include growth in the health informatics area [99]. Recently, Apple (e.g., Apple’s research kit) and Amazon have announced their entry into the field. This has happened in the past with attempts by both Google and Microsoft; however, these efforts were short lived due to lack of user adoption [100]. Ultimately, consumers will decide whether or not to change their current methods of using and storing their medical data. The factors involved in this decision include privacy, trust, cost, and willingness to share this information.

Conclusions

To support patient empowerment, even in the broadest sense, now means understanding the interactions among patients or consumers themselves and between consumers and the fragmented and increasingly complex health information environment they must navigate. We have long known that information alone, whether provided by an intermediary or accessed directly, does not necessarily lead to rational choice or informed decision-making [101]. For instance, the traditional “one size fits all” approach to public health campaigns is limited at best. Research in information behaviors continues to reveal that individuals facing serious health issues will seek out others with similar problems and that the notion of opinion

leaders is evolving in the new social networking environments emerging online. These patient-empowered research networks allow participants in clinical trials to “... unblind themselves, pool their data, parse literature, conduct statistical analysis, and post their findings online” [102]. New technologies are enabling a personalization of medicine that facilitates more quantitative assessment of one’s own progress toward some possible positive outcome and of one’s state measured against others. While there are concerns over an increasing influence of the private sector, direct-to-consumer marketing, and related social and ethical considerations, there is plenty of promising evidence suggesting a new model of clinical research is now possible: one that will help speed discovery and encourage participation. Patients are savvier and can make better decisions as to which trials might be a good fit for them; consequently, adverse events could be identified more quickly, thus helping to make clinical trials safer.

The underlying issues are not resolved but are becoming clearer, and this clarity will help guide future research. Information fields are becoming even more fluid as choices of sources and changing technologies become available and more ubiquitous. Collaboration among patients means enhanced knowledge sharing, and the citizen researcher can leverage this to help drive research relying on the wisdom of crowds to quickly correct erroneous information [64, 66]. These issues are part of the evolving role of consumers and the technologies and systems that support them and have risen to be particularly salient in the context of clinical research efforts.

References

1. Swan M. Emerging patient-driven health care models: an examination of health social networks, consumer personalized medicine and quantified self-tracking. *Int J Environ Res Public Health*. 2009;6:492–525. <https://doi.org/10.3390/ijerph6020492>.
2. Wallerstein N. What is the evidence on effectiveness of empowerment to improve health? World Health Organization Regional Office for Europe; 2006. <http://www.euro.who.int/en/what-we-do/data-and-evidence/health-evidence-network-hen/publications/pre2009/what-is-the-evidence-on-effectiveness-of-empowerment-to-improve-health>. Accessed Aug 2011.
3. Lemire M, Sicotte C, Paré G. Internet use and the logics of personal empowerment in health. *Health Policy*. 2008;88:130–40. <https://doi.org/10.1016/j.healthpol.2008.03.006>.
4. Johnson JD. Cancer-related information seeking. Cresskill: Hampton Press; 1997.
5. Rice RE, Atkin CK. Preface: trends in communication campaign research. In: Rice RE, Atkin CK, editors. *Public communication campaigns*. Newbury Park: Sage; 1989. p. 7–11.
6. Atkin C, Walleck L, editors. *Mass communication and public health*. Newbury Park: Sage; 1990.
7. Johnson JD, Andrews JE, Case DO, Allard SL, Johnson NE. Fields and/or pathways: contrasting and/or complementary views of information seeking. *Inf Process Manag*. 2006;42:569–82. <https://doi.org/10.1016/j.ipm.2004.12.001>.
8. Noar SM. A 10-year retrospective of research in health mass media campaigns: where do we go from here? *J Health Commun*. 2006;11:21–42. <https://doi.org/10.1080/10810730500461059>.
9. Hornik RC. Epilogue: evaluation design for public health communication programs. In: Hornik RC, editor. *Public health communication: evidence for behavior change*. Mahwah: Lawrence Erlbaum Associates; 2002. p. 385–405.
10. Noar SM. Challenges in evaluating health communication campaigns: defining the issues. *Commun Methods Meas*. 2009;3:1–11. <https://doi.org/10.1080/19312450902809367>.

11. Freimuth VS. Improve the cancer knowledge gap between whites and African Americans. *J Natl Cancer Inst.* 1993;14:81–92.
12. Freimuth VS, Stein JA, Kean TJ. Searching for health information: the cancer information service model. Philadelphia: University of Pennsylvania Press; 1989.
13. Alcalay R. The impact of mass communication campaigns in the health field. *Soc Sci Med.* 1983;17:87–94. [https://doi.org/10.1016/0277-9536\(83\)90359-3](https://doi.org/10.1016/0277-9536(83)90359-3).
14. Katz E, Lazarsfeld PF. Personal influence: the part played by people in the flow of mass communications. New York: Free Press; 1955.
15. Lichten I. Communication in cancer care. New York: Churchill Livingstone; 1987.
16. Rogers EM, Storey JD. Communication campaigns. In: Berger CR, Chaffee SH, editors. *Handbook of communication science*. Newbury Park: Sage; 1987. p. 817–46.
17. Allison M. Can web 2.0 reboot clinical trials? *Nat Biotechnol.* 2009;27:895–902. <https://doi.org/10.1038/nbt1009-895>.
18. Mills EJ, Seely D, Rachlis B, et al. Barriers to participation in clinical trials of cancer: a meta-analysis and systematic review of patient-reported factors. *Lancet Oncol.* 2006;7(2):141–8.
19. Atkinson NL, Massett HA, Mylks C, Hanna B, Deering MJ, Hesse BW. User-centered research on breast cancer patient needs and preferences of an internet-based clinical trial matching system. *J Med Internet Res.* 2007;9:e13. <https://doi.org/10.2196/jmir.9.2.e13>.
20. Marks L, Power E. Using technology to address recruitment issues in the clinical trial process. *Trends Biotechnol.* 2002;20:105–9. [https://doi.org/10.1016/S0167-7799\(02\)01881-4](https://doi.org/10.1016/S0167-7799(02)01881-4).
21. Fox S, Jones S. The social life of health information: Americans' pursuit of health takes place within a widening network of both online and offline sources. Pew Internet & American Life Project; 2009. <http://www.pewinternet.org/Reports/2009/8-The-Social-Life-of-Health-Information.aspx>. Accessed Aug 2011.
22. Barak A, Boniel-Nissim M, Suler J. Fostering empowerment in online support groups. *Comput Hum Behav.* 2008;24:1867–83. <https://doi.org/10.1016/j.chb.2008.02.004>.
23. Wicks P, Massagli M, Frost J, Brownstein C, Okun S, Vaughan T, Bradley R, Heywood J. Sharing health data for better outcomes on PatientsLikeMe. *J Med Internet Res.* 2010;12:e19. <https://doi.org/10.2196/jmir.1549>.
24. Cool C. The concept of situation in information science. *Annu Rev Inf Sci Technol.* 2001;35:5–42.
25. Johnson JD. Information seeking: an organizational dilemma. Westport: Quorum Books; 1996.
26. Rice RE, McCreadie M, Chang SL. Accessing and browsing information and communication. Cambridge, MA: MIT Press; 2001.
27. Sonnenwald DH, Wildemuth BM, Harmon GL. A research method to investigate information seeking using the concept of information horizons: an example from a study of lower socio-economic students' information seeking behavior. *New Rev Inf Behav Res.* 2001;2:65–85.
28. Scott J. Social network analysis: a handbook. 2nd ed. Thousand Oaks: Sage; 2000.
29. Kuhlthau CC. Inside the search process: information seeking from the user's perspective. *J Am Soc Inf Sci Technol.* 1991;42:361–71. [https://doi.org/10.1002/\(SICI\)1097-4571\(199106\)42:5<361::AID-ASI6>3.0.CO;2-#](https://doi.org/10.1002/(SICI)1097-4571(199106)42:5<361::AID-ASI6>3.0.CO;2-#).
30. Williamson K. Discovered by chance: the role of incidental information acquisition in an ecological model of information use. *Libr Inf Sci Res.* 1998;20:23–40. [https://doi.org/10.1016/S0740-8188\(98\)90004-4](https://doi.org/10.1016/S0740-8188(98)90004-4).
31. Fisher KE, Durrance JC, Hinton MB. Information grounds and the use of need-based services by immigrants in Queens, New York: a context-based, outcome evaluation approach. *J Am Soc Inf Sci Technol.* 2004;55:754–66. <https://doi.org/10.1002/asi.20019>.
32. Clifton A, Turkheimer E, Oltmanns TF. Personality disorder in social networks: network position as a marker of interpersonal dysfunction. *Soc Netw.* 2009;31:26–32. <https://doi.org/10.1016/j.socnet.2008.08.003>.
33. Cornwell B. Good health and the bridging of structural holes. *Soc Netw.* 2009;31:92–103. <https://doi.org/10.1016/j.socnet.2008.10.005>.

34. Adelman MB, Parks MR, Albrecht TL. Beyond close relationships: support in weak ties. In: Albrecht TL, Adelman MB, editors. *Communicating social support*. Newbury Park: Sage; 1987. p. 126–47.
35. Albrecht TL, Adelman MB. Communication networks as structures of social support. In: Albrecht TL, Adelman MB, editors. *Communicating social support*. Newbury Park: Sage; 1987. p. 40–63.
36. Becker MH, Rosenstock IN. Compliance with medical advice. In: Steptoe A, Mathews A, editors. *Health care and human behavior*. London: Academic; 1984. p. 175–208.
37. Rogers EM, Kincaid DL. *Communication networks: toward a new paradigm for research*. New York: Free Press; 1981.
38. Johnson JD. *Managing knowledge networks*. Cambridge, UK: Cambridge University Press; 2009.
39. Granovetter MS. The strength of weak ties. *AJS*. 1973;78:1360–80.
40. Fox S, Raine L. How internet users decide what information to trust when they or their loved ones are sick. Pew Internet & American Life Project; 2002. <http://www.pewinternet.org/Reports/2002/Vital-Decisions-A-Pew-Internet-Health-Report/Summary-of-Findings.aspx>. Accessed Aug 2011.
41. Taylor H, Leitman R. Four-nation survey shows widespread but different levels of Internet use for health purposes. Harris Interactive Healthcare Care News. 2002. http://www.harrisinteractive.com/news/newsletters/healthnews/HI_HealthCareNews2002Vol2_iss11.pdf. Accessed Aug 2011.
42. Lowery W, Anderson WB. The impact of web use on the public perception of physicians. Paper presented to the annual convention of the Association for Education in Journalism and Mass Communication. Miami Beach; 2002.
43. Johnson JD. Dosage: a bridging metaphor for theory and practice. *Int J Strateg Commun*. 2008;2:137–53. <https://doi.org/10.1080/15531180801958204>.
44. Parrott R, Steiner C. Lessons learned about academic and public health collaborations in the conduct of community-based research. In: Thompson TL, Dorsey AM, Miller K, Parrott RL, editors. *Handbook of health communication*. Mahwah: Lawrence Erlbaum Associates; 2003. p. 637–50.
45. Case D, Johnson JD, Andrews JE, Allard S, Kelly KM. From two-step flow to the internet: the changing array of sources for genetics information seeking. *J Am Soc Inf Sci Technol*. 2004;55:660–9. <https://doi.org/10.1002/asi.20000>.
46. Shapiro AL. The control revolution.....: how the internet is putting individuals in charge and changing the world we know. New York: Public Affairs; 1999.
47. Metoyer-Duran C. Information gatekeepers. *Annu Rev Inf Sci Technol*. 1993;28:111–50.
48. Burt RS. *Structural holes: the social structure of competition*. Cambridge, MA: Harvard University Press; 1992.
49. Katz E. The two step flow of communication: an up to date report on an hypothesis. *Public Opin Q*. 1957;21:61–78.
50. Paisley WJ. Knowledge utilization: the role of new communications technologies. *J Am Soc Inf Sci*. 1993;44:222–34.
51. Frost J, Massagli M. PatientsLikeMe the case for a data-centered patient community and how ALS patients use the community to inform treatment decisions and manage pulmonary health. *Chron Respir Dis*. 2009;6:225–9. <https://doi.org/10.1177/1479972309348655>.
52. Brownstein CA, Brownstein JS, Williams DS III, Wicks P, Heywood JA. The power of social networking in medicine. *Nat Biotechnol*. 2009;27:888–90. <https://doi.org/10.1038/nbt1009-888>.
53. Gustafson DH, Hawkins R, McTavish F, Pingree S, Chen WC, Volrathongchai K, Stengle W, Stewart JA, Serlin RC. Internet-based interactive support for cancer patients: are integrated systems better? *J Commun*. 2008;58:238–57. <https://doi.org/10.1111/j.1460-2466.2008.00383.x>.
54. Rains SA, Young V. A meta-analysis of research on formal computer-mediated support groups: examining group characteristics and health outcomes. *Hum Commun Res*. 2009;35: 309–36.

55. Aymé S, Kole A, Groft S. Empowerment of patients: lessons from the rare diseases community. *Lancet*. 2008;371(9629):2048–51.
56. Weijer C. Our bodies, our science: challenging the breast cancer establishment, victims now ask for a voice in the war against disease. *Sciences*. 1995;35:41–4.
57. Statement of Scott Gottlieb, M.D., Commissioner of Food and Drugs before the Subcommittee on Health, Committee on Energy and Commerce, US House of Representatives. 2017. <https://www.fda.gov/newsevents/testimony/ucm578634.htm>. Accessed 29 June 2018.
58. The Joint Commission J. ‘What did the doctor say?’: improving health literacy to protect patient safety. Oakbrook: The Joint Commission. 2007. http://www.jointcommission.org/What_Did_the_Doctor_Say/. Accessed Aug 2011.
59. McCray AT. Promoting health literacy. *J AHIMA*. 2005;12:152–63. <https://doi.org/10.1197/jamia.M1687>.
60. Siebert M, Gerbner G, Fisher J. The information gap: how computers and other new communication technologies affect the social distribution of power. New York: Oxford University Press; 1989.
61. Doctor RD. Social equity and information technologies: moving toward information democracy. In: Williams ME, editor. Annual review of information science and technology. Medford: Learned Information; 1992. p. 44–96.
62. Fortner RS. Excommunication in the information society. *Crit Stud Mass Commun*. 1995;12:133–54. <https://doi.org/10.1080/15295039509366928>.
63. Brubaker JR, Lustig C, Hayes GR. PatientsLikeMe: empowerment and representation in a patient-centered social network. Presented at the CSCW 2010 workshop on CSCW research in healthcare: past, present, and future. Savannah; 2007.
64. Ferguson T. e-patients: how they can help us heal health care. *e-patients.net*. 2007. <http://e-patients.net>. Accessed Aug 2011.
65. Frost J, Massagli M. Social uses of personal health information within PatientsLikeMe, and online patient community: what can happen with patients have access to one another’s data. *J Med Int Res*. 2008;10(3):e15.
66. Steinhubl SR, Muse ED, Topol EJ. The emerging field of mobile health. *Sci Transl Med*. 2015;7(283):283rv3. <https://doi.org/10.1126/scitranslmed.aaa3487>.
67. Sacristán JA, Aguarón A, Avendaño-Solá C, et al. Patient involvement in clinical research: why, when, and how. *Patient Prefer Adherence*. 2016;10:631–40. <https://doi.org/10.2147/PPA.S104259>.
68. Frank L, Forsythe L, Ellis L, et al. Conceptual and practical foundations of patient engagement in research at the patient-centered outcomes research institute. *Qual Life Res*. 2015;24(5):1033–41. <https://doi.org/10.1007/s11136-014-0893-3>.
69. Epstein RM, Street RL. The values and value of patient-centered care. *Ann Fam Med*. 2011;9(2):100–3. <https://doi.org/10.1370/afm.1239>.
70. DHHS, Patient Outcomes Research Trust Fund. <https://aspe.hhs.gov/patient-centered-outcomes-research-trust-fund> (2018). Last accessed June 29, 2018.
71. PCORI. About us, Our Mission. <https://www.pcori.org/about-us>. 2018. Accessed 29 June 2018.
72. PCORI. Fact sheet. <https://www.pcori.org/sites/default/files/PCORI-PCORnet-Fact-Sheet.pdf>. 2018. Accessed 29 June 2018.
73. PCORI. <https://www.pcori.org/research-results/pcornet-national-patient-centered-clinical-research-network>. 2018. Accessed 29 June 2018.
74. PCORI Engagement Rubric (Patient-Centered Outcomes Research Institute) website. <https://www.pcori.org/sites/default/files/Engagement-Rubric.pdf>. Published February 4, 2014. Updated June 6, 2016. Accessed 29 June 2018.
75. Crocker JC, Boylan AM, Bostock J, Locock L. Is it worth it? Patient and public views on the impact of their involvement in health research and its assessment: a UK-based qualitative interview study. *Health Expect*. 2017;20(3):519–28.
76. Demian MN, Lam NN, Mac-Way F, Sapir-Pichhadze R, Fernandez N. Opportunities for engaging patients in kidney research. *Can J Kidney Health Dis*. 2017;4:2054358117703070.

77. Dudley L, Gamble C, Allam A, Bell P, Buck D, Goodare H, Hanley B, Preston J, Walker A, Williamson P, Young B. A little more conversation please? Qualitative study of researchers' and patients' interview accounts of training for patient and public involvement in clinical trials. *Trials*. 2015;16(1):190.
78. Domecq JP, Prutsky G, Elraiayah T, Wang Z, Nabhan M, Shippee N, Brito JP, Boehmer K, Hasan R, Firwana B, Erwin P. Patient engagement in research: a systematic review. *BMC Health Serv Res*. 2014;14(1):89.
79. National Library of Medicine. What is direct-to-consumer genetic testing? 2018. <https://ghr.nlm.nih.gov/primer/testing/directtococonsumer>. Accessed 29 June 2018.
80. Society News. ASHG statement on direct-to-consumer genetic testing in the United States. *Am J Hum Genet*. 2007;81:637. http://www.ashg.org/pdf/dtc_statement.pdf. Accessed 29 June 2018.
81. All of Us Research Program. <https://www.joinallofus.org/en/program-overview>.
82. White House Archives. <https://obamawhitehouse.archives.gov/node/333101>. Accessed 29 June 2018.
83. National Library of Medicine. <https://ghr.nlm.nih.gov/primer/precisionmedicine/definition>. 2018. Accessed 29 June 2018.
84. Wenzel. 2017. Accessed at <http://www.clinicalinformaticsnews.com/2017/04/26/wearables-shaping-the-future-of-clinical-trials.aspx>.
85. Pew Research Center. <http://www.pewinternet.org/fact-sheet/mobile/> (February 5, 2018). Accessed 29 June 2018.
86. Li X, Dunn J, Salins D, et al. Digital health: tracking physiomes and activity using wearable biosensors reveals useful health-related information. In: Kirkwood T, editor. *PLoS Biol*. 2017;15(1):e2001402. <https://doi.org/10.1371/journal.pbio.2001402>.
87. Nash EL, Gilroy D, Srikanthanukul W, Abhayaratna WP, Stanton T, Mitchell G, Stowasser M, Sharman JE. Facebook advertising for participant recruitment into a blood pressure clinical trial. *J Hypertens*. 2017;35(12):2527–31.
88. Carter-Harris L, Bartlett Ellis R, Warrick A, Rawl S. Beyond traditional newspaper advertisement: leveraging facebook-targeted advertisement to recruit long-term smokers for research. *J Med Internet Res*. 2016;18(6):e117. <https://doi.org/10.2196/jmir.5502>.
89. Kayrouz R, Dear BF, Karin E, Titov N. Facebook as an effective recruitment strategy for mental health research of hard to reach populations. *Internet Interv*. 2016;4:1–0.
90. Moorcraft SY, Marriott C, Peckitt C, Cunningham D, Chau I, Starling N, Watkins D, Rao S. Patients' willingness to participate in clinical trials and their views on aspects of cancer research: results of a prospective patient survey. *Trials*. 2016;17(1):17.
91. Ryan A. Engaging consumers with musculoskeletal conditions in health research: a user-centred perspective. In: Integrating and connecting care: selected papers from the 25th Australian National Health Informatics Conference (HIC 2017), vol. 239. IOS Press; 2017. p. 104
92. Zanni MV, Fitch K, Rivard C, Sanchez L, Douglas PS, Grinspoon S, Smeaton L, Currier JS, Looby SE. Follow YOUR heart: development of an evidence-based campaign empowering older women with HIV to participate in a large-scale cardiovascular disease prevention trial. *HIV Clin Trials*. 2017;18(2):83–91.
93. Boote J, Baird W, Beecroft C. Public involvement at the design stage of primary health research: a narrative review of case examples. *Health Policy*. 2010;95(1):10–23.
94. Collins K, Boote J, Ardron D, Gath J, Green T, Ahmedzai SH. Making patient and public involvement in cancer and palliative research a reality: academic support is vital for success. *BMJ Support Palliat Care*. 2014;5(2):203–6. <https://doi.org/10.1136/bmjspcare-2014-000750>.
95. Chakradhar S. Many returns: call-ins and breakfasts hand back results to study volunteers. *Nat Med*. 2015;21:304–6. pmid:25849267.
96. Buckley JM, Irving AD, Goodacre S. How do patients feel about taking part in clinical trials in emergency care? *Emerg Med J*. 2016;33(6):376–80. <https://doi.org/10.1136/emermed-2015-205146>.
97. Healthit.gov. <https://www.apple.com/ios/health/>. Accessed 29 June 2018.

98. U.S. Department of Veterans Affairs. Blue button. Accessed at <https://www.va.gov/bluebutton/>
99. Monegain B. Amazon, Apple only part of ‘seismic change’ coming to healthcare. Healthcare IT News; 2018. <http://www.healthcareitnews.com/news/amazon-apple-only-part-seismic-change-coming-healthcare>. Accessed 29 June 2018.
100. The Economist. <https://www.economist.com/news/business/21736193-worlds-biggest-tech-firms-see-opportunity-health-care-which-could-mean-empowered>. February 3, 2018. Accessed 29 June 2018.
101. Johnson JD. Health-related information seeking: is it worth it? Inf Process Manag. 2014;50(5):708–17.
102. Wicks P, Vaughan T, Heywood J. Subjects no more: what happens when trial participants realize they hold the power? BMJ. 2014;348:g368. <https://doi.org/10.1136/bmj.g368>.



Clinical Research in the Postgenomic Era

8

Stephane M. Meystre and Ramkiran Gouripeddi

Abstract

Clinical research, being patient-oriented, is based predominantly on clinical data – symptoms reported by patients, observations of patients made by health-care providers, radiological images, and various metrics, including laboratory measurements that reflect physiological functions. Recently, however, a new type of data – genes and their products – has entered the picture, and the expectation is that given clinical conditions can ultimately be linked to the function of specific genes. The postgenomic era is characterized by the availability of the human genome as well as the complete genomes of numerous reference organisms. How genomic information feeds into clinical research is the topic of this chapter. We first review the molecules that form the “blueprint of life” and discuss the surrounding research methodologies. Then we discuss how genetic data are clinically integrated. Finally, we relate how this new type of data is used in different clinical research domains.

Keywords

Postgenomic era · Genetic data · Molecular biology genomic data · Bioinformatics Sequence ontology · Bioinformatics Sequence Markup Language · Sequence analysis data · Structure analysis data · Functional analysis data

S. M. Meystre, MD, PhD, FACMI (✉)

Biomedical Informatics Center, Medical University of South Carolina, Charleston, SC, USA
e-mail: meystre@musc.edu

R. Gouripeddi, MS, MBBS

Department of Biomedical Informatics, University of Utah, Salt Lake City, UT, USA
e-mail: ram.gouripeddi@utah.edu

Clinical research, being patient-oriented, is based predominantly on clinical data – symptoms reported by patients, observations of patients made by health-care providers, radiological images, and various metrics – including laboratory measurements that reflect physiological functions. Recently, however, a new type of data – genes and their products – has entered the picture, and the expectation is that given clinical conditions can ultimately be linked to the function of specific genes.

This new approach is a fruit of the pregenomic era. That era, which lasted from 1990 to 2003, was defined by the Human Genome Project effort to sequence the nucleotides that make up the human genome and identify its approximately 25,000 genes [1]. Since all humans have a unique nucleotide sequence, the data produced by this project represents not the genome of a single individual, but the aggregate genome of a small number of anonymous donors.

Completion of the effort ushered in the postgenomic era, characterized by the availability of the human genome as well as the complete genomes of numerous reference organisms. How genomic information feeds into clinical research is the topic of this chapter. We first review the molecules that form the “blueprint of life” and discuss the surrounding research methodologies. Then we discuss how genetic data are clinically integrated. Finally, we relate how this new type of data is used in different clinical research domains.

The Molecular Basis of Life

As first enunciated by Crick in 1958 [2], deoxyribonucleic acid (DNA) is responsible for transmitting structural information to proteins, the key structural and functional components of living cells. The DNA sequence information is transmitted to daughter cell DNA by replication and to proteins in a two-step process: transcription to messenger RNA (mRNA) and then translation (Fig. 8.1). The full DNA sequence of an organism is the genome, and the set of all mRNA molecules produced in a cell is called the transcriptome. The totality of proteins expressed by the genome is the proteome, and the network of their interactions is called the

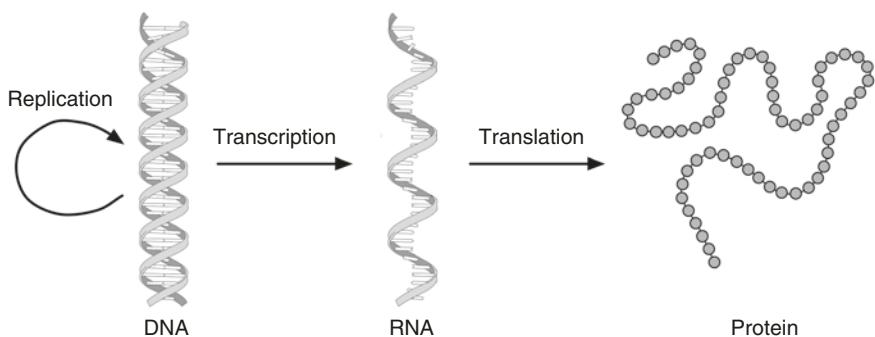


Fig. 8.1 Central dogma of molecular biology

interactome. The epigenome includes all heritable genome modifications that alter its expression. Finally, by-products and end products of metabolic pathways – metabolites – constitute the metabolome. For more information on these molecules of life, a resource such as the Genetics Home Reference [3] can be consulted.

The *omes* mentioned are the subjects of several fields of study. Genomics focuses on the genome and increasingly on comparative genomics (genetics focuses primarily on genes and their mutations and regulation). Transcriptomics focuses on the transcriptome, and proteomics on the proteome and proteins. Functional genomics focuses on the dynamic aspects of cell function – such as the timing and quantity of transcription, translation, and protein interactions – and therefore includes most of transcriptomics and proteomics. Metabolomics focuses on the metabolome, on how proteins interact with one another and with small molecules to transmit intra- and intercellular signals. Epigenomics centers on all epigenetic modifications of the genome. Microbiomics focuses on the microbiota of the human intestine, skin, and other body locations. Exposomics centers on the exposome, on the totality of human environmental exposures from conception onwards. Recent definitions of the exposome include endogenous processes within the body, biological responses of adaptation to the environment, and socio-behavioral factors beyond assessment of exposures [4].

Molecular Biology and Genomics Data

Molecular biology produces vast amounts of data. Currently, more than 1000 public molecular biology databases are available. Prominent examples and their Web addresses are listed in Table 8.1.

The flood of data (one RNA analysis, e.g., can produce an uncompressed image of more than 2000 MB) requires specialized tools for capture, visualization, and analysis. Computational tools and database development, and their application to the generation of biological knowledge, are the primary subdomains of bioinformatics. Bioinformatics, a term coined in 1978, is a discipline in which biology, computer science, and information technology merge. Bioinformatics uses computers for storage, retrieval, manipulation, and distribution of information related to biological macromolecules [5]. Bioinformatics tools are used extensively in three areas of molecular biological research – sequence analysis, structural analysis, and functional analysis.

Sequence Analysis Data

Knowledge of DNA, RNA, and gene and protein sequences is now indispensable in most biomedical research domains. In the clinical domain, this knowledge is used for studying disease mechanisms, for diagnosing and evaluating disease risk, and for treatment planning. Sequence analysis typically consists in searching for sequences of interest in specialized databases such as GenBank [6] or in identifying

Table 8.1 Selection of molecular biology databases

<i>Nucleotide sequence databases</i>	
GenBank	www.ncbi.nlm.nih.gov/Genbank
EMBL-Bank (European Molecular Biology Laboratory)	https://www.ebi.ac.uk
DDBJ (DNA Data Bank of Japan)	www.ddbj.nig.ac.jp
Relate DNA sequences with their location on chromosomes and corresponding genes, their products, official names and synonyms, and scientific publications. Used, for example, to identify the products (e.g., proteins) of a DNA sequence and develop methods to measure these products and therefore the activity of this sequence	
Many journals require submission of sequence information prior to publication, stimulating the growth of these databases	
<i>Amino acid sequence and proteomics databases</i>	
UniProt (Universal Protein Resource)	www.uniprot.org
PDB (Protein Data Bank)	www.rcsb.org/pdb
PRIDE (PRoteomics IDEntifications database)	www.ebi.ac.uk/pride
Ensembl	www.ensembl.org
InterPro	www.ebi.ac.uk/interpro
Relate proteins with their gene(s), function(s), structure, tissue specificity, involvement in diseases, official name and synonyms, variants, and scientific publications. Used, for example, to develop a new drug targeting a cell receptor with a known structure and predicting the structure the new drug should have.	
<i>Genes databases</i>	
OMIM (Online Mendelian Inheritance in Man)	www.ncbi.nlm.nih.gov/omim
Entrez Gene	www.ncbi.nlm.nih.gov/gene
Relate genes with their location, structure, function, interactions, associated phenotypes and diseases, markers, official name and synonyms, and scientific publications. Used, for example, to find all oncogenes related to a specific cancer and the markers that exist to detect them, to eventually develop a laboratory test predicting the behavior and outcome of this cancer	
<i>Gene and protein functional databases</i>	
OPHID (Online Predicted Human Interaction Database)	ophid.utoronto.ca
GEO (Gene Expression Omnibus)	www.ncbi.nlm.nih.gov/geo/
ArrayExpress	www.ebi.ac.uk/microarray-as/ae/
HMDB (Human Metabolome Database)	www.hmdb.ca
Relate genes and proteins with their expression profiles and corresponding scientific publications, official name and synonyms, diseases, and interactions. Used, for example, to link the expression profile of a set of genes in a patient with organ transplant with a graft rejection risk and subsequently adapt the treatment to prevent a rejection, therefore enabling “personalized medicine”	

(continued)

Table 8.1 (continued)

<i>Databases combining different types of molecular biology data</i>	
Entrez cross-database search	www.ncbi.nlm.nih.gov/sites/gquery
HGPD (Human Gene and Protein Database)	hupex.hgpd.jp/hgpd/cgi/index.cgi
KEGG (Kyoto Encyclopedia of Genes and Genomes)	www.genome.jp/kegg
GWAS Catalog	www.ebi.ac.uk/gwas/
Relate the genome with biological systems and the environment, integrate genes, proteins, and their interactions. Used, for example, to combine risk loci (DNA sequence) with diseases to suggest potential new therapies based on molecular genetic information. They support molecular biology research, functional genomics research, and systems biology in general	

sequence features that could be extended to structural or functional properties. Sequences are annotated with information such as binding sites, exons, or experimental features. The annotations can be represented by standardized terminologies and information models such as the Sequence Ontology [7] and the Bioinformatics Sequence Markup Language. The former provides a structured, controlled terminology for annotating sequences, exchanging annotation data, and for describing sequence objects in databases. It is also part of the Open Biomedical Ontologies Foundry [8], which groups interoperable reference ontologies to describe features such as anatomy, phenotypes, biochemistry, diseases, and molecular functions and provides mappings between them.

GenBank is an annotated collection of all publicly available DNA sequences. In 2017, it contained data on over 250 billion nucleotide pairs and about 206 million sequence records and doubled in size every 18 months [6]. GenBank and the European Molecular Biology Laboratory (EMBL) database were launched in 1982. GenBank merged with the National Center for Biotechnology Information (NCBI) when it was established, and EMBL is now managed by the European Bioinformatics Institute and included in the European Nucleotide Archive. Both also collaborate with the DNA Data Bank of Japan (DDBJ) and exchange new and updated data daily. Many scientific journals now require submission of sequence information to a database prior to publication, supporting database growth.

Computerized amino acid sequence databases such as the National Biomedical Research Foundation protein sequence database managed by the Protein Information Resource were started around 1980. Swiss-Prot, created in 1986, developed methods and tools to ensure high-quality data. It contains rich annotations (e.g., protein functions, variants, and posttranslational modifications) and numerous links to other databases, including GenBank/EMBL/DDBJ and the Protein Data Bank, and assures good data curation. Swiss-Prot collaborates with the EMBL, and its computer-annotated nucleotide sequence database (trEMBL) complements

Swiss-Prot. Since 2002, Swiss-Prot, trEMBL, and the Protein Information Resource protein sequence database have been combined in Universal Protein Resource, or UniProt, the world's largest protein information catalog.

Structure Analysis Data

The three-dimensional structure of nucleic acids and proteins follows thermodynamically from the sequence of their component nucleotides or amino acids, respectively. Structure prediction relies mostly on observed sequence-structure relationships that are based on actual protein structures previously determined by X-ray crystallography or nuclear magnetic resonance spectroscopy and is realized by comparative modeling or by fold recognition.

Protein structure can be described at different levels. The primary structure is the amino acid sequence. The secondary structure is the stable substructures – mostly alpha helices and beta sheets – caused by local peptide folding. The tertiary structure is the three-dimensional configuration of the entire protein and is stabilized by bonds between amino acids that are not close to each other in the primary structure. The quaternary structure involves stable interactions among multiple folded proteins to form a functional complex. Sequence information is stored in the Protein Data Bank, along with atomic coordinates, literature citations, chemical characteristics, links with other databases, and classification of the structure according to terminologies such as the CATH Protein Structure Classification [9] and is represented in XML format as PDBML [10]. The data can be analyzed with the aid of viewers that create three-dimensional representations of the proteins. Good examples are PyMOL [11] and the NGL Viewer [12]. The Structural Biology Knowledgebase was developed by the Protein Structure Initiative with the aim of making the three-dimensional structures of most proteins easily obtainable from their corresponding DNA sequences. It ended its services in 2017, partly replaced by the Protein Data Bank.

Functional Analysis Data

The first gene database, Mendelian Inheritance in Man, was published in 1966 by the late Victor McKusick and has been available online as OMIM since 1987. It contains information about all known Mendelian disorders and their almost 16,000 associated genes. OMIM is linked to NCBI's Entrez Gene [13], which contains over 17 million entries about known and predicted genes from a wide range of species. Genes are identified by gene finding, a process that relies on the complete human genome sequence and on computational biology algorithms to identify DNA sequence stretches that are biologically functional. Determining the actual function of a found gene, however, requires *in vivo* research (creating “knockout” mice is one possibility), although bioinformatics is making it increasingly possible to predict the function of a gene based on its sequence alone, aided by a computational analysis of similar genes in other organisms.

Genetic data include chromosomal localization (locus), product, markers, phenotypes, and interactions and are based on several terminologies and annotations such as Gene Ontology [14], the classification of the Human Genome Organization Gene Nomenclature Committee (HGNC) [15], and a growing body of information about epigenetic factors (factors that modify genes without changing their DNA sequence) [16] and interactions with other genetic elements. Gene Ontology includes gene product annotation with respect to molecular function, cellular location, and biological role. HGNC links to OMIM, Entrez Gene, GenBank/EMBL/DDBJ, UniProt, PubMed, GENATLAS [17], GeneCard [18], and other gene databases.

Gene expression profiling measures the relative amount of mRNA expressed by thousands of genes at the same time, creating a global picture of cellular function. The most common (and least costly) technology is DNA microarray analysis, but the development of next-generation sequencing has increased the use of sequence-based techniques such as serial analysis of gene expression, or SAGE. Microarray analysis depends on the binding of an RNA sequence to its complementary DNA sequence. A DNA microarray is a slide or “chip” on which tiny amounts of thousands of different short DNA sequences (“probes”) are arranged. When a clinical sample of extracted cellular RNA is applied to the slide, the amount of mRNA that binds to each sequence is measured with specialized scanners, and values are often stored in a vendor-specific format.

Microarray data can be represented in two-dimensional “heat maps” where values are represented by colors, but the exchange of microarray data is difficult due to the lack of standardization. Several groups are working on the problem. The Functional GEnomics Data (FGED) (formerly known as the MGED) Society has defined the minimum information needed to document a DNA microarray experiment (Minimal Information About a Microarray Experiment, or MIAME) [19] and addresses ways to describe microarray designs, manufacturing information, experimental protocols, gene expression data, and data analysis results (Microarray Gene Expression Markup Language, or MAGE-ML, and MAGE-TAB). The MGED society collaborates with the Protein Structure Initiative and the Metabolomics Standards Initiative to develop the Functional Genomics Ontology, now combined with clinical and epidemiological research and biomedical imaging concepts in the Ontology for Biomedical Investigations [20]. Gene and protein expression results are stored in a MIAME-compliant format in public repositories such as the Gene Expression Omnibus at NCBI [21] and the ArrayExpress at the European Bioinformatics Institute.

Protein expression is significantly more complex than gene expression. The genome is relatively constant, while the proteome differs from cell to cell and over time, and the approximately 25,000 human genes correspond to about 1,000,000 proteins [22]. Additional complexity follows from the fact that mRNA is not always translated, proteins undergo posttranslational modifications, and many different proteins are created from splice variants of a single stretch of DNA.

The techniques used to identify proteins, measure their expression, and study their modifications and cellular localization are protein microarrays and mass

spectrometry. Protein microarrays [23] resemble DNA microarrays and conventionally use monoclonal antibodies or purified proteins as probes. Recent advances allow protein arrays to be created by *in situ* synthesis from corresponding DNA arrays. Proteins and their multiple forms produced by splice variants from a gene can be represented with the Protein Ontology [24], another member of the Open Biomedical Ontologies Foundry.

Metabolomics data are even more variable and complex than gene expression and protein expression data. Metabolomics databases such as the Human Metabolome Database [25] and BiGG (Biochemical, Genetic, and Genomic) models [26] combine chemical and molecular biology data with links to other proteomics and genomics databases.

Several knowledge bases combine different types of molecular biology elements and functional data. An example is the Kyoto Encyclopedia of Genes and Genomes, a knowledge base for linking genomes to biological systems and to the environment and for integrating genes and proteins, ligands, and molecular interactions and reaction networks. These databases, along with the gene and protein functional data resources discussed above, support molecular biology and functional genomics research. All of these resources are also used in the field of systems biology, which aspires to understand the organisms via complex biological system simulations.

Human Variation

With the possible exception of monozygotic twins, no two human beings are genetically identical. A common source of genetic difference between individuals is single-nucleotide polymorphisms, or SNPs (pronounced “snips”). SNPs are gene variations that involve a single nucleotide – that is, an A, T, C, or G in one or both copies of a gene, replaced by C, G, A, or T, respectively. SNPs can occur within the coding and noncoding regions of the genome. Not all coding region SNPs lead to changes in peptide sequences because of genetic code degeneracy. SNPs in noncoding regions can lead to changes in expression of genes. Most SNPs do not have effects on health and development; others have been found to be advantageous. SNPs lead to variations in susceptibility and development of common diseases, response to certain drugs, and effect of various environmental factors. Genome-wide association studies (GWAS) consider the statistical association between specific genome variations and human health conditions and analyze specific chromosome regions or whole genomes for those health-associated sites.

Structural variants are another source of genetic variation among humans. They include sequence inversions, insertions, deletions, copy number variations, and complex rearrangements.

The International HapMap Project [27] was the first to systematically explore human SNPs and is currently cataloging those found in different groups of people worldwide. The project is an open resource that helps scientists explore associations between haplotypes (a set of associated SNP alleles in a single region of a chromosome) found in different populations and common health concerns or diseases. The

project uses representative SNPs in the region of the genome referred to as Tag SNPs to determine the collection of haplotypes present in each subject.

dbSNP is a database maintained by the National Center for Biotechnology Informatics along with the National Human Genome Research Institute [28]. dbSNP includes other polymorphisms apart from SNPs. It includes both polymorphisms associated with known phenotypes and neutral polymorphisms. As of February 3, 2017, dbSNP contained 325.7 million reference SNPs.

The 1000 Genomes Project [29] (which is actually sequencing 2000 genomes) is investigating structural variants as well as SNPs in human population samples from Europe, Africa, East and South Asia, and the Americas. The 1000 Genomes Project sought to find genetic variants with frequencies of at least 1%. In its 7-year course, the project analyzed 2504 genomes from 26 populations [30, 31]. It is now available as the International Genome Sample Resource [32].

A catalog of GWAS studies and their disease-gene associations was created by the National Human Genome Research Institute [33]. The European Bioinformatics Institute (EMBL-EBI) maintains this database since 2015. The new catalog includes a graphical user interface, ontology supported search functionalities, and an improved curation interface. The catalog also includes ancestry and recruitment information for all studies [34].

The Human Gene Mutation Database maintains a catalog of germline mutations in nuclear genes that are associated with human inherited diseases [35]. As of January 2018, the database contained 220,270 mutation entries, accruing new entries at the rate of about 10,000 per year. Somatic mutations are covered by the COSMIC system, which is especially relevant for cancer [36], and mitochondrial mutations are covered by the MITOMAP database [37]. The Human Variome Project is an overarching initiative focused on collecting and curating all human genetic variation affecting human health [38]. It is considered the successor to the Human Genome Project [39] and the HapMap project. The Human Variome Project catalogs genome sequences and variations in the human species and develops standards associated with the use of genetic information in health care and clinical research communities. Other projects and databases involved with human variation include dbSAP (single amino-acid polymorphism database for cataloging protein variations [40]), dbVAR (database of structural variants [41]), GWAS Central (supports visual querying of summary-level association data in one or more genome-wide association studies [42]), OMIM, and SNPedia (wiki-style database with personal genome annotation [43]).

Translating from the Molecular World to the Clinical World

Clinical Application of -Omics Data

Molecular biology data are becoming increasingly important in clinical research, with a prominent example being cancer research. Cancer is a somatic genetic disease in which a series of mutations provide a cell with a reproductive advantage.

Cancer is therefore a logical target for research based on genomic, epigenetic, proteomic, and functional data. Cancer genomics, or oncogenomics, focuses on the genome associated with cancer, on identifying new oncogenes (growth-promoting genes that can lead to cancer when mutated) and tumor suppressor genes (growth-regulating genes that can lead to cancer when mutated), and on improving the diagnosis, prognosis, and treatment of cancer. Cancer markers (such as prostate-specific antigen – PSA) are cancer-associated products found in the blood or urine that are used for early detection of cancer, to classify cancer types, or to predict outcomes. Cancer-associated proteins can be used as targets for drug therapies (as tyrosine kinase is for imatinib in chronic myelogenous leukemia or HER2 is for tamoxifen in breast cancer).

Clinical research informatics plays a crucial role in these efforts, facilitating translation between the basic sciences, such as all the -omics discussed above, and clinical research. This translation and the use of molecular biology data for clinical applications require the integration of data from both worlds, the molecular biology and bioinformatics world, and the clinical research and medical informatics world, using new methods and resources, as described by Martin-Sánchez and colleagues [44] and demonstrated in examples cited below.

Integration of Molecular and Clinical Research Data

Researchers have made significant advances in the use of -omics data to describe and investigate how genes are expressed under various conditions. As mentioned earlier, however, gene expression varies between individuals and at different times even within the same individual [45]. Therefore, knowing the genomic signature of an individual is frequently not sufficient to predict the presence or probability of a given condition. This has a profound impact on clinical research and informs basic science. Demographic and clinical information (such as age, sex, symptoms, comorbidities, diagnostic test results, tobacco and alcohol use, and reactions to therapies) characterize a phenotype more precisely [46]. Early investigations [47, 48] demonstrated that simply using annotation data (semantic categories such as “Amino Acid, Peptide, or Protein,” “Pharmacologic Substance,” “Disease or Syndrome,” and “Organic Chemical”) within publicly available gene expression databases such as Gene Expression Omnibus allowed researchers to associate phenotypic data with gene expression data and discover gene-disease relationships. Combining clinical and environmental data with genomic data enables more efficient and accurate identification of how genes are expressed under specific conditions and how genetic makeup may affect treatment outcomes.

This translation and the use of multi-omics data for clinical research require their integration, interrogation, and assimilation. Novel informatics methods and tools are being developed to address these growing needs of clinical research [44, 49, 50]. Methods involved in the integration phase include resolving identities and linking various assets involved in research, semantics, and metadata standards for storage of data to support their secondary use, data quality assessment methods, and

integration platforms. Interrogation and assimilation phases include methods such as knowledge representation, information extraction, machine learning and data mining for knowledge discovery, simulation, indexing, and other Big Data methods. Some prominent examples of these methods are presented below.

Integrating, storing, and searching vast amounts of data and knowledge generated from -omics research require development, selection, and evaluation of existing or new semantic representations [51]. These resources set standards for how -omics terms are named, defined, and associated and how new knowledge is modeled, shared, and stored. The PhenoGO database, for example, contains gene-disease annotations that were derived from literature using several Gene Ontology annotation databases, the Unified Medical Language System, and other specialized ontologies [52]. The Unified Medical Language System Metathesaurus and the National Cancer Institute Thesaurus map annotation fields within genomic databases to standard concepts for integrating data used in translational research [48]. The Unified Medical Language System is also used to map textual annotations across microarray studies in order to join similar phenotypes and automatically construct disease classes [53]. Many ontological resources used in health settings utilize incompatible formats and different modeling languages, making it difficult to integrate those resources in projects that span biomedical domains, such as clinical and translational research on genotype-phenotype associations. The Lexical Grid (LexGrid) project seeks to bridge multiple ontologies and provides standard application programming interfaces for more robust access to the underlying terminologies and their concept associations [54]. The National Center for Biomedical Ontology's Bioportal [55] and the Open Biomedical Ontologies (OBO) consortium's Foundry [8] provide resources for managing biomedical ontologies representing various facets of -omics data as needed for data integration. The exposome domain requires extensive work for developing and evaluating existing ontologies [56]. The Utah Pediatric Research using Integrated Sensor Monitoring Systems (PRISMS) Center is evaluating existing ontologies such as the Chemical Entities of Biological Interest ontology (ChEBI) for use in exposomic research [57].

As an example of integration, pharmacogenetics is the study of genetically based responses to drugs. The Pharmacogenomics and Pharmacogenetics Knowledge Base (PharmGKB) was developed to store the genomic, phenotypic, and clinical information that was rapidly being generated [58]. PharmGKB contains both primary study data and derived knowledge about genes associated with drug responses and their associated phenotypes. Interactive online tools facilitate research on the way genomics affects drug responses.

In addition to the semantics, proper sharing and reuse of -omics data in clinical research require biomedical data to be findable, accessible, interoperable, and reusable according to the FAIR guiding principles [59]. The capture of sufficient metadata from heterogeneous data sources is a key requirement for successful data harmonization and integration [60]. The biomedical and healthCAre Data Discovery Index Ecosystem (bioCADDIE) project through the DataMed platform provides a means to discover various datasets for reuse [61] and describes metadata of various

omics datasets [62]. Other efforts such as those lead by the Research Data Alliance are developing standards for sharing various biomedical data.

Key challenges to data integration in translation research include the needs to support diverse translational research archetypes using heterogeneous data with varying semantic complexity. There is a need to support different security, privacy, and data governance policies involved when using these data. Clinical and -omics data can be integrated using the following methodologies: federation (where data is queried from distinct resources without copying or transferring the original data), aggregation (where data is compiling from different resources with intent to prepare combined datasets for data processing and analysis), and complex integration, assimilation, and interrogation (where certain facets of data are combined from each resource and include sequential ordering of querying the resources and reasoning with existing knowledge resources). Essential constructs in data that need to be described for any -omics integration for clinical research include the identities of persons (patients, participants, and providers) and organizations they belong to, metadata and semantics of the data, and its persistence, workflows, and infrastructure for integrating the various data. Examples of infrastructures for such integration are presented below. They can be classified as those offering static aggregation (e.g., i2b2), static federation (e.g., caBIG), and dynamic federation (e.g., OpenFurther).

Informatics for Integrating Biology and the Bedside (i2b2) was a National Center for Biomedical Computing research based at Brigham and Women's Hospital (Boston, MA) [63] and focused on building an informatics framework to bridge clinical research data and the vast data banks arising from basic science research in order to better understand the genetic bases of complex diseases. The i2b2 Center developed a computational infrastructure and methodological framework that allows institutions to store genomic and clinical data in a common format and use innovative query and analysis tools to discover cohorts and visualize potential associations. The system can be used in early research design to generate research hypotheses, to validate potential subjects, and to estimate population sizes. Once data have been collected, the same framework can be used for deeper analysis and discovery. The inclusion of genomic data allows clinical researchers to study genetic aspects of diseases and facilitates the translation of their findings into new diagnostic tools and therapeutic regimens. This framework has been implemented in numerous institutions such as the University of Utah [64] and is used by research groups to, for example, study the genetic mechanisms underlying the pathogenesis of Huntington disease [65] or predict the response to bronchodilators in asthma patients [66].

In light of the growing amount of cancer genomic data and basic and clinical research data, the National Cancer Institute sponsored the development of the cancer Biomedical Informatics Grid (caBIG) to accelerate research on the detection, diagnosis, treatment, and prevention of cancer [67]. caBIG's goal was to develop a collaborative information infrastructure that links data and analytic resources within and across institutions connected to the cancer grid (caGrid [68]). caBIG resources include clinical, microarray (caArray), and tissue (caTissue) data objects and

databases in standardized formats, clinical trial software, data analysis and visualization tools, and platforms for accessing clinical and experimental data across multiple clinical trials and studies. The National Mesothelioma Virtual Bank, a biospecimen repository of annotated cases that includes tissue microarrays and genomic DNA that supports basic, clinical, and translational research, incorporated portions of the caBIG infrastructure [69].

OpenFurther (OF [70]) is an informatics platform that supports federation and integration of data from heterogeneous and disparate data sources. It uses informatics and industry standards and is open and sharable. It systematically supports federated and centralized data governance models by using dynamic federation. OF links heterogeneous data types, including clinical, biospecimen, and patient-generated data. It also empowers researchers with the ability to assess feasibility of particular clinical research studies, export biomedical datasets for analysis, and create aggregate databases for comparative effectiveness research and exposomic research. With the added ability of probabilistic linking of unique individuals from these sources, OF is able to identify cohorts for clinical research and reduce enrollment issues. The main components of OF include an ontology server (OS) that stores local and standard terminologies as well as inter-terminology mappings. It also includes an in-house developed metadata repository (MDR) that stores metadata artifacts for each data source and the relationships between different data models. A query tool that researchers can leverage to design a clinical research query is also included, as well as a federated query engine that orchestrates queries between the query tool, MDR, OS, and the data sources. Finally, data source adapters to facilitate interoperability with data sources, administrative, and security components are also part of OF. More recently OF has been developed to perform Big Data integration of Internet of Things devices to perform exposomic research in the PRISMS project.

In addition to storing data generated from genotype-phenotype studies, new messaging standards are also needed so that information between systems can be shared for clinical collaboration. The Health Level 7 Clinical Genomics Special Interest Group (HL7 CG SIG) was formed to address this gap. While message standards have been developed separately for genomic and clinical data, the HL7 CG SIG's goal was to associate personal genomic data with clinical data. A data storage message encapsulates all raw genomic data as static HL7 information objects. As this stored information is accessed for clinical care or research purposes, a data access or display message retrieves the most relevant raw genomic data as determined by associated clinical information, and those data are combined with updated knowledge. Thus, the presented information is dynamic, embodies the most up-to-date genomic research, and is based on a patient's clinical or research record at the time of access [71]. In parallel to the HL7 CG SIG, the Clinical Data Interchange Standards Consortium (CDISC) was formed in order to develop data standards that enable interoperability of clinical research systems [72]. Additionally, the Biomedical Research Integrated Domain Group Project, a collaborative effort of stakeholders from the Clinical Data Interchange Standards Consortium, Health Level 7, the National Cancer Institute, and the US Food and Drug Administration,

is producing a “shared view of the dynamic and static semantics that collectively define the domain of clinical and preclinical protocol-driven research and its associated regulatory artifacts,” such as the data, organization, resources, rules, and processes involved [73]. As of this writing, neither the CDISC nor the Biomedical Research Integrated Domain Group have specifically addressed genomic information collected during clinical research, but both groups are likely to focus on this area in the near future.

Molecular Data to Support Clinical Research

Incorporation of -omics into clinical trials recruitment can help with the refinement of participant selection for clinical trials as responses of those with specific phenotypes can be evaluated. For example, people with differences in their genes for cytochrome P450 oxidase (CYP) vary in the way they metabolize certain drugs, and people who metabolize drugs slowly are at greater risk of adverse drug effects than those who metabolize them rapidly. Clearance of the antidepressant drug imipramine, for example, depends on CYP2D6 gene dosage. To achieve the same effect, patients with less active CYP2D6 alleles (“poor metabolizers”) require less drug than those with very active CYP2D6 alleles (“ultra rapid metabolizers”) [74]. Thus, selecting patients according to their metabolizing genotype when evaluating drug effects yields more useful information.

Molecular data is also applied to the randomization and stratification of patients selected for clinical trials according to prognostic and predictive markers. Several trials have discovered and validated such markers in oncology, and others are ongoing; a marker for breast cancer treatment is one example [75]. When trastuzumab – a monoclonal antibody against HER2 – was analyzed in a breast cancer population, no major response was seen, but when patients with an overexpressed HER2 receptor protein were targeted, significant responses could be observed [76]. If these trials would have been realized only on a population without genetic or proteomic selection criteria, this excellent new drug would have been discarded.

Application of Molecular Data to Disease

Mechanisms of Disease

Some diseases are mostly caused by genetic disorders, such as single-gene diseases (e.g., familial hypercholesterolemia, sickle cell anemia) or chromosomal disorders (e.g., Down’s syndrome). Other diseases, such as hypertension and diabetes mellitus, have an important genetic component. Molecular pathogenesis offers new understandings of the mechanisms involved in such diseases. For example, genes that enhance susceptibility to Type 1A diabetes have been identified and can predict disease risk [77]. A large amount of the research conducted

on the mechanisms of diseases is nonclinical in nature but offer useful findings for development of novel interventions.

At the genomic level, the Cancer Genome Project [78] aims at identifying sequence variants and mutations in somatic cells that are involved in the development of human cancers. Among its resources are the sequenced human genome and the COSMIC database. At the functional genomics level, the National Cancer Institute's Cancer Genome Anatomy Project is determining the expression profiles of normal cells, precancerous cells, and cancer cells [79], and at the proteomic level, the Clinical Proteomics Program of the National Cancer Institute and the US Food and Drug Administration [80] is searching for and characterizing new circulating cancer biomarkers. Recent efforts in understanding oncogenic mechanisms are attempting to integrate multi-omics data such as onco-proteogenomic studies traversing the cancer genome, proteome, and genome [81].

Diagnostic Methods and Therapeutic Application Studies

Combining genomic and clinical data provides opportunities to tailor therapy on an individual basis laying the foundations for personalized medicine [82]. Single-gene tests are being developed at a very rapid pace and are the bellwether of postgenomic diagnostic development. The NIH genetic testing registry hosted at the NCBI maintains a database for describing disease-gene relationships and available genetic tests, including information on purpose of testing, clinical utility, and analytical methods.

Many diagnostic tests are being developed by high-throughput techniques exemplified by microarrays. These studies provide information about biochemical changes in tissues and are especially useful for chronic diseases when they relate to modifications in disease states. Many such studies focus on neoplasms and have led to the development of molecular signatures that recognize clinically indistinguishable subtypes of cancers as well as subtype aggressiveness. This has included lymphomas [83] as well as leukemias [84], bladder cancer [85], sarcomas [86], head and neck cancers [87], kidney cancers [88], ovarian cancers [89], neuroblastoma [90, 91], and melanoma [92]. Many clinical trials involve therapeutic interventions. In breast cancer in particular, several commercial genomic assays for outcome prediction such as the MammaPrint are available [93]. The ongoing TAILORx and MINDACT clinical trials concentrate on outcomes [94]. Molecular therapies for lymphomas are also undergoing clinical testing [95].

Transplantation is another active research area. Heart transplant studies and microarray-based biomarker signatures have been ongoing, resulting in the CARGO clinical trials using an 11-gene signature called the Allomap genes. Recent studies indicate that the number and frequency of cardiac biopsies can be reduced when the Allomap signature indicates a low risk of rejection. The US Food and Drug Administration has cleared Allomap for use in transplant management [96]. New studies of other organ transplants indicate a similar promise of monitoring the risk of transplant rejection [97].

The application of molecular profiling appears to hold promise for autoimmune diseases. Clinically distinct rheumatic diseases, for example, show dysregulation of the type I interferon pathway that correlates with disease progression. Pharmacogenomic studies based on such profiling are underway [98, 99]. Infectious disease is another area which has been altered by molecular data and the associated technologies. Resequencing arrays can now rapidly identify bacteria and viruses in body fluids based on their gene sequences, thus eliminating the need for time-consuming culturing techniques [100].

Selecting appropriate doses of drugs metabolized by some CYPs has been simplified by a chip that detects a standard set of CYP2C19 and CYP2D6 mutations [101]. The chip, called AmpliChip, predicts how rapid a metabolizer a patient is. The chip is best used for selecting the initial dose of medications such as warfarin to attain optimal therapy as quickly as possible. This pharmacogenetic test is regulated as a medical device by the US Food and Drug Administration.

The growing population of consumers contributing their health data to directly access genetic testing resources provides opportunities for clinical investigators. Today, consumers can send a saliva or cheek swab sample to companies such as 23andMe [102], Navigenics (acquired by Thermo Fisher) [103], and deCODEme [104] for genotyping and a risk analysis for a wide variety of health conditions. Consumers can also obtain an ancestral path based on their DNA. They can gain detailed information about their genetic conditions at Web sites such as the National Library of Medicine's Genetics Home Reference [3]. They can also join groups of people with similar conditions on the 23andMe or PatientsLikeMe [105] and share their specific health and genetic data. Researchers affiliated with these sites use the contributed patient data to promote research on rare conditions and on conditions with limited research funding. Clinical investigators are utilizing this consumer-centric initiative for performing novel research projects.

Molecular Epidemiological Data

Molecular epidemiology is the study of how genetic and environmental risk factors, at the molecular level, contribute to diseases within families and in populations. In the cancer domain, molecular epidemiology studies explore the interactions between genes and the environment and their influence on cancer risk. "Environment" includes exposures to foods and chemicals as well as lifestyle factors. The new field of nutrigenomics focuses on how diet influences genome expression [106].

Genealogical data allows for the study of the familiarity of diseases and risk factors. A prominent genealogical resource is the Utah Population Database (UPDB), a computerized integration of pedigrees, vital statistics, and medical records of millions of individuals that helped demonstrate the heritability of many diseases, including cancers – some before the genetics was established [107]. Recent studies have combined the pedigree-based linkage studies with genome-wide association studies. One example demonstrated the linkage of bipolar disorder with loci on chromosomes 1, 7, and 20 [108]. Another demonstrated linkage of rheumatoid arthritis with several chromosomes [109].

The Future of Molecular Data in Clinical Research

Molecular data has clearly made its way into clinical research and rapidly into standard care for various diseases, health conditions, and therapies. This trend is likely to accelerate for many decades as the postgenomic era matures. The large number of single-gene tests is being augmented by multigene testing techniques. The Lynch syndrome test for nonpolyposis hereditary colon cancer involves full sequencing of four genes and two associated laboratory tests. The panel of 17 genes involved in testing for hypertrophic cardiomyopathy is in the final stages of development and clinical trials [110]. Proteomics tests via tandem mass spectroscopy form the basis for mandatory screening of newborns. Molecular signatures based on microarray functional analyses are used routinely in breast cancer and in the final stages of clinical trials for many other cancers. Patients who have undergone organ transplants are being monitored by blood tests and associated molecular signature analysis that indicates the risk of rejection. Other disorders are similarly being transformed by these new and powerful sets of genomic information.

The next frontier in the postgenomic era may involve integration of exposomic, epigenomic, microbiomic, and metagenomic data, as well as nanoparticle technology. Nanoparticles are measured in nanometers, which is the size domain of proteins. They are being investigated for many applications such as potential drug delivery vehicles [111, 112]. Specific particles can interact with tumors of a specific genotype.

The Precision Medicine Initiative is a recent initiative of the National Institutes of Health to revolutionize health by integrating these various types of -omics data for accelerating biomedical discoveries. Informatics needs for this integration include understanding their semantics and metadata and their representation to reflect direct biological pathway alterations as well as mutagenic and epigenetic mechanisms of genomic and environmental influences on the phenotype. For example, exposomics clearly lack semantic standards for use in clinical research [113]. In addition, current approaches to metadata discovery are highly dependent on manual curation, an expensive and time-consuming process. Automatic or semiautomatic approaches for metadata discovery are necessary to enhance heterogeneous biomedical data integration [114]. The Utah PRISMS integration platform is providing generalizable infrastructure for integrating various -omics data to perform clinical research [115]. The future will undoubtedly involve utilizing these novel data sources with novel informatics methods for next generation clinical research and development of new therapies.

References

1. Collins FS, Morgan M, Patrinos A. The human genome project: lessons from large-scale biology. *Science*. 2003;300(5617):286–90.
2. Crick FH. On protein synthesis. *Symp Soc Exp Biol*. 1958;12:138–63.
3. Mitchell JA, Fomous C, Fun J. Challenges and strategies of the genetics home reference. *J Med Libr Assoc*. 2006;94(3):336–42.

4. Miller GW, Jones DP. The nature of nurture: refining the definition of the exposome. *Toxicol Sci.* 2014;137(1):1–2.
5. Luscombe NM, Greenbaum D, Gerstein M. What is bioinformatics? A proposed definition and overview of the field. *Methods Inf Med.* 2001;40(4):346–58.
6. Benson DA, Cavanaugh M, Clark K, et al. GenBank Nucleic Acids Res. 2018;46(D1):D41–7.
7. Eilbeck K, Lewis SE. Sequence ontology annotation guide. *Comp Funct Genomics.* 2004;5(8):642–7.
8. Smith B, Ashburner M, Rosse C, et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol.* 2007;25(11):1251–5.
9. Cuff AL, Sillitoe I, Lewis T, et al. The CATH classification revisited – architectures reviewed and new ways to characterize structural divergence in superfamilies. *Nucleic Acids Res.* 2009;37(Database issue):D310–4.
10. Westbrook J, Ito N, Nakamura H, et al. PDBML: the representation of archival macromolecular structure data in XML. *Bioinformatics.* 2005;21(7):988–92.
11. PyMOL. <http://www.pymol.org>.
12. Rose AS, Hildebrand PW. NGL Viewer: a web application for molecular visualization. *Nucleic Acids Res.* 2015;43(W1):W576–9.
13. Maglott D, Ostell J, Pruitt KD, Tatusova T. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.* 2005;33(Database issue):D54–8.
14. Ashburner M, Ball CA, Blake JA, et al. Gene ontology: tool for the unification of biology. *Gene Ontol Consortium Nat Genet.* 2000;25(1):25–9.
15. White JA, McAlpine PJ, Antonarakis S, et al. Guidelines for human gene nomenclature (1997). HUGO Nomenclature Committee. *Genomics.* 1997;45(2):468–71.
16. You MH. Case study of a patient with Parkinson's disease. *Taehan Kanho.* 1991;30(5):56–60.
17. Frezal J. Genatlas database, genes and development defects. *C R Acad Sci III Sci Vie.* 1998;321(10):805–17.
18. Rebhan M, Chalifa-Caspi V, Prilusky J, Lancet D. GeneCards: a novel functional genomics compendium with automated data mining and query reformulation support. *Bioinformatics.* 1998;14(8):656–64.
19. Brazma A, Hingamp P, Quackenbush J, et al. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat Genet.* 2001;29(4):365–71.
20. Bandrowski A, Brinkman R, Brochhausen M, et al. The ontology for biomedical investigations. *PLoS One.* 2016;11(4):e0154556.
21. Edgar R, Domrachev M, Lash AE. Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* 2002;30(1):207–10.
22. Oh JE, Krapfenbauer K, Fountoulakis M, et al. Evidence for the existence of hypothetical proteins in human bronchial epithelial, fibroblast, amnion, lymphocyte, mesothelial and kidney cell lines. *Amino Acids.* 2004;26(1):9–18.
23. Stoevesandt O, Taussig MJ, He M. Protein microarrays: high-throughput tools for proteomics. *Expert Rev Proteomics.* 2009;6(2):145–57.
24. Natale DA, Arighi CN, Barker WC, et al. Framework for a protein ontology. *BMC Bioinform.* 2007;8(Suppl 9(Suppl 9)):S1.
25. Wishart DS, Tzur D, Knox C, et al. HMDB: the human metabolome database. *Nucleic Acids Res.* 2007;35(Database issue):D521–6.
26. King ZA, Lu J, Dräger A, et al. BiGG Models: a platform for integrating, standardizing and sharing genome-scale models. *Nucleic Acids Res.* 2016;44(D1):D515–22.
27. The International HapMap Project. *Nature.* 2003;426(6968):789–96.
28. dbSNP. www.ncbi.nlm.nih.gov/projects/SNP/.
29. Kaiser J. DNA sequencing. A plan to capture human diversity in 1000 genomes. *Science.* 2008;319(5862):395.
30. 1000 Genomes Project Consortium, Auton A, Brooks LD, et al. A global reference for human genetic variation. *Nature.* 2015;526(7571):68–74.
31. Sudmant PH, Rausch T, Gardner EJ, et al. An integrated map of structural variation in 2,504 human genomes. *Nature.* 2015;526(7571):75–81.

32. IGSR. The international genome sample resource. <http://www.internationalgenome.org>.
33. A Catalog of Published Genome-Wide Association Studies. <http://www.genome.gov/gwastudies/#1>.
34. MacArthur J, Bowler E, Cerezo M, et al. The new NHGRI-EBI catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* 2017;45(D1):D896–901.
35. Stenson PD, Mort M, Ball EV, et al. The human gene mutation database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Hum Genet.* 2014;133(1):1–9.
36. Bamford S, Dawson E, Forbes S, et al. The COSMIC (catalogue of somatic mutations in cancer) database and website. *Br J Cancer.* 2004;91(2):355–8.
37. MITOMAP. A human mitochondrial genome database. <http://www.mitomap.org>.
38. What is the human variome project? *Nat Genet.* 2007; 39(4):423.
39. Cotton RG, Auerbach AD, Axton M, et al. GENETICS. The human variome project. *Science.* 2008;322(5903):861–2.
40. Cao R, Shi Y, Chen S, et al. dbSAP: single amino-acid polymorphism database for protein variation detection. *Nucleic Acids Res.* 2017;45(D1):D827–32.
41. Phan L, Hsu J, LQM T, et al. dbVar structural variant cluster set for data analysis and variant comparison. *F1000Res.* 2016;5:673.
42. Beck T, Hastings RK, Gollapudi S, et al. GWAS Central: a comprehensive resource for the comparison and interrogation of genome-wide association studies. *Eur J Hum Genet.* 2014;22(7):949–52.
43. Cariaso M, Lennon G. SNPedia: a wiki supporting personal genome annotation, interpretation and analysis. *Nucleic Acids Res.* 2012;40(Database issue):D1308–12.
44. Martin-Sanchez F, Iakovidis I, Norager S, et al. Synergy between medical informatics and bioinformatics: facilitating genomic medicine for future health care. *J Biomed Inform.* 2004;37(1):30–42.
45. Butte AJ, Kohane IS. Creation and implications of a phenome-genome network. *Nat Biotechnol.* 2006;24(1):55–62.
46. Chen DP, Weber SC, Constantinou PS, et al. Clinical arrays of laboratory measures, or “clin-arrays,” built from an electronic health record enable disease subtyping by severity. *AMIA Annu Symp Proc.* 2007:115–119.
47. Butte AJ, Chen R. Finding disease-related genomic experiments within an international repository: first steps in translational bioinformatics. *AMIA Annu Symp Proc.* 2006:106–110.
48. Shah NH, Jonquet C, Chiang AP, et al. Ontology-driven indexing of public datasets for translational bioinformatics. *BMC Bioinforma.* 2009;10(Suppl 2)(Suppl 2):S1.
49. Kahn MG, Weng C. Clinical research informatics: a conceptual perspective. *J Am Med Inform Assoc.* 2012;19(e1):e36–42.
50. Tenenbaum JD, Avillach P, Benham-Hutchins M, et al. An informatics research agenda to support precision medicine: seven key areas. *J Am Med Inform Assoc.* 2016;23(4):791–5.
51. Tenenbaum JD, Sansone S-A, Haendel M. A sea of standards for omics data: sink or swim? *J Am Med Inform Assoc.* 2014;21(2):200–3.
52. Sam LT, Mendonca EA, Li J, et al. PhenoGO: an integrated resource for the multiscale mining of clinical and biological data. *BMC Bioinforma.* 2009;10(Suppl 2)(Suppl 2):S8.
53. Liu CC, Hu J, Kalakrishnan M, et al. Integrative disease classification based on cross-platform microarray data. *BMC Bioinforma.* 2009;10(Suppl 1)(Suppl 1):S25.
54. Pathak J, Solbrig HR, Buntrock JD, et al. LexGrid: a framework for representing, storing, and querying biomedical terminologies from simple to sublime. *J Am Med Inform Assoc.* 2009;16(3):305–15.
55. Whetzel PL, Noy NF, Shah NH, et al. BioPortal: enhanced functionality via new Web services from the national center for biomedical ontology to access and use ontologies in software applications. *Nucleic Acids Res.* 2011;39(Web Server issue):W541–5.
56. Mattingly CJ, Boyles R, Lawler CP, et al. Laying a community-based foundation for data-driven semantic standards in environmental health sciences. *Environ Health Perspect.* 2016;124(8):1136–40.

57. Burnett N, Gouripeddi R, Cummins M et al. Towards a molecular basis of exposomic research. AMIA joint summits on translational science proceedings AMIA Summit on Translational Science, San Francisco. 2018;320.
58. Hewett M, Oliver DE, Rubin DL, et al. PharmGKB: the pharmacogenetics knowledge base. *Nucleic Acids Res.* 2002;30(1):163–5.
59. Wilkinson MD, Dumontier M, Aalbersberg IJJ, et al. The FAIR guiding principles for scientific data management and stewardship. *Sci Data.* 2016;160018:3.
60. Gouripeddi R, Schultz D, Bradshaw R, Facelli J. FURTheR: an infrastructure for clinical, translational and comparative effectiveness research. *AMIA Annu Symp Proc.* Washington, DC: 2013;513.
61. Chen X, Gururaj AE, Ozyurt B, et al. DataMed – an open source discovery index for finding biomedical datasets. *J Am Med Inform Assoc.* 2018;25(3):300–8.
62. Sansone S-A, Gonzalez-Beltran A, Rocca-Serra P, et al. DATS, the data tag suite to enable discoverability of datasets. *Sci Data.* 2017;4:170059.
63. Murphy SN, Mendis ME, Berkowitz DA, et al. Integration of clinical and genetic data in the i2b2 architecture. *AMIA Annu Symp Proc.* 2006;1040.
64. Deshmukh VG, Meystre SM, Mitchell JA. Evaluating the informatics for integrating biology and the bedside system for clinical research. *BMC Med Res Methodol.* 2009;9(1):70.
65. Lee J-M, Ivanova EV, Seong IS, et al. Unbiased gene expression analysis implicates the huntingtin polyglutamine tract in extra-mitochondrial energy metabolism. *PLoS Genet.* 2007;3(8):e135.
66. Himes BE, Wu AC, Duan QL, et al. Predicting response to short-acting bronchodilator medication using Bayesian networks. *Pharmacogenomics.* 2009;10(9):1393–412.
67. caBIG Tools. <https://biospecimens.cancer.gov/caBigTools.asp>.
68. Saltz J, Oster S, Hastings S, et al. caGrid: design and implementation of the core architecture of the cancer biomedical informatics grid. *Bioinformatics.* 2006;22(15):1910–6.
69. Amin W, Parwani AV, Schmandt L, et al. National mesothelioma virtual bank: a standard based biospecimen and clinical data resource to enhance translational research. *BMC Cancer.* 2008;8(1):236.
70. OpenFurther. <http://openfurther.org>.
71. Shabo A. The implications of electronic health record for personalized medicine. *Biomed Pap Med Fac Univ Palacky Olomouc Czech Repub.* 2005;149((2):suppl):251–8.
72. Clinical Data Interchange Standards Consortium (CDISC). <http://www.cdisc.org/>.
73. Biomedical Research Integrated Domain Group (BRIDG). <https://bridgmodel.nci.nih.gov/about-bridg>.
74. Schenk PW, van Fessem MA, Verploegh-Van Rij S, et al. Association of graded allele-specific changes in CYP2D6 function with imipramine dose requirement in a large group of depressed patients. *Mol Psychiatry.* 2008;13(6):597–605.
75. Loi S, Buyse M, Sotiriou C, Cardoso F. Challenges in breast cancer clinical trial design in the postgenomic era. *Curr Opin Oncol.* 2004;16(6):536–41.
76. Vogel CL, Cobleigh MA, Tripathy D, et al. Efficacy and safety of trastuzumab as a single agent in first-line treatment of HER2-overexpressing metastatic breast cancer. *J Clin Oncol.* 2002;20(3):719–26.
77. Jahromi MM, Eisenbarth GS. Cellular and molecular pathogenesis of type 1A diabetes. *Cell Mol Life Sci.* 2007;64(7–8):865–72.
78. Cancer Genome Project. <http://www.sanger.ac.uk/science/groups/cancer-genome-project>.
79. Cancer Genome Anatomy Project. <http://cgap.nci.nih.gov>.
80. FDA-NCI Clinical Proteomics Program. <http://home.ccr.cancer.gov/ncifdaproteomics/default.asp>.
81. Dimitrakopoulos L, Prassas I, Diamandis EP, Charakes GS. Onco-proteogenomics: multi-omics level data integration for accurate phenotype prediction. *Crit Rev Clin Lab Sci.* 2017;54(6):414–32.
82. Mancinelli L, Cronin M, Sadée W. Pharmacogenomics: the promise of personalized medicine. *AAPS PharmSci.* 2000;2(1):E4–41.

83. Leich E, Hartmann EM, Burek C, et al. Diagnostic and prognostic significance of gene expression profiling in lymphomas. *APMIS*. 2007;115(10):1135–46.
84. Codony C, Crespo M, Abrisqueta P, et al. Gene expression profiling in chronic lymphocytic leukaemia. *Best Pract Res Clin Haematol*. 2009;22(2):211–22.
85. Chan KS, Espinosa I, Chao M, et al. Identification, molecular characterization, clinical prognosis, and therapeutic targeting of human bladder tumor-initiating cells. *Proc Natl Acad Sci U S A*. 2009;106(33):14016–21.
86. Hoffman AC, Danenberg KD, Taubert H, et al. A three-gene signature for outcome in soft tissue sarcoma. *Clin Cancer Res*. 2009;15(16):5191–8.
87. Gold KA, Kim ES. Role of molecular markers and gene profiling in head and neck cancers. *Curr Opin Oncol*. 2009;21(3):206–11.
88. Petillo D, Kort EJ, Anema J, et al. MicroRNA profiling of human kidney cancer subtypes. *Int J Oncol*. 2009;35(1):109–14.
89. Yoshihara K, Tajima A, Komata D, et al. Gene expression profiling of advanced-stage serous ovarian cancers distinguishes novel subclasses and implicates ZEB2 in tumor progression and prognosis. *Cancer Sci*. 2009;100(8):1421–8.
90. Volchenboum SL, Cohn SL. Are molecular neuroblastoma classifiers ready for prime time? *Lancet Oncol*. 2009;10(7):641–2.
91. Vermeulen J, De Preter K, Naranjo A, et al. Predicting outcomes for children with neuroblastoma using a multigene-expression signature: a retrospective SIOPEN/COG/GPOH study. *Lancet Oncol*. 2009;10(7):663–71.
92. Ugurel S, Utikal J, Becker JC. Tumor biomarkers in melanoma. *Cancer Control*. 2009;16(3):219–24.
93. Kim C, Taniyama Y, Paik S. Gene expression-based prognostic and predictive markers for breast cancer: a primer for practicing pathologists. *Arch Pathol Lab Med*. 2009;133(6):855–9.
94. Sotiriou C, Pusztai L. Gene-expression signatures in breast cancer. *N Engl J Med*. 2009;360(8):790–800.
95. Rabson AB, Weissmann D. From microarray to bedside: targeting NF-kappaB for therapy of lymphomas. *Clin Cancer Res*. 2005;11(1):2–6.
96. XDX's AlloMap(R) Gene Expression Test Cleared By U.S. FDA For Heart Transplant 'Recipients. <http://www.medicalnewstoday.com/articles/119546.php>.
97. Khatri P, Sarwal MM. Using gene arrays in diagnosis of rejection. *Curr Opin Organ Transplant*. 2009;14(1):34–9.
98. van Baarsen LG, Bos CL, van der Pouw Kraan TC, Verweij CL. Transcription profiling of rheumatic diseases. *Arthritis Res Ther*. 2009;11(1):207.
99. Bauer JW, Bilgic H, Baechler EC. Gene-expression profiling in rheumatic disease: tools and therapeutic potential. *Nat Rev Rheumatol*. 2009;5(5):257–65.
100. Lin B, Malanoski AP. Resequencing arrays for diagnostics of respiratory pathogens. *Methods Mol Biol*. 2009;529(Chapter 15):231–57.
101. Individualize Drug Dosing Based on Metabolic Profiling with the AmpliChip CYP450 Test. <http://www.amplichip.us/>.
102. 23andMe. Genetics just got personal. <https://www.23andme.com/>.
103. There's DNA. And then there's what you do with it. <http://www.thermofisher.com/us/en/home.html>.
104. deCODE your health. <https://www.decode.com>.
105. PatientsLikeMe. Patients helping patients live better every day. <http://www.patientslikeme.com>.
106. Kaput J, Rodriguez RL. Nutritional genomics: the next frontier in the postgenomic era. *Physiol Genomics*. 2004;16(2):166–77.
107. Cannon-Albright LA, Thomas A, Goldgar DE, et al. Familiality of cancer in Utah. *Cancer Res*. 1994;54(9):2378–85.
108. Hamshere ML, Schulze TG, Schumacher J, et al. Mood-incongruent psychosis in bipolar disorder: conditional linkage analysis shows genome-wide suggestive linkage at 1q32.3, 7p13 and 20q13.31. *Bipolar Disord*. 2009;11(6):610–20.

109. Hamshere ML, Segurado R, Moskvina V, et al. Large-scale linkage analysis of 1302 affected relative pairs with rheumatoid arthritis. *BMC Proc.* 2007;1(Suppl 1):S100.
110. Bos JM, Towbin JA, Ackerman MJ. Diagnostic, prognostic, and therapeutic implications of genetic testing for hypertrophic cardiomyopathy. *J Am Coll Cardiol.* 2009;54(3):201–11.
111. la Fuente de M, Csaba N, Garcia-Fuentes M, Alonso MJ. Nanoparticles as protein and gene carriers to mucosal surfaces. *Nanomedicine (Lond).* 2008;3(6):845–57.
112. Emerich DF, Thanos CG. Targeted nanoparticle-based drug delivery and diagnosis. *J Drug Target.* 2007;15(3):163–83.
113. Martin-Sanchez F, Gray K, Bellazzi R, Lopez-Campos G. Exosome informatics: considerations for the design of future biomedical research information systems. *J Am Med Inform Assoc.* 2014;21(3):386–90.
114. Wen J, Gouripeddi R, Facelli JC. Metadata discovery of heterogeneous biomedical datasets using token-based features. *IT Convergence and Security 2017*, Singapore: Springer. *Dermatol Sin.* 2018;449(6):60–7.
115. University of Utah PRISMS informatics center. <http://prisms.bmi.utah.edu/project/>.

Part II

Data and Information Systems Central to Clinical Research



Clinical Research Information Systems

9

Prakash M. Nadkarni

Abstract

Information systems can support a host of functions and activities within clinical research enterprises. We consider issues and workflows unique to clinical research that mandate the use of a Clinical Research Information System and distinguish its functionality from that provided by electronic medical record systems. We then describe the operations of a CRIS during different phases of a study. We finally discuss briefly the issues of standards and certification.

Keywords

Clinical research information systems · Clinical study data management · Research data management · Regulatory support systems · Research logistics support · Real-time electronic data validation

Clinical Research Information Systems (CRISs) are software applications intended to handle one or more aspects of supporting clinical research. Their effective use can play an important role in reducing the costs of conducting research studies [1]. While distinct from electronic health record (EHR) systems, they must typically interoperate bidirectionally with EHRs. Initially, many systems supported only individual aspects of clinical research, such as primary data capture, study logistics, patient recruitment, and so on, but over time, the systems have tended to grow more monolithic and pricier. Nonetheless, despite the proliferation of commercial software, special problems still arise that only custom software development can solve.

P. M. Nadkarni, MD (✉)

Interdisciplinary Graduate Program in Informatics and College of Nursing, University of Iowa, Iowa City, IA, USA

e-mail: prakash-nadkarni@uiowa.edu

In this chapter, we provide the reader with a feel for the various issues and processes related to CRISs. We also emphasize practical issues of CRIS operation that have little to do with informatics per se but which can be ignored only at one's peril.

CRIS Vendor Models

The larger commercial CRISs require a financial investment almost as formidable as that for an institutional EHR and require a sizeable team with information technology (IT), informatics, and clinical and clinical research expertise to administer and support, and are therefore viable only when deployed institution wide. They can manage an arbitrary number of clinical studies within a single physical database design. These are sold outright to the customer, installed at a customer site, and operated by customer personnel.

Institution-level IT projects have high risk and require organizational commitment. Many customers are understandably squeamish and also lack the budgets to hire or retain skilled staff. Therefore, certain CRIS vendors (like some EHR vendors) follow a different approach, hosting the software in the Cloud, with Web-based or non-Web-based client software on customer machines. Initial Institutional Review Boards' (IRB) concerns with Cloud data privacy and security have been addressed through the use of private clouds, and with the number of institution-level security breaches increasing, IRBs are now more accepting of remotely hosted data that is managed by skilled security professionals.

With remote hosting, vendor staff perform all administrative and software-development functions, including software upgrades, billing the customer based on the number of supported studies, the complexity of the processes needing support, and the number of electronic data capture instruments that have to be developed. This model has much lower up-front costs but potentially higher ongoing costs. However, the vendor can perform administrative/developer tasks much more cheaply than the customer, and a single developer's or administrator's time can be fully utilized in supporting multiple customer sites.

A third alternative in terms of ready-made software is an open-source system. While the software purchase costs are zero, the people who have to learn and run the software locally are not free. Such systems are rarely plug-and-play and require a team commitment much greater than that required to run a commercial package: while user groups may help answer simple questions, free software cannot be supported by its authors 24/7. (Some open-source vendors will, however, provide support for a fee.)

As with any software, one must study the documentation carefully to ensure that the software is a match for one's needs: the capability of different systems vary widely. REDCap, for example, while an excellent package for simple study designs such as surveys, currently has somewhat limited support for complex longitudinal study designs such as multiple arms of a study, where the forms, or the questions within a form, that are applicable at a particular time point vary with the patient. REDCap is also not intended for regulated studies such as those involving the Food

and Drug Administration (FDA), where there is a horde of electronic paperwork that must be maintained or generated.

Why Have Clinical Research Information Systems Evolved?

EHRs are not entirely suitable for supporting clinical research needs by themselves, for reasons relating to the differences between clinical research and patient-care process. (There is, however, a class of studies called “pragmatic clinical trials,” which are almost entirely EHR-based. We will discuss these later.) We describe these differences below while emphasizing that workflows involve interoperation with EHR-related systems.

In the account below, we will use the words “subject” and “patient” interchangeably while accepting that participants in a study may often be healthy. We will use “case report form” (CRF) to refer to either a paper or an electronic means of capturing information about a set of related parameters. The parameters are often called *questions* when the CRF is a questionnaire but may also be clinical findings or results of laboratory or other investigations.

The Concept of a Protocol Is Fundamental to CRISs

CRISs differ from EHRs in that their design is based on the concept of a *study*. The details of a given study – the experimental design, the CRFs used, the time points designated for subject encounters, and so on – constitute the *study protocol*. Sometimes, a *project* may involve multiple related studies performed by a research group or consortium, typically involving a shared pool of subjects, so that certain common data on these subjects – such as demographics or screening data – is shared between studies within the same project.

A CRIS must provide two essential functions: representing a protocol electronically and supporting electronic data capture. EHRs are not a good fit for the former objective, because patients typically show up when they are sick, at unanticipated times. Some EHR vendors have tried to adapt the EHR subcomponent related to cancer therapy protocols – which are also rigid with respect to time points and the interventions and tests applicable to each time point – to support basic CRIS functions. Such efforts have, to date, met with only a very modest degree of success. Individual CRIS offerings differ in how fully they can model a variety of protocols and the sophistication of their data capture tools.

CRISs Implement User Roles That Are Specific to Research Designs

Supporting Differential Access to Individual Studies

For an institutional CRIS that supports multiple studies, only a handful of individuals – typically, administrators and developers – will have access to all studies. In the

EHR setting, where a patient can be seen by almost any healthcare provider in the organization – though such access is audited – access to research subjects’ data must be limited to those individuals involved in the conduct of the study or studies in which that subject is participating. The vast majority of users, after logging on, will therefore see only the studies or projects to which they have been given access. Even here, their *privileges* – the actions they can perform once they are within a study – will vary. For example, an investigator may be a principal investigator in one study, but only a co-investigator in another: therefore certain administrative-type privileges may be denied in the latter study.

While EHRs also support user roles, these are related to functions related to patient care (e.g., clinician, nurse, nurse assistant, pharmacist, lab tech, supervisor vs. non-supervisor, etc.), and roles related to clinical research are quite different. Considerable design and implementation effort would be required to support clinical research roles, and hospitals not in the research business would balk at having to pay for functionality that they will never use.

Representing Experimental Designs

Clinical research often involves an experimental design. In some designs, two or more groups of subjects are given different therapeutic agents (including placebo) or procedures. The designs are typically double-blinded. That is, neither the patient nor the caregiver/s dealing with the patient (nor even the chief investigator) knows what the patient is receiving: the patient simply receives a custom-formulated medication with one’s name on the container. It is occasionally necessary to break the blinding for a given patient, e.g., if serious adverse effects develop and the patient needs specific therapy to counteract it. Therefore, some individuals (typically pharmacists who dispense the medication) are aware of the blinding scheme. CRIS software is aware of the study-specific privileges of the currently logged on user with respect to blinded data: EHR software lacks this capability, because traditional patient care is never blinded.

The Scope of a CRIS May Cross Institutional or National Boundaries

A given clinical study may often be conducted by a research consortium that crosses institutional boundaries, with multiple geographically distributed sites. Very often, certain investigators in the consortium happen to be professional rivals who are collaborating only because a federal agency initiates and finances the consortium, selecting members through competitive review. Individual investigators would not care to have investigators from other sites access their own patients’ data. However, neutral individuals, such as the informatics and biostatistics team members and designated individuals affiliated with the sponsor, would have access to all patients.

Even if all consortium investigators trusted each other fully, regulations such as those related to the Health Insurance Portability and Accountability Act (HIPAA) limit unnecessary access of personal health information (PHI) to individuals not

directly involved in a patient's care. So, biostatisticians intending to analyze the data would generally not care to have access to it. Sometimes, selective PHI such as patient address might be necessary, e.g., if one is studying the fine-grained geographical distribution of the condition of interest.

The concept of enforcement of selective access to individual patients' data (*site restriction*) as well as *selective access to part of a patient's data* (PHI) based on the user's role and affiliation is again a critical issue that EHRs do not address.

For trans-institutional studies, CRIS solutions must increasingly use Web-technology to provide access across individual institutional firewalls. By contrast, EHRs, even when used in a geographically distributed setting (as for a network of community-based physicians), are still institutional in scope. Therefore, EHR vendors have been relatively slow to provide access this way: most still employ two-tier (traditional "fat" client-to-database server) access or access using remote log-in (through mechanisms such as Citrix). (One of the few vendors that provide Cloud-based Web access is Eclipsys.)

When a multi-site study is conducted across countries with different languages, the informatics challenges can be significant, as well-described in [2]. Besides co-ordination challenges, the same physical CRIS (which is hosted in the country where the main informatics team is located) must ideally present its user interface in different languages based on who has logged in. This feature, called *dynamic localization*, is possible to implement with relatively modest programming effort using Web-based technologies such as Java Enterprise Edition and Microsoft ASP.NET.

Localization relies on *resource files* containing text-string elements of the user interface (e.g., user prompts, form labels, error messages, etc.) for each language of use. In the software application, the programmer refers to these elements symbolically, rather than hard-coding prompts or messages in the code in a specific language. At runtime, the appropriate language-specific elements are pulled from the resource file and integrated into the user interface. The programming framework also automatically takes care of issues such as direction of text (e.g., left to right vs right to left in Hebrew and Arabic) and display of dates and times (e.g., mm/dd/yyyy vs dd/mm/yyyy) without the programmer having to worry about these issues.

The language that the application's user interface will use depends on machine and Web-browser default-language settings, though some applications may also rely on a configuration file setup by the user. While several commercial Web sites such as Google and Amazon implement dynamic localization, to the best of our knowledge no existing commercial CRIS has employed it currently, though it is not too difficult to do.

Certain Low-Risk Clinical Studies May Not Store Personal Health Information

In EHR-supported processes involving patient care, the Joint Commission on the Accreditation of Hospital Organizations recommends the use of *at least* two personal identifiers [3] to ensure that errors due to treatment of the wrong patient are

minimized. In contrast, in certain multi-site clinical studies that involve minimal risk to the patient (such as purely observational studies), Institutional Review Boards will not permit PHI entry into a CRIS: patients are often identified only by a machine-generated “Study ID,” and the correspondence between a Study ID and an actual patient is stored in a separate system. In anonymous surveys, the respondents will not even volunteer PHI – at best, they may supply gender and age range.

Note: Especially for longitudinal studies, using IDs this way with extra manual processes risks the error of entering/editing data for the wrong patient, unless the Study ID incorporates extra check digits to prevent an invalid Study ID, e.g., one that is digit-substituted or digit-transposed, from being accepted. The author has witnessed the fear of allowing PHI to be entered even in multi-site studies where physical injury, e.g., dose escalations of a toxic drug, would result from decisions accidentally made for the wrong patient. In other words, patient safety trumps patient privacy: IRBs need to be gently educated about this fact.

Workflow in Clinical Research Settings Is Mostly Driven by the Study Calendar

Most research studies are conducted in ambulatory (outpatient) settings: the expense of continuous subject monitoring through admission to a hospital or research center is rarely mandated. Consequently, patient visits to the clinic or hospital are scheduled based on the study’s design. The schedule of visits, worked out relative to a reference “time zero” (such as the date of the baseline screening and investigations), is called the *Study Calendar*. Obviously, all patients do not enroll in a given study at the same time: they typically trickle in. The application of the study calendar to a single patient creates a *Subject Calendar* for that patient.

In a simple study design such as a one-time survey, there is only one event, so a calendar is not needed. However, for any longitudinal study, whether observational or interventional, calendar capability is essential. CRISs also typically allow for “unscheduled” visits that do not fall on calendar time points, such as those required to treat medical emergencies due to adverse drug effects.

Some CRIS software uses the more general term “Event” instead of “Visit” to reflect the fact that certain critical time points in the study calendar may not necessarily involve actual visits by a subject but will still drive workflow. For example, 1 week before the scheduled visit date, a pre-visit reminder event will drive a workflow related to mailing of form-letter reminders. Thus, the Subject Calendar is really a *Calendar of Events* rather than a calendar of visits.

Time Windows Associated with Events

One should note that, in order to allow for subjects’ convenience, and because certain scheduled days may fall on weekends or public holidays, subject calendar dates in longitudinal studies need to have some slack. Investigators determine the permissible slack or *window* based on the event, as well as the protocol or type of study. Thus, the “1-year follow-up” event may be allowed to occur between 11and

13 months. In a natural history/observational study, the windows might be 6 months wide, while for pharmacokinetic studies of fast-acting IV drugs, the acceptable window might be measured in minutes.

Based on the study calendar, a given subject can, soon after enrollment, receive a pre-computed calendar in advance, with minor adjustments made to suit the subject's convenience either at participation commencement or as the study progresses. Such adjustments are permissible as long as the event/visit falls within the permissible time window.

The Event-CRF Cross-Table

At each event, specific actions are performed – e.g., administration of therapy, particular evaluations – and units of information gathered in individual CRFs. The “Event-CRF Cross-Table” records the association of individual events with individual CRFs. For expense and patient-safety reasons, all investigations are not carried out at all events or with equal frequency: costly and/or highly invasive tests (e.g., organ biopsies) are much fewer than cheaper or routine tests.

CRISs must *enforce the Study's Event-CRF cross-table constraints*. That is, a research-team member should not be allowed to accidentally schedule a 3-month MRI when the protocol mandates a 6-month MRI instead. Similarly, accidentally creating a CRF instance for an event where it doesn't apply should be disallowed. Cross-table constraint enforcement allows accurate pooling, and accurate interpretation, of multiple patients' data because the corresponding data points for all patients are properly aligned chronologically.

The CRIS should also provide advance alerts for the research staff about which subjects are due for a visit and what event that visit corresponds to, so that the appropriate workflow (e.g., scheduling of use of a scarce resource like a PET scanner) can be planned. This allows *advance reminders* to subjects either through form letters, phone messages, or e-mail. (Reminders are one feature that today's EHRs support very well: missed office visits translate into lost revenue.) Timely alerts about *missed visits* are particularly critical, because even if a subject shows up after, the data for the delayed visit may not be usable if it falls outside that event's time window.

Clinical Research Subjects Are Not Typical “Patients”

Clinical research subjects differ from the typical patients whose care an EHR supports.

- EHRs support processes where caregivers (rather than research staff) interact with patients in processes that are either preventive (e.g., annual physical exams) or therapeutic in nature. In many clinical studies, by contrast, the subjects may be healthy volunteers who are involved in processes that have no direct relationship to caregiving, such as performing cognitive tasks or responding to standard questionnaires in anonymous surveys.

- In most studies, a large number of potential subjects are screened for recruitment. Many individuals eligible on initial criteria may, on detailed screening via a questionnaire, fail to meet the study's eligibility criteria. Even among eligible individuals, it often takes persistent persuasion over several encounters, via phone calls or personal interviews, to secure their participation, and many potential subjects still decline. All the while, the CRIS must record contact information about potential subjects and keep a log of all encounters, so that recruiting staff are paid for the time and effort invested.
- In genetic-disease research, one type of study design involves large groups of subjects who are related to each other through marriage and common ancestors (i.e., *pedigrees*). In such situations, to increase the power of eventual data analysis, one may include "pseudo-subjects": long-deceased ancestral individuals (e.g., great-grandparents) who connect smaller families, even though almost nothing is known about them.

CRISs Often Need to Support Real-Time Self-Reporting of Subject Data

In research studies involving surveys or patient-reported outcomes, it is necessary to support self-entry by subjects or to support bulk import of data from external survey systems. Many subjects are more than capable of using Web-based computer applications for work or personal purposes, so it is preferable and more resource-efficient to allow such patients to fill up such CRFs via the Web at a time and location (e.g., home) convenient to them, rather than mandate a visit or a phone interview. CRISs that support self-entry by subjects allow informatics staff to provide a limited log-in to subjects and also to specify which forms are subject-enterable. When the subject logs in, only such forms will be presented for data entry.

Clinical Research Data Capture Is More Structured Than in Patient Care

In clinical care, a patient may present with any disease: even in clinical specialties, a broad range of conditions are possible. Especially in primary or emergency care, the only sufficiently flexible way to capture most information other than vital signs or lab tests is through the narrative text of clinical notes. Structured data only arises when a patient is being worked up through a specific protocol where the required data elements are known in advance, e.g., for coronary bypass, cataract surgery, or when partial structure can be imposed (e.g., for a chest X-ray examination).

Information extraction from narrative text into analyzable, structured form is difficult because of issues such as medical-term synonymy and the telegraphic, often non-grammatical nature of the notes. By contrast, in most clinical research, patients are preselected for specific clinical condition, with the desired data elements known

in advance. Therefore, CRFs maximize use of elements that require numeric or discrete responses (e.g., values selected from a list of choices).

Occasionally, in studies that have dual objectives – i.e., research combined with clinical care – such forms will occasionally contain narrative text elements like “Additional Comments” and “If Other, Please Specify,” but such elements are relatively modest in number. A good research team will monitor the contents of such fields continuously, looking for frequently occurring textual responses that can be discretized in later versions of the CRF, thereby making it faster to fill and more analyzable.

CRIS Electronic Data Capture Needs to Be Robust and Flexible and Efficient to Setup

Data capture in many research settings (notably psychiatry/psychology) is typically far more extensive than in EHRs. Numerous questionnaires designed specifically for research are too lengthy for convenient use by busy caregivers or by patients who are not compensated for their study participation. The long length of such CRF risks inconsistency during data capture. Consequently, CRISs must provide extensive support for real-time data validation; for instance:

- Validation at the individual field level includes data type-based checks for dates and numbers, range checking, preventing out of range values by presenting a list of choices, regular expression checks for text, spelling check for the rare circumstances where narrative text must be supported, and mandatory field check (blank values not permitted). Certain values (especially dates) can be designated as approximate – accurate only to a particular unit of time such as month or year – if the subject does not recall a precise date. Fields can also be designated as having their contents missing for specified reasons such as failure of subject to recall, refusal to answer the answer, or change in a form version (a new question is introduced, so that data created with older version does not have the response for this question). Such reasons may often be specific to a given study.
- Cross-field validation can occur within a form through simple rules – e.g., the sum of the individual field values of a differential WBC count must equal 100.
- The more powerful packages will even support consistency checks across the entire database, e.g., by comparing a value entered for a specific parameter with the value entered for the previous event where the CRF applies.
- Support of computations where the values of certain items are calculated through a formula based on other questions in the form whose values are filled in by the user.

In addition to simple validation techniques, which rely on the software pointing out a user’s mistake, many facilities are ergonomic aids that are both preventive and streamline data entry.

- The use of *default values* for certain fields can speed data entry.
- *Skip logic* is employed when a particular response to a given question (e.g., an answer of “No” to “Have you been diagnosed with cardiovascular disease?”) causes subsequent questions for details of this disease to become disabled or invisible. Conversely, to minimize screen clutter, the detail questions may be invisible by default, and a Yes response makes them visible.
- *Dynamic (conditional) lists*: Certain lists may change their contents based on the user’s selection from a previous list. For example, some implementations of the National Bone Marrow Donor Program screening form will ask about the broad indication for transplant: based on the indication chosen, another list will change its contents to prompt for the specific sub-indication. This feature, typically implemented using Web-based technologies such as asynchronous JavaScript over XML (AJAX) [4], reduces the original 15-page paper questionnaire (which contains instructions such as “If you chose Hodgkin’s disease, go to page 6”), into a two-item form.
- Certain experimental designs, as described in [5, 6], require more than one research team member to evaluate the same subject (or the same tissue from the same subject) for the same logical encounter. Each team member performs an evaluation or rating, and this design intends to estimate interobserver variability or agreement in an attempt to increase reliability.
- Issues of privileges specific to individual user roles arise. Some users may only be allowed to view the data in forms; others may also edit their contents, while some with administrator-level privileges may be permitted to lock CRF data for individual forms or subjects to prevent retrospective data alteration. Certain designated CRFs may be editable only by those responsible for creating their data. Certain fields within certain CRF can be populated during primary data entry only by specific personnel, e.g., adjudicators.
- Finally, certain research designs, such as those involving psychometrics, may require the order of questions in a particular electronic CRF to be changed randomly. In computerized adaptive testing [7], even the questions themselves are not fixed: depending on how the subject has responded to previous questions, different new questions will appear.

While EHRs increasingly allow sophisticated data capture and also allow self-entry through patient portals, CRF features are very primitive compared to the best CRISs, especially with respect to adaptive forms, such as developed by the PROMIS consortium [8].

Use of Data Libraries

A significant part of the effort of electronic protocol representation involves CRF design. To speed up the process, many CRISs use a *data library*, which is essentially a type of metadata repository. That is, the definitions of questions, groups of questions, and CRFs are stored so as to be reusable. For example, the definition of a question (including its associated validation information) can be used in multiple

CRFs. (Thus, Hemoglobin's definition can be used in a form for anemia as well as traumatic blood loss).

Similarly, the same CRF can be used across multiple studies dealing with the same clinical domain: standard CRFs, such as laboratory panels, can be used in a variety of research domains. For the last situation, some CRISs will allow study-level customization, so that, for a given study, only a subset of all questions in a CRF will be shown to the user: questions that the investigator considers nonrelevant can be hidden.

This is one area where REDCap is currently somewhat deficient, making it less suitable for institutions that perform a vast number of studies in a single medical sub-domain – e.g., digestive diseases or cancer. While entire CRF definitions can be exported to Excel and stored externally, reusing single elements (such as Hemoglobin, above) across multiple studies is somewhat tedious and involves multiple manual steps that must be performed outside REDCap prior to importing a modified CRF definition. (On the other hand, the package is free, which may make the additional effort acceptable, given the high cost of commercial packages.)

Data Entry in Clinical Research May Not Always Be Performed in Real-Time: Quality Control Is Critical

EHRs capture patient-encounter data in real time or near-real time: CRISs are more adaptable to individual needs, supporting offline data entry with transcription from a source document if real-time capture is not possible, or bulk import of data such as laboratory values from external systems.

Having said this, other than the bulk-electronic-import scenario, there is virtually no excuse, in today's era of ubiquitous mobile devices, for offline entry and delayed CRF validation (other than highly unreliable internet connectivity). The major source of error in CRISs is overwhelmingly the source document. Delayed entry can result in missing data when source documents are misplaced or damaged. Also, if the absence of interactive validation results in source document errors, such errors are hard or impossible to salvage later: querying the source document's human originator works only if the operator remembers the encounter, which is likely only for very recent encounters. In these circumstances, double data entry (DDE), an archaic quality-control method based on comparing identical input created by two different human operators transcribing the same source document separately to ensure the fidelity of transcription, is useless [9].

Today, if delayed transcription is unavoidable, best quality control (QC) practices involve close to real-time data entry with CRFs maximally using interactive validation, followed by very timely *random audits* of a statistical sample of CRFs against the source documents. The proportion of audited CRFs depends on criteria such as the criticality of a particular CRF for the study's aims and clinical decision-making: the study's stage (early on, the sampling percentage is higher so as to get an idea of the error rate) and site in a multi-site study (some sites may be more lackadaisical). All questions on a single CRF are not equally

important, and therefore only some (typically critical items used for analysis or decision-making) are audited.

This approach, based on QC guru W. Edwards Deming's approach, allows concentration of limited resources in the areas of most potential benefit, as opposed to DDE, which indiscriminately weights every question on every CRF equally. In delayed-entry scenarios, a useful CRIS report will list which CRFs have not yet been entered for scheduled patient visits or which have been created after a delay longer than that determined to be acceptable.

CRIS-Related Processes During Different Stages of a Study

After discussing the special needs that CRISs meet, we now consider CRIS-related matters that arise in the different stages of a study. In chronological sequence, these stages are:

1. Study planning and protocol authoring
2. Recruitment/eligibility determination (screening)
3. Protocol management and study conduct (including patient monitoring and safety)
4. Analysis and reporting

Study Planning and Protocol Authoring

While clinical investigators are ultimately responsible for the overall study plan, a study plan must be developed in close collaboration with the biostatistics and informatics leads at the outset, rather than approaching them after a study plan has already been determined without their inputs. While experimental expert-type systems have been developed with the idea of helping clinical investigators design their own trials [10–12], their scope is too limited to address the diverse issues that human experts handle.

For example, a skilled biostatistician will work with the investigator to conduct a study of the relevant literature to determine previous research, availability of research subjects, relative incidence in the population of the condition(s) of interest, epidemiology of the outcome, the time course of the condition, risk factors, and vulnerable populations. Knowledge of these factors will provide a guide as to an appropriate experimental design. If the design involves two or more groups of subjects, knowledge of the risk factors and comorbidities will suggest strata for randomization. A power analysis can determine how many subjects need to be recruited for the study to have a reasonable chance of being able to prove its main hypothesis. If data is available on the annual number of cases presenting at the institution, sample size determined will provide an idea as to how long the study must remain open for enrollment of new subjects or even if it is possible to accrue all subjects from a single institution: sometimes, multiple sites will need to be involved to get sufficient

power. A useful freeware package for power analysis is PS, developed at Vanderbilt University by Dupont and Plummer [13].

Data security considerations should be part of the study plan. Other than the study-specific considerations discussed earlier, the issues of physical security, data backup/archiving, user authentication, audit trails for data changes and user activity, and data locking are not significantly different from those applying to EHRs. An informatics support team should have all these issues worked out in advance.

Informaticians work with investigators and biostatisticians to give them an idea of the extent to which their experimental design can be supported by the software that is currently in use at the institution and what aspects require custom software development. The latter understandably expensive, but even if they were zero for a given study, a CRIS will not run itself. The informatician should therefore provide a cost estimate for the informatics component of the study. In our experience, some naïve clinical investigators greatly underestimate the human resources required for informatics support tasks such as CRF and report design, administrative chores, end-user training, documentation, and help-desk functions. Meeting with the investigator while the idea for the study is still being developed minimizes the risk of underbudgeting. For an informatics team, participation in a study where the members find themselves expending more resources than they are being financially compensated for becomes, in the immortal words of Walt Kelly's Pogo, an insurmountable opportunity.

Electronic protocol design involves the following tasks:

- Setting up the Study Calendar.
- Designing the CRFs for the study (or reusing other CRFs that have been previously created for other studies).
- Designating which CRFs apply to which event on the calendar.
- Designating user roles and the privileges associated with each.
- Specifying the options required for a given experimental design, such as blinding, hiding of PHI.
- Specifying eligibility criteria. (More on this shortly.)
- Identifying the types of reports that will be needed, and designing these, as well as devising a data analysis plan. (More on this later.)
- Determining QC parameters for timeliness and accuracy of CRF entry.
- Creating a manual of operations. CRIS Software typically does not have support for multiple authoring and version control. However, tools such as Adobe RoboHelp™ are more than capable for this task, and they can create the documentation in formats such as HTML (and automatically generating a searchable Web site) as well as generate indexed, searchable help files that can be downloaded and installed on a user's local machine. In addition, one can create context-sensitive help that is accessible from individual CRFs.
- Devising and documenting a data safety monitoring plan (DSMP), which ensures adequate oversight and monitoring of study conduct, to ensure participant safety and study integrity. At the least, the DSMP should include a plan for adverse event reporting (see later) and a Data Safety Monitoring Board if the intervention has the potential of significant risk to the patient.

- Testing the resulting functionality and revising the design until it works correctly. Most CRISs will let you simulate study operation in a test mode using fictitious patients. Once everything works correctly, one can throw a “go live” switch that enables features such as audit trails.
- Role-based user training and certification. Note that this will be an ongoing process as new personnel join the research team.

Recruitment and Eligibility Determination

Most CRIS software will support eligibility determination based on a set of criteria. For simple criteria, they will allow creating questions with Yes/No responses: for a subject to be considered eligible, responses to all inclusion criteria must be Yes, and responses to exclusion criteria must be No. For more complex cases, one can utilize the CRF-design capabilities to design a special “eligibility determination” CRF. Standalone systems also exist: some of these are experimental, e.g., [14], while others, such as the Cancer Center Participant Registry [15], are domain-specific.

The most effective approach to recruitment for subjects with a clinical condition (as opposed to healthy volunteers) involves close integration with the EHR. Information about patients who would meet the broader eligibility criteria (e.g., based on diagnosis codes or laboratory values) can be determined computationally by queries against the EHR data, though other criteria (such as whether the patient is currently pregnant) would have to be ascertained through subject interviews or further tests. Most automation efforts have involved custom, study-specific programming. Though it is possible to build a general-purpose framework that would be study-independent, such a framework would still be specific to a given EHR vendor’s database schema.

When a subject agrees to participate in the study, s/he is given a calendar of visits. As stated earlier, the exact dates may be changed to suit patient convenience: CRIS software may often provide its own scheduler but should ideally be well integrated with an EHR’s scheduling system if the subjects are patients and the hospital (as opposed to a clinical research center) is primarily responsible for providing care.

Robust software generates reminders for both staff and subjects and also allows rescheduling within an event’s window. The period of time prior to a visit date for which changes to the visit date are allowed depend on the nature of the visit: if the visit involves access to a relatively scarce and heavily used resource such as a Positron Emission Tomography scanner, changes to the schedule must be made well in advance.

Protocol Management and Study Conduct

Many of the issues related to recruitment continue through most of the study, since all patients never enroll in the study at the same time. Issues specific to this part of the study include:

- Tracking the overall enrollment status by study group, demographic criteria, and randomization strata.

- Transferring external source data into the CRIS, using electronic rather than manual processes where possible.
- Monitoring and reporting of protocol deviations, which are changes from the originally approved protocol, such as off-schedule visits. Protocol violations are deviations that have not been approved by the IRB. Major violations affect patient safety/rights, or the study's integrity. While protocol deviations related to issues such as major CRF revisions or workflow issues may be prevented simply by the informatics staff resisting changes to the electronic protocol without official approval. Some major violations, such as failure to document informed consent in the CRIS, or enrolling subjects who fail to meet all eligibility criteria, can also be forestalled by the software refusing to proceed with data capture for that patient until these issues are fixed.
- Supporting occasional revisions to the protocol to meet scientific needs. Including CRF modification. (Note that significant protocol revisions require IRB approval.)
- Creating new reports to answer specific scientific questions. (More on this shortly.)
- Monitoring the completeness, timeliness, and accuracy of data entry.
- The workflow around individual events based on the Study Calendar. In addition to reminders to patients to minimize the risk of missed or off-schedule visits, CRISs may also generate a checklist for research staff, e.g., a list of things to do for a given patient based on the event.

Patient Monitoring and Safety

In clinical studies involving therapeutic interventions, monitoring for adverse effects (AEs) is critical. It is not enough to record the mere presence of an AE: its severity in a given patient is also important. For cancer studies, the National Cancer Institute has devised a controlled terminology called the Common Toxicity Criteria for Adverse Events (CTC AE) [16]. Here, the gradation of each concept is specified unambiguously, typically on a 5-point scale for most AEs (5=death). The severity of an AE dictates workflow: in cancer studies, a grade 3 or greater AE must be reported to the sponsor and other collaborating sites as well as to the local IRB. (Failure to do so is a major protocol violation: good CRIS software, by automating the workflow as soon as a grade 3+ AE is detected, prevents such violations.)

An important aspect of CTC AE is that the grades are based on anchored (i.e., objectively defined, often quantitative) criteria that minimize interobserver variability. Therefore, CTC AE has often been used in noncancer studies where AE grading, especially of physical findings and laboratory values, is necessary. CTC AE's use is less appropriate for subjective symptoms or in studies of psychiatric disorders, where the scale lacks sufficient sensitivity and discrimination.

In cases where the study is being conducted in a hospital rather than a clinical research setting, effective interoperability between the EHR software and the CRIS can simplify AE tracking. Some AE data originates from laboratory tests or structured data based on subject interviews/examinations where specific AEs are looked for: here, either the CRIS or the EHR may be the primary system for AE capture: Richesson et al. have devised software that facilitates AE capture and grading and automates the related workflows [17]. In hospital settings, AEs are also recorded in the narrative text

of progress notes. Processing these is much more challenging, but Wang et al. [18] describe an approach for pharmacovigilance based on narrative EHR data.

Analysis and Reporting

Most CRISs implement a variety of standard reports. Among these are:

- Reports related to enrollment of subjects, subcategorized by demographics or randomization strata. Reporting details of subjects screened vs. subjects actually enrolled
- Reports of screened subjects who failed individual eligibility criteria
- Reports related to adverse events
- Reports related to completeness, accuracy, and timeliness of data capture/entry
- Reports summarizing the numbers of patients in different stages of the study (based on events)
- Reports of patients who terminated from the trial abnormally, e.g., because of refusal to continue, adverse events, etc.
- Workflow reports related to the calendar – which patients are due for visits over a forthcoming time interval and what needs to be done for each

In addition, each study will generally require specific, custom-designed reports related to its scientific objectives.

For the purposes of *analysis*, a CRIS must provide bulk-export capabilities, with the data ideally being in a format that is directly acceptable as input by a statistical package. Since the internal data model of CRISs differs significantly from the flat file design that most statistics packages use, the CRISs must perform extensive transformation on their data. Also, in practically all cases, the data sets generated for statistical analysis must be *de-identified*, i.e., the subjects must be identified only by their machine-generated ID without any PHI being eliminated, because these are destined for a data analyst who does not need to know the PHI. By contrast, most reports related to workflow, as well as many study-specific reports, which are used by research staff who are in direct contact with their subjects, will contain PHI, especially because clinician decisions may be made on the basis of the reports' contents, and it is important to identify each subject accurately.

Miscellaneous Issues

Validation and Certification

CRISs are often used to make clinical decisions: therefore, defects should be minimized. We know of now-defunct CRIS software that in the 1990s that was priced at around \$3 million and crashed several times a day with a “blue screen.”

Certification of CRISs has been proposed in a manner similar to that used by the Certification Commission for Hospital Information Technology (CCHIT) for EHRs. As many EHR customers have learned painfully, however, CCHIT certification does not actually mean that the software will meet an organization's needs, or even that it will be usable. The criteria for CRIS certification may be based on whether a CRIS has particular features or not – but if the implementation of individual features is clumsy, use of those features will be nonintuitive and error-prone.

A detailed testing plan is obviously important in helping to establish a CRIS as a robust product. However, as Kaner, Falk, and Nguyen's classic "Testing Computer Software" [19] emphasizes, the absence of detected errors does not prove conclusively the absence of defects. Also, software that fully meets its specifications on testing is not defect-free if the specification itself was incomplete or flawed. Further, CRISs are built on top of existing operating systems, commercial database engines, transaction managers, and communications technology. Defects in any of these – is any user of Microsoft Windows unaware of periodic discoveries of bugs and vulnerabilities? – could affect their operation.

Finally, even if a CRIS is itself defect-free, flawed implementations at a particular institution by an insufficiently trained or knowledgeable CRIS support team may cause major usability problems. For example, CRF design is essentially a kind of high-level programming, typically using a GUI so that nonprogrammers can accomplish most tasks. Errors of both commission – e.g., a mistake in a formula for a computed field – and omission, e.g., forgetting to add sufficient validation checks so that bad data creeps in, will cause problems.

The point we are trying to make is that there are no simple solutions to the matter of system validation and certification.

Standards

Lack of standards has been one limiting factor in CRISs: as in several other areas of computing, they result in an uncomfortably tight dependency of a customer on a given vendor. Several chapters of this book deal with the issue of standards in greater detail, so we will just give you our take on data-library standards.

There are efforts toward standardizing the contents of data libraries, such as by the Clinical Data Interchange Standards Consortium (CDISC). However, data libraries are where individual CRIS vendors differentiate themselves the most, especially for complex validation (but in highly incompatible ways), and CDISC makes no attempt to represent complex validation rules. Even if it eventually did, we doubt that it would have significant impact: vendors have no compelling reason to change (which would require overhauling their infrastructure completely). The fact is that complex validation in CRISs is not easy to implement in a manner that is readily learnable by nonprogrammers. It is harder still to represent in a metadata interchange model.

Pragmatic Clinical Trials: Use of EHRs Instead of CRISs

While we have focused on the use of CRISs, there is one situation where EHRs are used instead for primary data capture. “Pragmatic trials” differ from traditional clinical trials in that the conditions of the trial are more lax than in traditional controlled clinical trials, which are termed “explanatory.” In pragmatic trials, established medications already employed in clinical practice are the interventions under study rather than investigational drugs, with the intervention being performed by the clinicians who would normally see the subjects/patients as part of their job, rather than specially designated research personnel: the subjects are never completely healthy volunteers.

The motivation of a pragmatic trial is to study interventions (typically medications) as used in actual practice – under imperfect research conditions – rather than in the ideal but highly constrained situations that, for example, employ double-blind designs.

The best overview of pragmatic trials that we have seen is by Patropoulous [20], who describes the difference between pragmatic and explanatory trials in multiple areas:

1. *Patient selection criteria:* A far wider variety of patients is studied.
2. *Personnel:* The clinicians involved in caregiving may lack research expertise.
3. *Comparison-therapy choice flexibility:* Unless the primary goal is to compare two medications, the comparison treatment may simply be the best available alternative treatment.
4. *Therapeutic flexibility:* Clinicians have full freedom to adjust the medication dose for both the intervention medication and comparison medication/s – unless study of one dose versus another is the study’s primary goal – and administer other medications as needed.
5. *Follow-up intensity:* A rigid patient calendar may not be enforced: only a final time-point is critical.
6. *Outcome:* The primary outcome is one that is clinically meaningful and can be assessed readily by non-researcher clinicians.
7. *Subject compliance:* Rigorous monitoring – e.g., pill-counting dispensers or chemical surrogates (e.g., riboflavin combined with the medication) – is not enforced, since in real life patients may miss occasional doses.
8. *Practitioner adherence:* Since participating caregiving clinicians are compensated minimally if at all (to reduce costs), there is little or no pressure on them to adhere strictly to the study protocol.
9. *Analysis:* All participants who are intended to be treated may be included.

No pragmatic trial can be completely slack with respect to all of the above criteria, and typically one or more constraints may be enforced; thus, most pragmatic trials actually fall on a continuum between pragmatic and explanatory. In any case, the relatively lax conditions in which pragmatic trials are performed mean that pragmatic trials have less *internal validity* than explanatory trials. That is, because all the other variables that can influence outcome, such as concurrent medications, are

not controlled rigorously, an inference that the intervention under study was actually responsible for all or most the observed outcome/s is more dubious. However, *external validity* is greater – i.e., the trial’s results are more likely to be generalizable across more healthcare settings and across more populations. (Note: greater external validity is not guaranteed: in some multicentric pragmatic trials, for example, outcomes have varied very greatly across individual sites.)

The reason for using EHRs, with all their limitations, rather than CRISs is that, given caregiving clinicians’ minimal reimbursement, their workflow must change as little from normal clinical practice as possible (or not at all). Forcing them to learn CRIS software or fill in specially designed paper forms would increase their workload unacceptably and bring on mass revolt. The destination for collated data across multiple sites will typically be a custom database, where it is cleansed prior to export to a statistical package.

Concluding Remarks

An important aspect of evaluation of a CRIS that is a candidate for purchase is its usability. A CRIS is a complex piece of software, and it will understandably have a significant learning curve simply because a lot of its functionality must be learned in order to set up an electronic protocol correctly. A less than intuitive user interface can greatly compound the difficulty in learning it. Such software should ideally follow the principles of user-centered design [21], which is a fancy way of describing a design process that emphasizes the perspectives, needs, and the limitations of the intended users of the software. The reality, however, is that the CRIS software market is simply not as competitive as that of mass-produced microcomputer software, and so one may often find that the user (and organizational processes) must adapt to the software rather than vice versa. A similar situation prevails in EHRs – after spending tens of millions of dollars, the vendor controls users’ fates rather than vice versa.

With competitive pressure due to the entry of open-source CRIS software, this situation may change for the better. However, it is important to ensure that the software one is considering is a good fit for one’s needs, and some forward thinking is necessary: it should not only be a good fit for the studies one is conducting presently but also for studies one may conduct in future.

References

1. Eisenstein EL, Collins R, Cracknell BS, Podesta O, Reid ED, Sandercock P, Shakhov Y, Terrin ML, Sellers MA, Califf RM, Granger CB, Diaz R. Sensible approaches for reducing clinical trial costs. *Clin Trials*. 2008;5(1):75–84.
2. Frank E, Cassano GB, Rucci P, Fagiolini A, Maggi L, Kraemer HC, Kupfer DJ, Pollock B, Bies R, Nimgaonkar V, Pilkonis P, Shear MK, Thompson WK, Grochocinski VJ, Scocco P, Buttenfield J, Forgione RN. Addressing the challenges of a cross-national investigation: lessons from the Pittsburgh-Pisa study of treatment-relevant phenotypes of unipolar depression. *Clin Trials*. 2008;5(3):253–61.

3. Joint Commission on Accreditation of Hospital Organizations. National Patient Safety Goals. Available at http://www.jointcommission.org/PatientSafety/NationalPatientSafetyGoals/08_hap_npsgs.htm. Last accessed 12/03/09.
4. Crane D, Pascarello E, James D. AJAX in action. Greenwich: Manning Publications Co; 2005.
5. Van den Broeck J, Mackay M, Mpontshane N, Kany Kany Luabeya A, Chhagan M, Bennish ML. Maintaining data integrity in a rural clinical trial. *Clin Trials*. 2007;4(5):572–82.
6. Thwin SS, Clough-Gorr KM, McCarty MC, Lash TL, Alford SH, Buist DS, Enger SM, Field TS, Frost F, Wei F, Silliman RA. Automated inter-rater reliability assessment and electronic data collection in a multi-center breast cancer study. *BMC Med Res Methodol*. 2007;18(7):23.
7. Wikipedia. Computerized Adaptive Testing. Available at: en.wikipedia.org/wiki/Computerized_adaptive_testing. Last accessed 12/1/09.
8. Celli D, Yount S, Rothrock N, Gershon R, Cook K, Reeve B, Ader D, Fries JF, Bruce B, Rose M. The patient-reported outcomes measurement information system (PROMIS): progress of an NIH roadmap cooperative group during its first two years. *Med Care*. 2007;45(5):S3–S11.
9. Day S, Fayers P, Harvey D. Double data entry: what value, what price? *Contemp Clin Trials*. 1998;19(1):15–24.
10. Wyatt JC, Altman DG, Heathfield HA, Pantin CF. Development of design-a-trial, a knowledge-based critiquing system for authors of clinical trial protocols. *Comput Methods Prog Biomed*. 1994;43(3–4):283–91.
11. Modgil S, Hammond P. LinksDecision support tools for clinical trial design. *Artif Intell Med*. 2003;27(2):181–200.
12. Rubin DL, Gennari J, Musen MA. LinksKnowledge: representation and tool support for critiquing clinical trial protocols. *Proc AMIA Symp*. 2000:724–8.
13. Dupont WD, Plummer WD. Power and sample size calculations: a review and computer program. *Control Clin Trials*. 1990;11:116–28.
14. Gennari JH, Sklar D, Silva J. LinksCross-tool communication: from protocol authoring to eligibility determination. *Proc AMIA Symp*. 2001:199–203.
15. National Cancer Institute. Cancer center participant registry. Information available at: <https://cabig.nci.nih.gov/tools/c3pr>. Last accessed 12/01/09.
16. National Cancer Institute. Common Terminology Criteria for Adverse Events (CTCAE) and Common Toxicity Criteria (CTC). 2009 Available from: http://ctep.cancer.gov/protocolDevelopment/electronic_applications/ctc.htm. Last accessed 1/2/18.
17. Richesson RL, Malloy JF, Paulus K, Cuthbertson D, Krischer JP. An automated standardized system for managing adverse events in clinical research networks. *Drug Saf*. 2008;31(10):807–22.
18. Wang X, Hripcak G, Markatou M, Friedman C. Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: a feasibility study. *J Am Med Inform Assoc*. 2009;16:328–37.
19. Kaner C, Falk J, Nguyen HQ. Testing computer software. 2nd ed. New York: Wiley; 1999.
20. Patsopoulos NA. A pragmatic view on pragmatic trials. *Dialogues Clin Neurosci*. 13(2):217–24. PMC ID: PMC3181997.
21. Wikipedia. User-centered design. Available at http://en.wikipedia.org/wiki/user_centered_design. Last accessed: 1/2/18.



Study Protocol Representation

10

Joyce C. Niland and Julie Hom

Abstract

Clinical research is an extremely complex process involving multiple stakeholders, regulatory frameworks, and environments. The core essence of a clinical study is the *study protocol*, an abstract concept that comprises a study's investigational plan—including the actions, measurements, and analyses to be undertaken. The “planned study protocol” drives key scientific and biomedical activities during study execution and analysis. The “executed study protocol” represents the activities that actually took place in the study, often differing from the planned protocol, and is the proper context for interpreting final study results. To date, clinical research informatics (CRI) has primarily focused on facilitating electronic sharing of text-based study protocol documents. A much more powerful approach is to instantiate and share the abstract protocol information as a computable protocol model, or *e-protocol*, which will yield numerous potential benefits. At the design stage, the *e-protocol* would facilitate simulations to optimize study characteristics and could guide investigators to use standardized data elements and case report forms (CRFs). At the execution stage, the *e-protocol* could create human-readable text documents; facilitate patient recruitment processes; promote timely, complete, and accurate CRFs; and enhance decision support to minimize protocol deviations. During the analysis stage, the *e-protocol* could drive appropriate statistical techniques and results reporting and support proper cross-study data synthesis and interpretation. With the average clinical trial costing millions of dollars, such increased efficiency in the design and execution of clinical research is critical. Our vision for achieving these major CRI advances through a computable study protocol is described in this chapter.

J. C. Niland, PhD (✉) · J. Hom, BS

Department of Diabetes and Cancer Discovery Science, City of Hope, Duarte, CA, USA

e-mail: jniland@coh.org; jhom@coh.org

Keywords

Clinical research informatics · Study protocol · E-protocol · Case report form
Executed study protocol · Computable study protocol · Web ontology language
Unified Modeling Language

Overview

The Study Protocol: Core Essence of a Clinical Research Study

A clinical research study is a planned investigation in which a series of prespecified actions are carried out on study participants, their data, and/or their biospecimens, to collect information that can be analyzed to increase our understanding of human health and disease. The study's investigational plan—including the actions to be undertaken, the measurements, and the analysis procedures to be followed—is called the *study protocol*. The study protocol is an abstract concept, which manifests as two related states during the life cycle of the research study. The “planned study protocol” is the core essence of any clinical research study, representing the study's conceptual scientific structure. In proposed and ongoing studies, the planned protocol drives the key scientific and biomedical activities that take place during study execution and analysis. In completed studies, the “executed study protocol” represents the activities that actually took place, which often may differ from the study's planned protocol. Understanding and documenting the executed protocol is vital for interpreting the final study results. In any case, the study protocol is the single most valuable and distinguishing assembly of information to define a clinical research study.

In common usage, the term *study protocol* often conflates the abstract notion of the planned research, as described above, with the textual documents that traditionally describe the abstract protocol. That is, investigators write study protocol *documents* in text, not study protocols. Study protocol documents are artifacts generated to guide the conduct of clinical research during the course of a study. These *documents* are filed for human subjects approval applications, and sponsors sometimes post study protocol *documents* on the web. While they do describe a study's planned activities with varying accuracy and completeness, study protocol documents are not the core essence of a study's scientific structure in the way that the abstract study protocol is.

Table 10.1 defines several key terms used in this chapter. We distinguish between the study protocol, which is the abstract investigational plan for a study, and the *computable protocol model*, which is a generic computable representation of the abstract elements and decision rules commonly found in study protocols. There already exist multiple computable protocol models of various depth and complexity, as reviewed below. Standardization of the underlying computable protocol model into a *common computable protocol model* has been a “holy grail” of clinical research informatics (CRI) for many years, to facilitate the interoperation of

Table 10.1 Clinical research informatics terms relating to study protocol

Study protocol	The abstract specification of a study's investigational plan, including the actions to be undertaken, the variables to be assessed, and the analysis procedures to be followed
Study protocol document	A textual description of the study protocol, often in the form of a PDF or other document format, such as MS Word
Computable protocol model	A generic computable representation of the information contained within a clinical research study protocol
Common computable protocol model	A <i>shared and standardized</i> computable representation of study protocols that serves as a reference semantic across all clinical research studies
E-protocol	An instantiation of an individual study plan in a specific singular computable protocol model ("e-protocol") or ideally going forward, in the common computable protocol model ("e-protocol")

computable protocols across disparate systems for advanced information and knowledge management in clinical research. Efforts are ongoing to establish such an agreed upon common computable protocol model, as described below. These efforts are likely to be an active area of CRI for years to come.

When a specific study's protocol is instantiated in a computable protocol model, we introduce here a new term for this representation, an *e-protocol*. The e-protocol is defined as an instantiation of an individual study plan as an electronic computable protocol representation, based on a specific singular computable protocol model. Ideally going forward, once a commonly defined and accepted computable protocol model is in place, we propose that a study plan that utilizes this common model will be so designated by the term *e-protocol*.

The Study Protocol Enabled by Clinical Research Informatics

Clinical research is an extremely complex process involving multiple stakeholders acting within a number of regulatory frameworks. Study management in such a complex environment necessitates the sharing of information generally represented to date as documentary artifacts. Given the current state of CRI tools, the focus has been on facilitating the electronic sharing of such study protocol documents. However, because study protocol documents are only derivatives representing the abstract study protocol, this focus on document management is narrower and less powerful than direct information management of the abstract study protocol and its metadata.

To advance CRI tools, the abstract study protocol must be made directly computable, without the intermediary of textual descriptions in the form of protocol documents. Rather, it should be the other way around, that is, sharable protocol documents should be generated via creation of a computable study protocol.

With the average cost of commercial clinical trials being in the millions, efficiency in the design and execution of clinical research is not a luxury. There is an increasingly urgent and outstanding opportunity to apply the power of computers

beyond clinical research document management, to provide true information management in full-spectrum support of the design, execution, analysis, and reporting of clinical research. Such computable study protocols would yield many benefits throughout the clinical study life cycle. For example, at the design stage, a computable protocol would facilitate conducting simulations of varying design characteristics to help an investigator iteratively optimize the design to lower study duration and costs. User interfaces that help investigators capture study plans as computable protocols also afford the opportunity for ensuring standardized data elements and case report forms (CRFs). At the execution stage, as has been shown in clinical research management systems, the computable protocol can be used to create human-readable text and paper documents; facilitate distributed patient recruitment processes; monitor patient recruitment against desired sample size to avoid over- or under accrual; provide timely, complete, and more accurate CRFs for greater study quality assurance; and drive decision support to help minimize protocol deviations such as ineligible patients, missed visits, or inappropriate doses [1]. During the analysis stage, the computable protocol can drive the use of appropriate statistical analytic techniques and computable reporting of results [2].

The protocol elements and rules that need to be computable to create a functional e-protocol are described in the next section, followed by several examples of use cases for which the e-protocol will offer substantial benefits. Further benefits would accrue if e-protocols could easily be instantiated across multiple systems, as the average multicenter clinical trial typically now enrolls thousands of patients from over 20 participating sites. The current status of efforts to standardize major elements of computable protocol models will be presented later in the chapter.

Current Inefficiencies in Study Protocol Informatics

One of the greatest sources of inefficiency in instantiating the computable protocol is the lack of well-accepted and adopted standards. The typical clinical trial protocol document contains many implied meanings and unclear instructions, often leading to misinterpretations, errors, and inconsistencies in trial conduct. This issue becomes especially critical when companies and protocol sponsors ascribe different meanings to the same term. Standardized encoding of data capture will help to improve clinical trial capabilities to drive operational efficiency and allow centers to mount multisite studies much more rapidly and efficiently.

Among existing clinical research databases and systems, most have developed independently with tremendous variability in nomenclature, data content, and analytical tools, leading to silos that impede efficient solutions even as clinical research information systems, rules, processes, and vocabularies are becoming increasingly interdependent over time. In short, there is no unifying architecture to support the desired interoperability and enforce the technological and lexical standards upon which these systems depend. The structured protocol created via the e-protocol can serve as the semantic CRI foundation that adds value through improved clarity and

communication. While standardization can confer benefits even in a paper-based world, these benefits are leveraged and magnified in a computer-assisted clinical research environment.

For biomedical data to be effectively exchanged, integrated, and analyzed, the need for standardization must be addressed first. Although establishing a common structured protocol representation is a formidable task, it is prudent to address this problem before it becomes even more intractable and costly to solve, in the face of tens of thousands of human studies going on worldwide at any one time. The application of computational and semantic standards is essential for information integration, system interoperability, workgroup collaboration, and the overall exploitation of significant prior investments in biomedical information resources.

Currently, there is substantial redundancy of data collection, entry, and storage throughout clinical research institutions, overlapping with processes and data in the clinical care arena. The lack of data sharing and integration across systems is exacerbated by the absence of universally adopted clinical research standards. Vocabularies differ, and there has been no clear emergence of a complete clinical research semantic system. Further, the discipline is lacking in the comprehensive metadata required to appropriately address and resolve these issues. The standards embedded within the e-protocol will enforce such unifying approaches to enable rapid design of protocols, mounting of multicenter initiatives, and integration and interpretation across studies to speed discoveries.

Figure 10.1 illustrates that a single clinical study involves many different people and many different information systems over its life cycle. To take advantage of the computable e-protocol of the study, each system will need to interface with the underlying computable protocol model in which the e-protocol is instantiated. Clearly, clinical research informatics would be well served if there was a *common* computable protocol model in which all e-protocols are instantiated, so that systems would not have to build separate interfaces to a multitude of protocol models. Given that the average clinical trial is conducted in 23 different sites, each possibly using local configurations of clinical trial management and other systems, substantial resources will be required for protocol and data integration in order to provide decision support across the life cycle across multiple sites and multiple systems for a

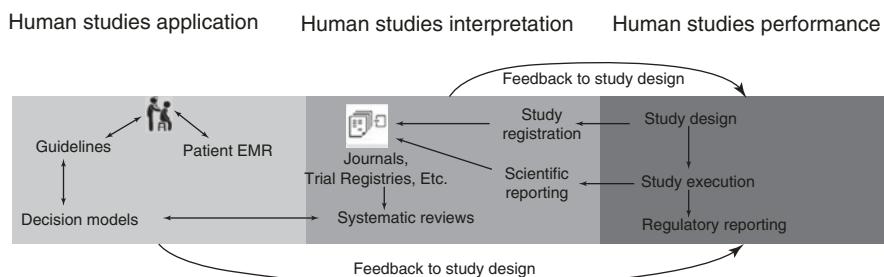


Fig. 10.1 Life cycle of human studies

single study. A common computable protocol model can virtually eliminate that resource overhead and has therefore been a “holy grail” of clinical research informatics. The next section of this chapter highlights the necessary elements and beneficial use cases for the computable study protocol.

Elements and Use Cases for a Computable Study Protocol

Most clinical researchers are intimately familiar with study protocol documents, which may be paper-based or completely electronic (e.g., PDF). These documents are used for a multitude of tasks, ranging from obtaining funding to securing human subjects approval and to guiding study execution. The documents vary greatly in length and content but generally should include detailed background rationale and objectives; carefully stated scientific hypotheses; clear and complete eligibility criteria; well-specified outcomes, measurements, data collection, and variables; and robust statistical design and analysis plans.

Despite the importance of their content, far too often protocol documents include only cursory descriptions of the study population and primary variables. There are no broadly accepted standards for the contents of protocol documents at the design stage, although at least one has been proposed [3]. The International Conference on Harmonization E3 standard, while important, is meant for a different audience and purpose, as it applies to describing the executed protocols of completed studies, rather than planned protocol documents created before study initiation.

The major elements of e-protocols overlap with the elements contained within study protocol documents but are of necessity broader reaching and more standardized. While study protocol documents are for human use, e-protocols are for supporting computational approaches to data structure and organization, information management, and knowledge discovery. Thus, to support a broad range of clinical research use cases, e-protocols must satisfy both domain modeling (content requirements) as well as requirements for computability. By considering what is required of the e-protocol to meet particular use cases, we illuminate the abstract common requirements for more generic computable protocol models.

The computable study protocol that will be enabled through the e-protocol will confer numerous benefits and eliminate many of the inefficiencies that exist today due to the usage of paper protocol documents and a “mishmash” of CRI systems to guide study conduct. Content requirements for the e-protocol are dictated by the ultimate functionality to be supported. The e-protocol’s purpose is to (1) capture the complete study plan in computable form, (2) provide decision support during study conduct, (3) facilitate timely and accurate data capture and storage, (4) support appropriate statistical analysis and reporting, (5) support appropriate interpretation and application of results, (6) facilitate reuse of study data and artifacts (e.g., biosamples), and (7) allow comparisons and metanalyses across studies of the same interventions for common indications. Out of scope for the e-protocol content requirements will be the tracking of the scientific and regulatory review and approval processes. However, amendments to the study protocol content will of necessity and

naturally be captured as a self-documenting audit trail within the e-protocol. The minimal content requirements for each area of desired functionality are described in the following sections.

Capturing the Complete Study Plan in Computable Form

A first step toward computable study plans is to capture the complete study plan in electronic, if not necessarily computable, form. Absent widely accepted guidelines on study protocol contents, Table 10.2 provides a typical table of contents that we will use to discuss the protocol data elements necessary to facilitate all further functionality. Complete capture of this content in e-text will allow the rendering of the study protocol in human-readable form(s), such as PDF or MS Word documents that humans will always need to conduct studies. However, capture of this content as fully coded machine-readable standardized data elements is ideal and will enable much richer and more powerful decision support and enhanced workflow functionality. Based on today's state of the computable study protocol, we also suggest in

Table 10.2 Example table of contents and data formats for a clinical research e-protocol^a

Study protocol content	Data format
Study objectives	Text-based, possibly templated
Background	Text-based, possibly templated
Hypotheses	Text-based, possibly templated
Patient eligibility	Coded core eligibility criteria to enable patient-protocol filtering (e.g., per ASPIRE standards) and fully coded complete eligibility criteria (e.g., per ERGO)
Study design	Coded data elements per emerging standards (e.g., TrialDesign component of CDISC model or OCRe)
Sample size	Coded enrollment numbers, per arm
Registration guidelines	Text-based, possibly templated
Recruitment and retention	Templated (e.g., CONSORT flowchart)
Intervention description	Templated, for different types of interventions (e.g., RxNorm codes for drug names, model numbers for devices)
Intervention plan	Text-based, possibly templated
Adverse event (AE) management	Coded data for AE terms reporting intervals, regulatory agencies
Outcome definitions	Coded baseline, primary, and secondary outcome variables and coding
Covariates	Coded main covariates (e.g., stratification variables, adjustment factors)
Statistical analyses	Coded data and algorithms per emerging standards (e.g., StatPlan component of CDISC model)
Data submission schedule	Coded data submission intervals

^aThese data elements are meant to be illustrative, not exhaustive

Table 10.2 the data formats that are currently realistic for the electronic e-protocol, even if the e-protocol is not yet fully computable.

As work progresses on the computable model and related rule sets (mostly within the Biomedical Research Integrated Domain Group [BRIDG] model activities, mentioned in Chap. 18 and described below), more discrete data elements will be captured for each content category in ever more structured and coded format. The definition, modeling, and standardization of these more discrete data elements are being driven by the work to support the following e-protocol functionalities.

Providing Decision Support During Study Conduct

Modern clinical research protocols can be very complex, arguably too complex to be generalizable to daily clinical care [4]. As a result, study coordinators and front-line staff have many complex protocol rules to follow (e.g., who to enroll, when to assess outcomes, and how/when to grade and report AEs). Because standardized study processes can increase the internal validity of studies, decision support to regularize study conduct serves scientific as well as regulatory goals. Broadly speaking, the constructs that need to be computable to support this functionality include (1) eligibility criteria, (2) decision rules for triggering specific study actions (e.g., AE reporting), and (3) participant-level and study data referenced by eligibility criteria and decision rules. The following sections discuss the representation of eligibility criteria and the requirements for achieving computability, focusing on the content requirements for criteria, rules, and clinical data.

As clinical research studies cover the entire range of health and disease, the broad answer to the question “what are the content requirements for study protocol decision support?” is “all of medicine.” The need for robust standardized representations for all medical concepts is as much a challenge for CRI as it has been a challenge for health informatics over many decades, requiring the exchange and use of knowledge from multiple domains. Several controlled terminologies may be used for subdomains in medicine (e.g., RxNorm for drugs; see Table 10.2); however there should be no bounds on the permissible domain content for e-protocols. Indeed, clinical research studies often require content from outside of medicine, for example, eligibility criteria that require residence within a certain county or decision rules in health services research studies that are triggered by changes in patient insurance status. Clearly, the scope of decision support will be driven by the domain coverage of the clinical data that are coded and formally represented in e-protocols.

Another category of content requirement for decision support is semantic relationships between multiple encoded concepts. Thus, an inclusion criterion for patients with renal failure *due to* diabetes is semantically different from one that includes patients with renal failure *coexisting with* but not necessarily due to diabetes. In other words, a decision support system that attempts to fully determine whether a particular patient satisfies the first criterion above needs to have access to standardized data elements for renal failure, diabetes, and the causal relationship between them.

The representation of semantic relations is currently very rudimentary. The first version of the open biomedical ontologies (OBO) relations ontology details ten relations: two foundational ones (*is_a*, *part_of*) and other physical (e.g., *located_in*), temporal (e.g., *preceded_by*), and participant (e.g., *has_participant*) relations [5]. In accordance with the underlying OBO philosophy, the relations ontology includes only “relations that obtain between entities in reality, independently of our ways of gaining knowledge about such entities,” which would exclude many clinical relevant relations such as “*due_to*.” The Unified Medical Language System (UMLS) has approximately 100 semantic relations, but without a formal structure, it is impossible to fully reason across semantic relations themselves (e.g., “*that are due_to*” and “*caused_by*” are similar, but this similarity is not fully represented). This lack of formal structure in turn limits opportunities to fully reason across protocols and protocol content encoded in this way. Better decision support for clinical research awaits additional advances in the representation and codification of clinically relevant semantic relations.

Facilitating Timely and Accurate Data Capture and Storage

When fully and appropriately executed, the e-protocol will greatly enhance the ability to capture and store data in an accurate, complete, and timely manner. Electronic CRFs should be designed such that the metadata, including user definitions and allowable code lists for each field, are encoded within the e-protocol. The ability to export the metadata from the system should be in place, for integration within a metadata repository, facilitating the ability to draw upon this repository to create standard data elements. Ideally in the future, CRI tools will evolve in the future such that the “forms metadata” would also include the ordering, labeling, and placement of the data elements within the electronic CRFs. These forms could then be automatically generated via the system. Embedding the technical metadata into the e-protocol could facilitate the design and creation of the data storage tables as well.

In the e-protocol, metadata describing the specifications for data capture should include the core and full eligibility criteria, treatments received, treatment deviations, routine monitoring results for subject health status, AEs, primary and secondary endpoint measurements, and any covariates or adjustment factors for the analysis. Efforts at standardized data elements for CRFs are underway and will greatly improve and speed the process of creating CRFs within electronic data capture systems, as documented through the e-protocol [6, 7]. The data model could be exported to electronic data capture (EDC) systems to automatically instantiate the fields and constructs needed to collect study data in the EDC as the research progresses.

Currently, uneven data quality frequently limits the effectiveness and efficiency of clinical trials execution. Improved data quality will be enhanced through programmatic data validations that can be specified in the e-protocol prior to initiation of data collection. Ideally, such validations also could be exported to EDC tools, to automatically program up-front data validations into the system. Global data

element libraries will allow for reuse in study development, resulting in more rapid study implementation. This process also will reduce the complexity and thereby facilitate within study or cross-study data analysis and integration by eliminating data “silos.”

Supporting Appropriate Statistical Analysis and Reporting

The goal of conducting clinical research studies is to collect unbiased data that can be analyzed to inform our understanding of health and disease. If inappropriate analytic methods are used, the findings will be uninformative or worse, misleading. E-protocols can mitigate these problems by enforcing clear definitions of study variables and their data types: for example, diabetes as a dichotomous variable ($\text{HbA1c} \geq 6.5\%$) should be analyzed using different statistical methods than diabetes as a continuous variable of HbA1c level.

The appropriate statistical tests to use depend on the data type of the independent and dependent variables. In turn, the data types and statistical tests used determine what aspects of the results should be reported (e.g., p value, beta coefficient) to maximally inform the scientific community of the study’s findings. Therefore, the content of e-protocols needed to support statistical analysis and reporting includes a clear definition of study variables (e.g., the primary outcome) and their data types, the relationship of raw data to these variables (e.g., censored, aggregated), a clear specification of the study analyses (e.g., models to be created, covariates to be included), and the role of individual variables as independent or dependent variables within specific study analyses. The definition of these elements and their interrelationships are defined in the Ontology of Clinical Research [8].

Supporting Appropriate Interpretation and Application of Results

One of the tenets of evidence-based medicine is that study results must be interpreted in light of how the data were collected. Thus, generations of students have learned the principles of critical appraisal and the hierarchy of evidence (e.g., that randomized controlled trials provide less internally biased results than observational studies). Readers of journal articles are exhorted to consider all manner of design and study execution features that might affect the reliability of the study results (e.g., Was allocation concealed? Were the intervention groups similar in baseline characteristics? Was there disproportionate lack of follow-up in one arm?). For computers to support results interpretation, the e-protocol representing the executed (not the planned) protocol must contain the data elements required for critical appraisal. Sim et al. identified 136 unique study elements required for critically appraising randomized controlled trials [9]. Comparable data elements are required for critically appraising observational and nonrandomized interventional studies. These data elements are modeled in the Ontology of Clinical Research (OCRe),

which was designed to support study interpretation and methodologically rigorous synthesis of results across multiple studies [8].

Facilitating Reuse of Study Data and Artifacts

The same design and execution elements needed for critical appraisal also are needed to properly reuse study data or biospecimens. For example, data from a trial enrolling only patients with advanced breast cancer will not be representative of breast cancer patients in general, and this must be recognized in any data reuse. Studies may even include subjects who do not have the condition of interest, for example, a study with a nonspecific case definition or a study with healthy volunteers. While sharing patient-level data from human studies would help investigators make more and better discoveries more quickly and with less duplication, this sharing must be done with equal attention to sharing study design and results data, preferably via computable e-protocols. Sharing of biospecimens will be facilitated through encoding of the type, quantity, processing, and other specific characteristics of the specimens to be collected during the conduct of the study.

Computability and Standardization Features and Requirements

The ability to reuse protocol elements across different studies requires standardized, formal representation of the “parts” of a protocol (see the constructs in Table 10.2). For standardizing the representations, bindings to appropriate clinical vocabularies are critical but not sufficient. There needs to be agreement on the conceptual elements in each construct as well as the specific codings that should be used. For example, for endpoint definitions, how exactly are primary endpoints different from secondary endpoints? Investigators sometimes change these designations over the course of a study for various reasons. The representational challenges here are reminiscent of those that have plagued clinical data representation and exchange in the electronic health record (her) context—clinical terminologies offer standardized value sets, but the meaning of the data field itself needs standardization for computability.

The e-protocol could be represented using a number of representational formalisms, with Unified Modeling Language (UML) and Ontology Web Language (OWL) being the dominant choices. OWL provides mechanisms that tend to encourage cleaner semantics, while UML has the practical benefit of coupling modeling to software development. E-protocol models do not have to be rendered in either UML or OWL but could utilize both. The BRIDG model is now both in UML and OWL, as is the Ontology of Clinical Research (OCRe). The Ontology for Biomedical Investigations project also defines, in OWL, entities relevant to e-protocols [10]. The achievement of a single unified model in corresponding OWL and UML forms across the breadth of clinical research is challenging but remains the holy grail of CRI. A critical gap is easy-to-use and widely accessible tools that allow distributed editing and harmonization of conceptual models expressed in various formalisms.

Protocol Representation Standards

In this section we summarize many of the protocol representation standard activities that are underway to enable the common computable study protocol.

Standards for Protocol Documents

HL7 Regulated Clinical Research Information Model Protocol Representation Group

Health Level 7 (HL7) is the preferred electronic exchange format for healthcare information, per the Department of Health and Human Services [11]. HL7 is a not-for-profit, American National Standards Institute (ANSI)-accredited standards development organization dedicated to providing a comprehensive framework and related standards for the exchange, integration, sharing, and retrieval of electronic health information. The HL7 exchange format is already used for several FDA messages, including the Structured Product Label (SPL), the Integrated Case Safety Report (ICSR), and the Regulated Product Submission (RPS) messages. HL7 messages also are the preferred exchange format for clinical observations captured within EMR systems, which will enable the integration and reuse of clinical care data within the e-protocol.

HL7 V3.0 is based on HL7's Reference Information Model (RIM), which incorporates standards for communications that document and manage patient care, as a fundamental part of the technologies needed to meet the global challenge of integrating healthcare information [12]. The RIM allows exchange of information on clinical care processes via *technical* interoperability. HL7 V3.0, as a standardized model to represent healthcare information, will yield *semantic* interoperability, based on consensus ballots worldwide. A subgroup called the HL7 Regulated Clinical Research Information Model (RCRIM) is working to utilize the RIM to evolve a standardized model for research, which would facilitate the creation and adoption of the e-protocol format.

The Clinical Data Interchange Standards Consortium Protocol Representation Group

The Clinical Data Interchange Standards Consortium (CDISC) was formed in 1997 through a collaboration of biopharmaceutical, regulatory, academic, and technology partners with a goal toward optimizing clinical research through the creation and adoption of standards [13]. The CDISC Protocol Representation Group (PRG) has developed the CDISC Protocol Representation Model (PRM) V1 to facilitate the exchange of clinical research data, allowing studies to be initiated more rapidly and supporting machine- and human-understandable decision support. Recognizing that the study protocol lies at the heart of clinical research, the primary goal of the PRG is to develop a standard interoperable protocol. The mission statement of the group is: "Development and publication of standard, structured, protocol concepts and content that will enable interchange and reuse

of protocol-required data and metadata among systems, stakeholders, and operations staff throughout the lifecycle of the study.”

The CDISC PRM must be both machine-readable and easily understood by its audience, namely, regulatory authorities, statisticians, data managers, and medical researchers. The PRM has the potential to add great value to the efficiency of clinical study conduct, diminishing time to author new protocols, improving the quality of study conduct through enhanced clarity and consistency of protocol information, and facilitating multicenter data exchange. However, as with other representation models, the full value of the PRM will not be realized unless it receives widespread acceptance and adoption across the stakeholder spectrum. The current directions for the PRG effort are to (a) leverage standards that had matured since the initiation of the PRM project, (b) align with the BRIDG model (see below) that had been initiated to harmonize CDISC standards, and (c) focus on an initial set of representative priority use cases out of the many that involve the clinical research protocol.

Furthermore, the CDISC Operational Data Model (ODM) standard has shown remarkable flexibility and extension to a broad range of use cases, data, and metadata, as the demand for interoperability has increased [14]. A classification schema of the use of ODM within the clinical research data life cycle shows that it has been extended to EDC and HER infrastructure, study planning, data collection, data tabulation and analysis, and study archival.

The Standard Protocols Items for Randomized Trials Initiative

The Standard Protocols Items for Randomized Trials (SPIRIT) initiative is defining an evidence-based checklist that defines the key items to be addressed in trial protocols, leading to improved quality of protocols and enabling accurate interpretation of trial results [3]. The SPIRIT group’s methodology is rigorous and similar to that of the CONSORT group that defines trial reporting standards [15]. The SPIRIT recommendations come from the academic epidemiology and evidence-based medicine community, not from clinical research informatics, and should complement the protocol document standards discussed above.

Standards for Protocol Model Representation

Biomedical Research Integrated Domain Group (BRIDG)

BRIDG is a collaborative effort engaging stakeholders from CDISC, the HL7 BRIDG Work Group, International Organization for Standardization (ISO), National Cancer Institute (NCI), and the Food and Drug Administration (FDA). The BRIDG model strives to be an overarching protocol-driven biomedical model in support of clinical research, with the goal of producing a shared view of the dynamic and static semantics for basic, preclinical, clinical, and translational research. The model is proposed to provide harmonization among standards within the clinical research domain, and between biomedical/clinical research and healthcare, with a focus on

supporting the day-to-day operational needs of those who run interventional clinical trials intended for submission to the FDA. With its 4+ releases, BRIDG now includes clinical and translational research concepts in its common, protocol representation, study conduct, AE, regulatory, statistical analysis, experiment, biospecimen, and molecular biology subdomains [16].

BRIDG has already been used by a number of groups as the underlying model for the development of clinical research systems, automated business process support for the conduct of research, and the representation to inform standardization of protocol data collection and conduct. The development of such standardized CRI tools also continually informs and advances the BRIDG model representation to be more useful and broadly applicable across all clinical research. The current BRIDG model version is in both Unified Modeling Language (UML) and OWL.

Ontology of Clinical Research

While the BRIDG model focuses on modeling the administrative and operational aspects of clinical trials to support clinical trial execution, the Ontology of Clinical Research (OCRe) focuses on modeling the scientific aspects of human studies to support their scientific interpretation and analysis [17]. Thus, OCRe allows the indexing of research studies across multiple study designs, interventions, exposures, outcomes, and health conditions [18]. The OCRe is a formal ontology for describing human studies, providing methods for binding to external information standards (e.g., BRIDG) and clinical terminologies (e.g., SNOMED CT). OCRe makes clear ontological distinctions between interventional and observational studies. It models a study's unit of analysis as distinct from the unit of randomization, and it models study endpoints more deeply than BRIDG does, that is, as an outcome phenomenon studied (e.g., asthma), the variable used to represent this phenomenon (e.g., peak expiratory flow rate), and the coding of that variable (e.g., as a continuous or dichotomized variable). OCRe imports operational constructs from BRIDG where possible (e.g., BRIDG's detailed modeling of actions, actors, and plans). OCRe is the semantic foundation for the Human Studies Database Project, a multi-institutional project to federate human studies design and results to support large-scale reuse and analysis of clinical research results [19]. OCRe is also modeled in both OWL and UML.

Other Protocol Models

Other protocol model representations include epoch and the primary care research object model (PCROM) [20, 21]. Like BRIDG, these models are primarily concerned with modeling clinical trials to support clinical trial execution. The WISDOM model represents clinical studies primarily for data analysis [22]. The Ontology for Biomedical Investigations (OBI) is a hierarchy of terms including some that are relevant to clinical research (e.g., enrollment, group randomization) [10]. OBI differs from BRIDG, OCRe, WISDOM, and other protocol models in that it is a

standardization and representation of *terms* in clinical research, but not a model of the *structure* of research studies. A common structured protocol model may come from blending the operational modeling of BRIDG, the scientific and statistical analysis modeling of OCRe and WISDOM, and the terminological modeling of OBI.

Eligibility Criteria Representation Standards

Eligibility criteria specify the clinical and other characteristics that study participants must have for them to be enrolled on the study. As such, eligibility criteria define the clinical phenotype of the study cohort and represent a protocol element of immense scientific and practical importance. Making eligibility criteria computable would offer substantial benefits for providing decision support for matching eligible patients to clinical trials and to improving the comparability of trial evidence by facilitating standardization and reuse of eligibility criteria across related studies. Hence, there have been many attempts to represent eligibility criteria in computable form, but there does not yet exist a dominant representational standard.

Part of the challenge of representing eligibility criteria is that they often are written in idiosyncratic free-text sentence fragments that can be ambiguous or under-specified (e.g., “candidate for surgery”). Indeed, in one study, 7% of 1000 eligibility criteria randomly selected from [ClinicalTrials.gov](#) were found to be incomprehensible [23]. The remaining criteria exhibited a wide range of complex semantics: 24% had negation, 45% had Boolean connectors, 40% included temporal data (e.g., “within the last 6 months”), and 10% had if-then constructs. Formal representations of eligibility criteria ideally should be able to capture all of this semantic complexity while capturing the clinical content using controlled clinical vocabularies. In addition, if the criteria are to be matched against EHR data (e.g., to screen for potentially eligible study participants), the representation needs a patient information model to facilitate data mapping from the criterion to the patient data (e.g., mapping a lab test value criterion to the appropriate EHR field). The major projects on eligibility criteria representation differ in the ways they address these needs.

The agreement on standardized protocol inclusion requirements for eligibility (ASPIRE) project defined key “pan-disease” (e.g., age, demographics, functional status, pregnancy) as well as disease-specific criteria (e.g., cancer stage) stated as single predicates (i.e., one characteristic, one value) [24]. For each criterion, ASPIRE defined the allowable values (e.g., stage I, II, III, or IV). This approach offers an initial high-level standardization of the most clinically important eligibility criteria in each disease area. Disease-specific standardized criteria had been defined for the domains of breast cancer and diabetes. ASPIRE does not aim to capture the complete semantics of eligibility criteria, nor does it include reference to a patient information model. ASPIRE would therefore not be sufficient as the sole formal representation for eligibility criteria in a fully computable protocol model but has the potential benefit of lower adoption barriers.

The Eligibility Rule Grammar and Ontology (ERGO) project takes a different approach than ASPIRE. ERGO aims to capture the full semantics of eligibility

criteria from any clinical domain in a template-based expression language; but encoding criteria into formal expression languages is difficult and time-consuming [25]. The ERGO investigators therefore developed ERGO Annotation, a lighter-weight template model that captures substantial semantic complexity (e.g., Boolean connectors, quantitative and temporal comparators) and that can be converted programmatically to OWL DL or SQL queries to execute against patient data [26]. In preliminary work, natural language processing (NLP) techniques were used to assist in transforming eligibility criteria from free text into ERGO Annotation.

Milian et al. have proposed a method for building a library of structured eligibility criteria, with the aim of being able to compare populations under study across different trials, reuse structured representations of eligibility criteria, and automatically suggest more relaxed criteria that could enhance recruitment [27]. Other eligibility criteria representations include caMatch, SAGE, and GLIF [28–30]. While the latter two representations are for practice guidelines, representing the conditional part of guidelines is conceptually identical to representing eligibility criteria. Weng et al. reviewed this very active area of clinical research informatics work and concluded that an expressive language is highly desired for clinical decision support uses of eligibility criteria (e.g., eligibility determination) and that a patient model is important for matching against patient data but less so for uses such as facilitating reuse of criteria in protocol authoring [31].

Because the vision for the computable structured protocol model includes driving operations at the individual patient level, a computable representation of eligibility criteria for the e-protocol should be based on an expressive language, should reference a standard patient information model (e.g., HL7 RIM), and should code to a broad controlled clinical vocabulary. Further research is needed on developing and testing practically useful and usable expression languages for eligibility criteria and on standardized approaches to applying complex criteria semantics to patient data in EHRs. Recently a computable eligibility criteria description has been proposed, the eligibility criteria (EC) representation, a CDISC-compliant schema for organizing criteria along with a patient-centric model for their formal expression, linked with international classifications and codifications [32].

Examples of Computable Protocol-Driven Research over the Study Life Cycle

Although e-protocols have most often been used to drive clinical research management systems, their uses in fact span the entire life cycle of clinical research. This section discusses several illustrative examples of the potential benefits of a common computable protocol model in actual implementation.

Improving Study Design

Design-a-Trial was one of the first examples of using a declarative study protocol to drive a system that helps investigators design new trials [33]. More recently,

WISDOM has similar aims. Such systems benefit from a computable protocol model on which to implement complex design knowledge to guide users to instantiate superior study plans [22]. For example, if a user designs a randomized trial of Surgery A versus Surgery B, the system can default the variable a patient's surgery assignment be the independent variable in the study's primary analysis and to restrict allowable statistical analyses to those that are appropriate for dichotomous independent variables. These systems could therefore be valuable in training new investigators or to introduce new research methods to established investigators (e.g., adaptive designs) [34].

Once instantiated, execution of an e-protocol could be simulated using data from other studies and sources on such execution parameters as recruitment rates and baseline disease rates to iteratively optimize the design for study duration and cost. For example, an e-protocol's computable eligibility criteria could be matched against an institution's patient data repository for automated cohort discovery [35, 36]. At the study design stage, an investigator could modify the eligibility criteria to balance recruitment time with the selectivity of the eligibility criteria. Simulation of e-protocols to optimize study time and costs could save valuable clinical research resources.

Protocols.io is an open access repository platform that allows the user to enter an existing MS Word or PDF protocol document in a structured form [37]. The platform is easily customizable to accommodate entry of different study protocol contents. Once the protocol is entered into the system, the user is able to edit the protocol collaboratively and share it with other researchers or sponsors and export the protocol as a PDF or as a JavaScript Object Notation (JSON) file. While originally designed for formalizing laboratory protocols, we are currently exploring to applying the use case for clinical trial protocols to Protocols.io.

Improving Clinical Study Efficiencies

Integration of electronic medical record (EMR) data for secondary use of this information within clinical research, and therefore improved study efficiency, will be greatly facilitated through the e-protocol. Such secondary use of EMR data has the potential to greatly enhance the efficiency, speed, and safety of clinical research. By clearly defining the protocol information as encoded fields within the e-protocol, mapping the fields required within the CRF to data that may exist within the EMR will advance the evaluation and discovery of new treatments, better methods of diagnosis and detection, and prevention of symptoms and recurrences. Clinical research can be enhanced and informed by data collected during the practice of care, such as comorbid conditions, staging and diagnosis, treatments received, recurrence of cancer, and vital status and cause of death.

A fully computable e-protocol, with structured coded data rather than free text, offers a solid foundation for integrating the clinical research workflow with data capture into the electronic medical record or other care systems. Such integration would offer at least two major benefits. First, study-related activities that generate

EMR data (e.g., lab tests, radiological studies) would be clearly indexed to an e-protocol, clarifying billing considerations. Second, with computable e-protocols, decision support systems could combine scheduled study activities with routine clinical care whenever possible (e.g., a protocol-indicated chest X-ray coinciding with a routine clinical visit), to increase participant convenience and therefore participant retention and study completion rates.

Improving Application to Care and Research

Clinical research is a multibillion dollar enterprise whose ultimate value is its contribution to improving clinical care and improving future research. E-protocols can support results application by capturing in computable form the intended study plan, the executed study plan, and the eventual results, to give decision support systems the information they need to help clinicians critically appraise and apply the study results to their patients. Existing systems for evidence-based medicine support either rely on humans to critically appraise studies and use computers to deliver the information (e.g., UpToDate) or build and manage their own knowledge bases of studies for their reasoning engines. Neither of these approaches is scalable to the tens of thousands of studies published each year. With computable e-protocols of completed studies publicly available, point-of-care decision support systems like MED could be more powerful in customizing the application of evidence to individual patients via the EMR [38].

Moreover, most clinical questions are addressed by more than one investigation, and the totality of the evidence must be synthesized with careful attention to the methodological strengths and weaknesses of the individual studies. Currently, such systematic reviews of the literature are a highly time-consuming and manual affair, which limits the pace of scientific knowledge, reduces the return on investment of clinical research, and delays the determination of comparative effectiveness of health treatments. The Human Studies Database Project is using OCRe as the semantic standard for federating human studies design data from multiple academic research centers to support a broad range of scientific query and analysis use cases, from systematic review to point-of-care decision support [17].

The Protocol Model-Driven Future

We conclude this chapter with a view to the future. The current patchwork, paper-driven approach to clinical research is inefficient, redundant, and is impeding the advance of science by squelching opportunities for data sharing and reuse of various resources. It is an approach that is overdue for reengineering. Critically, the full promise of CRI for achieving this reengineering demands that study protocols become fully structured and computable.

Study protocols specify all the major administrative and scientific actions in a study and drive how studies are conducted, reported, analyzed, and applied. Making

protocols fully computable would improve efficiencies and quality throughout the life cycle of a study, from study design to participant recruitment to knowledge discovery. Making protocols electronic in the form of PDF or word processor documents is better than paper protocol documents but is no substitute for e-protocols based on computable protocol models that are semantically rich and indexed to controlled clinical vocabularies. Ideally, however, all e-protocols would be based on one common computable protocol model to maximize interoperability and efficiencies for managing data, systems, and knowledge across the entire clinical research enterprise.

While there are many ongoing initiatives addressing various parts of the problem, there remain large challenges to achieving the overall vision of a protocol model-driven future. First, modeling work from the clinical trial execution and analysis communities (e.g., BRIDG and OCRe, respectively) needs to be merged to provide a semantic foundation for the entire study life cycle. Second, the use of clinical vocabularies (e.g., SNOMED, RxNorm, locally developed vocabularies) needs to be harmonized and processes for standardizing clinical constructs established and adopted (e.g., ASPIRE for eligibility criteria, cSHARE for study outcomes). Thirdly, user-friendly tooling is greatly needed to support modeling and harmonization work in this complex domain, and new methods and tools are needed to gracefully integrate the semantic standards into clinical research systems to enable systems interoperation and data sharing.

Finally, the many sociotechnical challenges cannot be downplayed. Clinical research involves a broad and complex group of stakeholders from industry to regulators to academia that represent multiple diseases, multiple countries, and multiple, sometimes conflicting, interests. The adoption of clinical research standards, like the adoption of electronic health record standards, will be in fits and starts but is already on its way through initiatives like CDISC and other efforts. These efforts show that there is general agreement on the broad constructs of the common computable protocol model, but specific terms, controlled terminologies, and data elements are harder to get consensus on, and representational challenges still loom large particularly for modeling eligibility criteria and the scientific structure of clinical research studies. Nevertheless, moving clinical research practice away from paper-based protocol drivers and toward being driven by a shared fully computable protocol model is a vital and worthwhile goal that would pay immense dividends for both clinical research and science.

Acknowledgment Authors thank Ida Sim for her substantial contributions to a previous version of this chapter that appeared in Springer 2012 version of this text.

References

1. Shankar R, O'Connor M, Martins S, Tu S, Parrish D, Musen M, Das A. A knowledge-driven approach to manage clinical trial protocols in the Immune Tolerance Network. In: American Medical Informatics Association symposium, Washington, DC 25 Oct 2005 [poster]; 2005.
2. Sim I, Owens DK, Lavori PW, Rennels GD. Electronic trial banks: a complementary method for reporting randomized trials. *Med Decis Mak.* 2000;20:440–50.

3. Chan AW, Tetzlaff J, Altman DG, Gøtzsche PC, Hróbjartsson A, Krleža-Jeric K, et al. The SPIRIT initiative: defining standard protocol items for randomised trials. *Ger J Evid Qual Health Care.* 2008;2008:S27.
4. Peto R, Collins R, Gray R. Large scale randomized evidence: large simple trials and overviews of trials. *J Clin Epidemiol.* 1995;48:23–40.
5. Smith B, Ceusters W, Klagges B, Köhler J, Kumar A, Lomax J, et al. Relations in biomedical ontologies. *Genome Biol.* 2005;6:R46.
6. Clinical Data Interchange Standards Consortium. CDASH. 2010. Available at <http://www.cdisc.org/cdash>. Accessed Aug 2011.
7. National Cancer Institute. Standardized Case Report Form (CRF) Work Group. 2009. Available at https://cabig.nci.nih.gov/workspaces/CTMS/CTWG_Implementation/crf-standardization-sig/index_html. Accessed Aug 2011.
8. University of California San Francisco. The Ontology of Clinical Research (OCRe). 2009. Available at <http://rctbank.ucsf.edu/home/ocre>. Accessed Aug 2011.
9. Sim I, Olasov B, Carini S. An ontology of randomized trials for evidence-based medicine: content specification and evaluation using the competency decomposition method. *J Biomed Inform.* 2004;37:108–19.
10. The Ontology for Biomedical Investigations. Home page. 2009. Available at http://obi-ontology.org/page/Main_Page. Accessed Aug 2011.
11. <https://www.hl7.org>.
12. <https://www.hl7.org/RIM>.
13. <http://cdisc.org/standards/protocol.html>.
14. Hume S, Aerts S, Sarnikar S, Huser V. Current applications and future directions for the CDISC operational data model standard: a methodological review. *J Biomed Inform.* 2016;60:352–62.
15. Hutton B, Wolfe D, Moher D, Shamseer L. Reporting guidance considerations from a statistical perspective: overview of tools to enhance the rigour of reporting of randomized trials and systematic reviews. *Evid Based Ment Health.* 2017;20(2):46–52.
16. Beclen LB, Hastak S, Ver Hoef W, Milius RP, Slack M, Wold D, Glickman ML, Brodsky B, Jaffe C, Kush R, Helton E. BRIDG: a domain information model for translational and clinical protocol-driven research. *J Am Med Inform Assoc.* 2017;24(5):882–90.
17. Sim I, Carini S, Tu S, Wynden R, Pollock BH, Mollah SA, Gabriel D, Hagler HK, Scheuermann RH, Lehmann HP, Wittkowski KM, Nahm M, Bakken S. The human studies database project: federating human studies design data using the ontology of clinical research. *AMIA Summits Transl Sci Proc.* 2010;2010:51–5.
18. <https://code.google.com/archive/p/ontology-of-clinical-research/>.
19. Human Studies Database (HSDB) Project Wiki. Home page. 2010. Available at https://hsdb-wiki.org/index.php/HSDB_Collaborative_Wiki. Accessed Aug 2011.
20. Shankar RD, Martins SB, O'Connor MJ, Parrish DB, Das AK. Epoch: an ontological framework to support clinical trials management. In: Proceedings of the international workshop on healthcare information and knowledge management, Arlington, November 11–11, 2006. HIKM '06. New York: ACM; 2006. p. 25–32. <https://doi.org/10.1145/1183568.1183574>.
21. Speedie SM, Taweele A, Sim I, Arvanitis T, Delaney BC, Peterson KA. The primary care research object model (PCROM): a computable information model for practice-based primary care research. *J Am Med Inform Assoc.* 2008;15:661–70.
22. CTSpedia. Web-based interactive system for study design, optimization and management (WISDOM). 2009. Available at <http://www.ctspedia.org/do/view/CTSpedia/WISDOM>. Accessed Aug 2011.
23. Ross J, Tu S, Carini S, Sim I. Analysis of eligibility criteria complexity in randomized clinical trials. *AMIA Summits Transl Sci Proc.* 2010;2010:46–50.
24. Niland J. ASPIRE: agreement on standardized protocol inclusion requirements for eligibility. In: An unpublished web resource. 2007.
25. Tu SW, Peleg M, Carini S, Rubin D, Sim I. ERGO: a template-based expression language for encoding eligibility criteria 2008. http://128.218.179.58:8080/homepage/ERGO_Technical_Documentation.pdf.

26. Tu S, Peleg M, Carini S, Bobak M, Ross J, Rubin D, Sim I. A practical method for transforming free-text eligibility criteria into computable criteria. *J Biomed Inform.* 2011;44(2):239–50. Epub 2010 Sep 17 PMID: 20851207.
27. Milian K, Hoekstra R, Bucur A, Ten Teije A, van Harmelen F, Paulissen J. Enhancing reuse of structured eligibility criteria and supporting their relaxation. *J Biomed Inform.* 2015;56:205–19.
28. Cohen E. caMATCH: a patient matching tool for clinical trials. In: caBIG 2005 Annual Meeting, Bethesda, MD. April 12–13, 2005.
29. Tu SW, Campbell JR, Glasgow J, Nyman MA, McClure R, et al. The SAGE guideline model: achievements and overview. *JAMA.* 2007;14:589–98.
30. Boxwala A. GLIF3: a representation format for sharable computer-interpretable clinical practice guidelines. *J Biomed Inform.* 2004;37:147–61.
31. Weng C, Richesson R, Tu S, Sim I. Formal representations of eligibility criteria: a literature review. *J Biomed Inform.* 2010;43(3):451–67. Epub 2009 Dec 23.
32. Chondrogiannis E, Andronikous EV, Tagaris A, Karanastasis E, Varvarigou T, Tsuji M. A novel semantic representation for eligibility criteria in clinical trials. *J Biomed Inform.* 2017;69:10–23.
33. Wyatt JC, Altman DG, Healthfield HA, Pantin CF. Development of design-a-trial, a knowledge-based critiquing system for authors of clinical trial protocols. *Comput Methods Prog Biomed.* 1994;43:283–91.
34. Luce BR, Kramer JM, Goodman SN, Conner JT, Tunis S, Whicher D, Sanford Schwartz J. Rethinking randomized clinical trials for comparative effectiveness research: the need for transformational change. *Ann Intern Med.* 2009;151:206–9. Available at <http://www.annals.org/cgi/content/full/0000605-200908040-00126v1?paper toc>. Accessed Aug 2011.
35. Murphy S, Churchill S, Bry L, Chueh H, Weiss S, et al. Instrumenting the health care enterprise for discovery research in the genomic era. *Genome Res.* 2009;19:1675–81.
36. Niland JC, Rouse LR. Clinical research systems and integration with medical systems. In: Ochs MF, Casagrande JT, Davuluri RV, editors. *Biomedical informatics for cancer research.* New York: Springer; 2010.
37. Teytelman L, Stoliartchouk A, Kindler L, Hurwitz BL. Protocols.io: virtual communities for protocol development and discussion. *PLoS Biol.* 2016;14(8):e1002538. <https://doi.org/10.1371/journal.pbio.1002538>.
38. Kawamoto K, Houlihan CA, Balas EA, Lobach DF. Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. *BMJ.* 2005;330(7497):765.



Data Quality in Clinical Research

11

Meredith Nahm Zozus, Michael G. Kahn,
and Nicole G. Weiskopf

Abstract

Every scientist knows that research results are only as good as the data upon which the conclusions were formed. However, most scientists receive no training in methods for achieving, assessing, or controlling the quality of research data—topics central to clinical research informatics. This chapter covers the basics of acquiring or collecting and processing data for research given the available data sources, systems, and people. Data quality dimensions specific to the clinical research context are used, and a framework for data quality practice and planning is developed. Available research is summarized, providing estimates of data quality capability for common clinical research data collection and processing methods. This chapter provides researchers, informaticists, and clinical research data managers basic tools to assure, assess, and control the quality of data for research.

M. N. Zozus, PhD (✉)

Department of Biomedical Informatics, College of Medicine,
University of Arkansas for Medical Sciences, Little Rock, AR, USA
e-mail: mzozus@uams.edu

M. G. Kahn, MD, PhD

Department of Pediatrics and the Colorado Clinical and Translational Sciences Institute,
University of Colorado Anschutz Medical Campus, Aurora, CO, USA
e-mail: MICHAEL.KAHN@UCDENVER.EDU

N. G. Weiskopf, PhD

Department of Medical Informatics and Clinical Epidemiology, School of Medicine,
Oregon Health & Science University, Portland, OR, USA
e-mail: weiskopf@ohsu.edu

Keywords

Clinical research data · Data quality · Research data collection · Processing methods · Informatics · Management of clinical data · Data accuracy · Secondary use

Clinical Research Data Processes and Relationship to Data Quality

Data quality is foundational to trusting the results and conclusions from human research. Data quality is so important that a National Academy of Medicine (then, Institute of Medicine) report [1] was written on the topic. Further, two key thought leaders in the industrial and clinical quality arenas, W. E. Deming and A. Donabedian, specifically addressed data quality [2–4]. Data quality in clinical studies is achieved through design, planning, and ongoing management. Lack of attention in these areas is an implicit assumption that errors will not occur; such inattention in turn further threatens data quality by inhibiting the detection of errors when they do occur [5].

Data quality is broadly defined as fitness for use [6]. Unfortunately, for clinical investigators and research teams, data use and thus appropriate quality vary from study to study. Moreover, in clinical research, data collection and acquisition processes are often customized according to the scientific questions and available resources, resulting in different processes for individual studies or programs of research. Because methods to assure and control data quality are largely dependent on how data are collected and processed, they are complicated by this customization. Science-driven customization of data collection and management processes will likely persist as will variability in study designs employed across the spectrum of the National Institutes of Health (NIH) definition of clinical research. Thus, methodology for data quality planning in clinical research must account for such expected variation.

Similar to the decreased property value of a house with a serious foundation problem, it is no surprise that research conclusions are only as good as the data upon which they were based. As plans and construction of a house help determine quality, well-laid research protocols are the start of data quality planning, for example, by specifying measures with sufficient precision and reliability and by designing error prevention, detection, and mitigation into study procedures. These might include collection of independent samples or assessments or a step to confirm that device acquired data are within expected limits prior to disconnecting the leads. Such “quality by design” is important because it is rare that the quality of data can exceed that with which it was initially collected. The quality of data affects how the data can be used and, ultimately, the level of confidence that can be reposed in research findings or other decisions based on the data. Thus, study and data collection design must be concerned with assuring data quality from the start.

The types of data collected in clinical research include data that are manually abstracted or electronically extracted from medical records, observed in clinical exams, obtained from laboratory and diagnostic tests, or from various biological

monitoring devices, and from patient-completed questionnaires. The near future holds an explosion of novel data sources including mobile devices, wearables, in-home sensors, and social media sources [7, 8]. Each data source is associated with one or more methods by which the data were acquired. After acquisition, these data are subject to further processing. Whether data are collected specifically for a research project or whether data collected for other purposes are used, a data management plan should take into account the data source, precollection processing, the data acquisition method, and, finally, postprocessing. These concepts that capture the full spectrum of data manipulations from acquisition to analytics are collectively called traceability. While these steps apply regardless of where the data were collected, the data sources will likely influence the plan. In other words, one method does not fit all. Using the same method to collect and process all data will overlook both the sources of error and opportunities to prevent them. For example, data recorded on a form may be retrospectively abstracted from medical records, may be written directly onto the form by the patient, or may be recorded directly on the form by a provider during a study visit. Each of these data acquisition processes is subject to different sources of error and, therefore, may benefit from different error prevention or correction strategies. The same is true of precollection processing, data acquisition, and postprocessing methods, thus the need to take these into account in planning for data quality. This chapter is primarily concerned with how to accomplish this and will elaborate on this framework for assuring and controlling quality regardless of the data source, acquisition method, or processing.

Data quality and the discipline of informatics are inextricably linked. Representation, observation and measurement, and data processing impact data quality (Fig. 11.1). In turn, data quality impacts use and vice versa. For example,

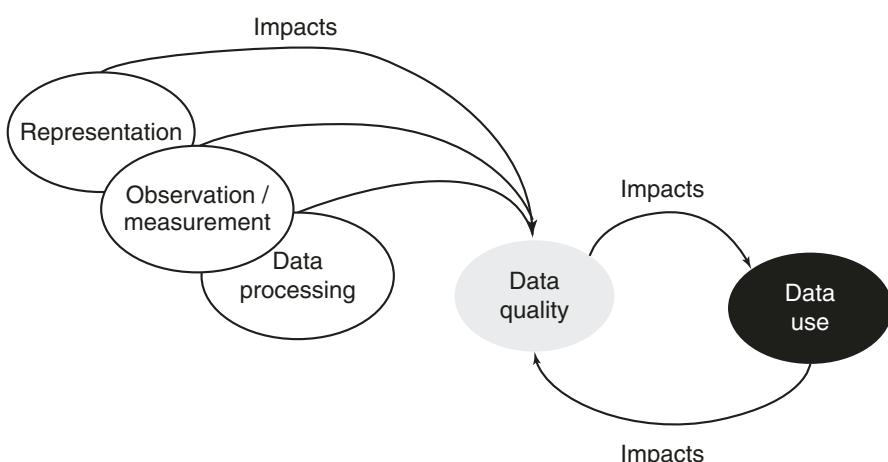


Fig. 11.1 The link between data quality and informatics. The way data are defined, collected, and handled impacts their quality. The quality of data impacts our willingness and ability to use them. Use of data and information by those who collect them causes more care to be taken in their definition, collection, and handling, increasing the quality

data used in clinical decision-making, performance measures, and quality improvement registries return benefit to those who record them; thus we expect that their quality is higher. In clinical research, data can be collected both prospectively and retrospectively, depending on the study protocol and local procedures. As such, information use in clinical care as well as information use in the study may impact data quality. Further, the goal of biomedical informatics is to improve human performance ostensibly through information use [9]. Thus, we posit that data quality is a consideration throughout all informatics-related disciplines, not merely a concern for data collection and processing.

Each step in the collection, handling, and processing of data has the potential to affect data quality. International Conference on Harmonization (ICH) guidelines state “Quality control should be applied to each stage of data handling to ensure that all data are reliable and have been processed correctly” and recommend a risk-based, process-oriented approach where researchers focus quality management on those processes that impact data integrity and human subject protection [10]. Though not always possible, as in the case of secondary analysis—also known as research using existing data—the gold standard in achieving quality is prevention rather than after-the-fact finding and fixing errors. Therefore, interventions aimed at preventing errors are typically designed into data collection and handling processes, i.e., part of the process rather than an after-the-fact checking activity applied to a data handling step. Similarly, methods for monitoring data quality are built into data collection and handling processes and should cover all operations performed on data.

Assuring data quality in clinical research is largely a focus on preventing or reducing *data errors that matter* – those that have the potential to adversely impact research results or safety of research subjects. We define a data error consistently with the International Organization for Standardization (ISO) 8000 family of standards as a data value that does not accurately reflect the true state of the thing being represented at the designated point in time [11]. Data represent things (or states or characteristics of things) in the real world. Thus, data that are correct at one time point will not necessarily remain so at a later time point. Therefore, the time at which a value was observed is necessary context to understand accuracy of that data point over time, for example, patient age as of the first study visit; air temperature in degrees Celsius at latitude 35.620252°N, at longitude –82.542933°W, and at an elevation of 2310 ft at noon last year on May 23rd or medications taken within the 10-day time window before the blood draw (see discussions of reliability and validity in Chaps. 4 and 11). The use of a broader definition than “inaccuracies created in data processing,” or “nonconformance to data specifications,” is intentional because inaccuracies from any source may render data values incorrect. Data quality can be compromised at any point along the continuum of data definition, collection, and processing, as demonstrated by the following examples adapted from actual cases. In fact, data accuracy may be thought of as having two components: (1) representational inadequacy such as an incomplete representation missing contextual information necessary for interpretation and (2) degradation or loss of information such as in the case of information reduction through mapping or errors introduced during

data collection. In this chapter, we develop and apply a framework for preventing and controlling data errors in prospectively collected data as well as for assessing data quality for secondary use of existing data. Consider the following scenarios.

Example 1

A large multisite clinical trial was sponsored by a pharmaceutical company to obtain marketing authorization for a drug. During the final review of tables and listings, an oddity in the electrocardiogram (ECG) data was noticed. The mean heart rate, QT interval, and other ECG parameters for one research site differed significantly from those from any other site; in fact, the values were similar to ones that might be expected from small animals rather than human subjects. The data listed on the table were found to match the data collection form and the data in the database, thereby ruling out data entry error; moreover, there were no outliers from that site that would have skewed the data. After further investigation, it was discovered that a single ECG machine at the site was the likely source of the discrepant values. Unfortunately, the site had been closed, and the investigator could not be contacted. This example was adapted from the Society for Clinical Data Management [12].

Example 2

In the course of a clinical research study, data were single entered at a local data center into a clinical data management system. During the analysis, the principal investigator noticed results for two questions that seemed unlikely. The data were reviewed against the original data collection forms, and it was discovered that on roughly half of the forms, the operator entering the data had transposed “yes” and “no.” Closer examination failed to identify any characteristics particular to the form design or layout that might have predisposed the operator to make such a mistake. Instead, the problem was due to simple human error, possibly from working on multiple studies with differing form formats. This example was adapted from the Society for Clinical Data Management [12].

Example 3

A clinical trial of subjects with asthma was conducted at 12 research sites. The main eligibility criterion was that subjects must show a certain percentage increase in peak expiratory flow rate following inhalation of albuterol using the inhaler provided in the drug kits. Several sites had an unexpectedly high rate of subject eligibility compared with other sites. This was noticed early in the trial by an astute monitor, who asked the site staff to describe their procedures during a routine monitoring visit. The monitor realized that the high-enrolling sites were using nebulized albuterol (not permitted under the study protocol), instead of the albuterol inhaler provided in the study kits for the eligibility challenge. Because nebulized albuterol achieves a greater increase in expiratory flow, these sites enrolled some patients who would not otherwise have been eligible. Whether due to misunderstanding or done deliberately to increase their enrollment rate (and financial gain), the result was the same: biased and inaccurate data. This example was adapted from the Society for Clinical Data Management [12].

Example 4

A multicenter pragmatic clinical trial was conducted to measure efficacy of a cancer screening process. The study planned to rely on health record data for cohort identification and outcome measures. There were multiple options for primary endpoints (1) whether the patient completed the initial screen and provided the sample in response to a mailed home screening kit and (2) whether patients with positive initial screening tests followed through and completed a second screening. The latter could not be used as an endpoint because the data across the multiple facilities were not consistently collected in routine care. In the planning stages, it was discovered that multiple facilities referred out for the second stage screening test and that some facilities did not routinely receive a follow-up report. Further, when follow-up reports were received, they were variously documented in the patient's record by methods including scanned images of faxed reports, entry of the result in text fields, and entry into structured coded fields.

Each of these scenarios describes a data quality problem, one in device-based data collection, one in data processing, one in a measurement procedure, and one in secondary data use. Despite the differences in setting and in the sources of the errors, the end result was the same: inaccurate data.

A 1999 National Academy of Medicine report [1] emphasized the importance of data quality to regulatory decision-making, in part, drawing conclusions from clinical trials. At the time, there was little in the literature base to synthesize in the report. Since the first edition of this text, there has been methodological progress toward data quality assurance, assessment, and control in clinical research. The approach presented here draws from a synthesis of experience and first principles.

Errors Exist

Errors occur naturally by physical means and human fallibility. Some errors cannot be prevented or even detected, for instance, a study subject who deliberately provides an inaccurate answer on a questionnaire or a measurement that is in range but inaccurate due to calibration drift or measurement error. Nagurney reports that, up to 8% of subjects in a clinical study could not recall historical items and up to 30% gave different answers on repeat questioning [13]. A significant amount of clinical data consists of information reported from patients. Further, as Feinstein eloquently states,

In studies of sick people, this [data accuracy] problem is enormously increased because (1) the investigator must contemplate a multitude of variables, rather than the few that can be isolated for laboratory research; (2) the variables are often expressed in the form of verbal descriptions rather than numerical dimensions; (3) the observational apparatus consists mainly of human beings, rather than inanimate equipment alone [14].

With clinician observation, reading test results, or interpreting images, human error and variability remain as factors. Simply put, where humans are involved, human error exists [15]. Reports of error or agreement rates can be found in the literature

for most types of assessment, observation, or interpretation of test results. These known and real errors and inconsistencies should be accounted for in data quality planning in clinical research.

Moreover, in every process, nature affects every project every day. As time passes, natural forces cause machines to wear, settings to drift, and attention to wander. Thus, while measurements and processes capable of achieving the desired levels of quality are often sought and employed in a research project, energy and vigilance must continuously be applied to maintain them.

Natural laws, logic, and empirical evidence together suggest that it is unwise to assume any data set is truly error-free. Still, respondents to a data quality survey conducted by the author [16] and others [17] have noted perfect data as their acceptance criterion. References to fear of consequences from regulators and potential data users observing obvious errors [1], such as a diastolic blood pressure of 10, suggest that the real concern may be the doubt that a user-discovered data error casts on the rest of the data set. The concern of obvious errors discrediting a data set will likely increase with more public data sharing, so methods such as looking at descriptive statistics, outliers, frequencies, and distribution graphs to efficiently scan a data set will persist.

Within the context of a given research project, pursuing data quality to a greater extent than needed to support the conclusions and foreseeable secondary use is unnecessary. Thus, data quality plans must be informed by the necessary level of data quality and must target the necessary level of data quality in the most cost-effective way. Two questions naturally result from this line of thought:

1. How clean do the data need to be to support the intended analysis?
2. What is the best method, given the study context, to achieve this?

The first is a statistical question, and the second is a design and engineering problem for the experienced informaticist or clinical research data manager to tackle.

Defining Data Quality

The National Academy of Medicine (NAM) defines quality data as “data strong enough to support conclusions and interpretations equivalent to those derived from error-free data” [1]. Like the “fitness for use” definition [6], the NAM definition is use dependent. Further, the robustness of statistical tests and decisions to data errors differs. Thus, applying the NAM definition requires *a priori* knowledge of how a statistical test or mode of decision-making behaves in the presence of data errors. For this reason, in clinical research, it is most appropriate that a statistician be involved in setting the acceptance criterion for data quality.

Further specification of the NAM definition of data quality is necessary for operational application. Other authors who have discussed data quality define it as a *multidimensional concept* [6, 18–25]. In clinical research, the dimensions most

commonly considered are *reliability*, *validity*, *accuracy*, and *completeness* [24, 26]. A more recent review of data quality assessment for electronic health record (EHR) data used in clinical research identifies the latter two plus concordance, plausibility, and currency [27]. Reliability and validity address the underlying concept being measured, i.e., is this question a reliable and valid measure of depressive mood? Accuracy is important with respect to and intrinsic to the data value itself. For example, does a heart rate of 92 represent the patient's true heart rate at the time of measurement? That is, *is it correct?* And completeness is a property of a set of data values; i.e., *are all of the data there?* More recently, as research methods have matured and data are increasingly used for monitoring and decision-making during a clinical study (as in the case of data and safety monitoring boards), *curren-*cy has emerged as an important data dimension. Concordance is defined by Weiskopf and Weng as agreement between data values in the EHR or between the EHR and another data source; they define plausibility as a value in the EHR seeming reasonable in terms of other knowledge [27]. Further, regulatory authorities are concerned with trustworthiness of the data and initially identified the following data quality dimensions for data submitted to the FDA: “electronic source data and source documentation must meet the same fundamental elements of data quality (e.g., attributable, legible, contemporaneous, original, and accurate) that are expected of paper records and must comply with all applicable statutory and regulatory requirements” [25].

These “fundamental elements,” *attributable*, *legible*, *contemporaneous*, *original*, and *accurate*, are commonly referred to as ALCOA. The European regulatory authority added complete, consistent, enduring, and available when needed to the ALCOA dimensions stated by the US FDA [28]. Registries commonly report data quality in terms of accuracy and completeness [26]. As secondary use of data has grown, so has the need for data to be *specified*, *accessible*, and *relevant*. Similarly, the dimension of *volatility*, or how quickly the data change, becomes a concern; for example, studies in adult populations seldom collect height at annual study visits, but studies in pediatric populations are likely to do so. These fundamental dimensions are different aspects of data quality. The relevance of each dimension waxes and wanes with various data uses, allowing users, especially secondary users, to evaluate the likelihood that data will support their specific (secondary) use. As we begin to see an increase in secondary, particularly research, uses of clinical data, the need for measuring and reporting on fundamental dimensions of data quality will become a necessary data itself.

The multidimensionality of data quality causes ambiguity because any given use of the term might refer to a single dimension or to some subset of possible dimensions. Further, different data users may emphasize some dimensions while excluding others; for instance, the information technology (IT) sector tends to assess data quality according to conformance to data definitions, valid values, valid formats, and stated business rules, while regulatory authorities are concerned with ALCOA and traceability [25]. Although accuracy and completeness historically have been emphasized in the clinical research literature, multiple dimensions ultimately affect and determine the usefulness of data. Each individual dimension describes an element of quality that

is necessary but usually not sufficient for data to be useful for their intended purpose. When maintained as metadata, dimensional measures can be used to assess the quality of the data for both primary and secondary uses of data.

Many dimensions may be calculated for any data, but often the circumstances surrounding a given use include built-in processes that obviate need for explicit measurement of one or more dimensions. For example, in a clinical trial, those who use data often have a role in defining it, meaning the *definition* is of little concern to the original study team. However, when data are considered for secondary uses, such as a pooled analysis spanning a number of studies, *relevance* and *definition* become primary concerns. By employing a dimension-oriented approach to data quality, these assumptions become transparent, helping us to avoid overlooking important considerations when working with new data or in new situations. In other words, describing data quality using dimensions increases the explicitness with which we measure, monitor, and make other decisions about the fitness for use of data.

Measuring data quality in an actionable way requires both operational definitions and acceptance criteria for each dimension of quality. An approach that facilitates collaboration across studies and domains includes standard operational definitions for dimensions, with project-specific acceptance criteria. For example, *timeliness* can be operationally defined as the difference between the date data were needed and the actual date they became available. The acceptance criterion—"How many minutes, days, or weeks late is too late?"—is set based on study needs. Further, some dimensions are inherent in the data, i.e., characteristics of data elements or data values themselves, while others are context dependent further increasing usefulness of standard operational definitions in conjunction with use-specific acceptance criteria. Table 11.1 contains common clinical research data quality dimensions, labels each dimension as inherent or context sensitive, labels the level at which it applies, and suggests an operational definition.

As highlighted by the previous sections, terminologies, definitions, and assessment methods are used inconsistently across publications, making it difficult to know how one publication relates to or builds upon previous literature. While a universal set of terms, definitions, and assessment methods currently do not yet exist, a recent effort by a large national collaborative focused on an initial set of data quality terms for describing three key data quality dimensions for secondary use of EHR data [29]. In the harmonized data quality terminology, data quality is segmented into three top-level dimensions: conformance, completeness, and plausibility. Each dimension builds upon the previous in specificity and complexity. Conformance focuses on the structural features of the data that are present without any reference to the meaning of the data. Structural features refer to adherence to the use of correct data formats and allowed data values. Data completeness focuses on the mere existence of data values (missingness, temporal and atemporal density) without reference to the accuracy or believability of the data values. Plausibility focuses on the believability of the data values, as individual values, as a temporal sequence of values, and/or as a set of interrelated values. The model also notes that data quality may be assessed using the existing data as its own reference (called the

Table 11.1 Data quality dimensions for clinical research

Dimension	Type	Natural language definition	Operational definition/metric
Accuracy	Inherent	The data value matches the true value	Number of errors divided by number of fields inspected (implies comparison with gold standard)
Currency	Inherent	Length of time a data value has been stored (since last update). <i>Length of time from a change in the real-world state to the time when the data reflect the change</i>	Use/need date minus date data last updated
Completeness	Inherent	<i>The extent to which every represented real-world state is reflected in the data</i>	Number of missing values divided by number of fields assessed
Consistency (internal)	Inherent	Data values representing the same real-world state are not in conflict	Number of discrepant values divided by number of values subject to data consistency checks
Timeliness	Context dependent	Availability of data when needed	Data need date minus date data ready for intended use
Relevance	Context dependent	Data can be used to answer a particular question	Percentage of data elements or data values applicable to intended use
Granularity	Context dependent	Level of detail captured in data	Percentage of data elements or data values at level of detail appropriate for intended use
Specificity (no ambiguity)	Inherent	<i>Each state in the data definition (metadata) corresponds to one (or no) state of the real world</i>	Number of values with full ISO 11179 metadata including definition divided by number assessed
Precision	Context dependent	Number of significant digits to which a continuous value was measured (and recorded); for categorical variables, the resolution of the categories	Percentage of values with precision appropriate for intended use
Attribution	Inherent	Source and individual generating and updating data are inextricably linked to data values	Percentage of data values linked to source and user ID of individual who generated and changed record

Italicized wording quoted from Wand and Wang [21]

verification context) or relative to one or more external data sources such as national data source or local gold standards (called the validation context). The harmonized framework explicitly ignores other key data quality dimensions mentioned in previous sections, most importantly timeliness and currency. Also note that the framework does not contain commonly used data quality terms such as accuracy, precision, validity, or truthfulness. These terms are widely used with significantly varying definitions in contexts outside of data quality assessment, such as in the development of psychometric instruments, psychosocial surveys, and biometric test

evaluations. The harmonized terminology specifically selected the more neutral word “plausibility” in an attempt to avoid using a term that had very specific technical usage in other scientific disciplines.

A common terminology for describing data quality features (dimensions) is useful for comparing computational methods (how a data quality measure was calculated) and for comparing data quality assessment methods and findings across data “runs” and across multiple data partners in a research network. For example, Callahan studied over 11,000 data quality checks across six research networks [30]. Using the harmonized data quality framework, Callahan et al. categorized each check by the data quality dimension(s) that each check was assessing. Their findings showed that each of the six networks had markedly different distributions of the data quality dimensions that they were assessing via their data quality checks. Callahan’s findings underscore both the diversity of data quality checks that are in practice and how each network must optimize their data quality efforts in alignment with their network’s mission and needs. A second effort that leverages the harmonized data quality terminology is the development of a suite of data quality tools based on the data quality categories [31].

Each data user focuses on specific variables of interest for data collection, management, analysis, and data quality assessment. Global data quality measures (called intrinsic data quality) focus on overall assessment of a data set, whereas fitness-for-use data quality measures focus on assessment of use-specific data elements. While the overall computational methods and data quality metrics are often the same, the broader focus of global data quality measures typically allows an investigator to only determine that a data set is unfit for use but does not provide sufficient drill-down insights to determine true fitness for use. The more narrow fitness-for-use data quality measures provide insights into the acceptability of a data set for a specific intended use but provide less insight into the fitness of a data set for some other use case that requires a different set of data elements. Thus both global and use-specific data quality measures are required to obtain a complete view of the data quality of a given data set for a specific use.

Related to but different than the harmonized terminology, the same data quality collaborative developed a set of recommendations to help guide both authors and readers toward a common set of data quality reporting components [32]. These recommendations can be used to help establish the features and components of a data quality plan, governance, and oversight structures.

Systematic Data Quality Planning

Over the past decade or more, the number and diversity of both new technology and new data sources have increased [33]. Managing new technology or data sources on a given project is now a normal aspect of clinical research data management. One of the largest challenges is preparing investigators and data managers to work with new technology and data sources. Methodology is needed that will help investigators and data managers (1) systematically assess a given data collection scenario,

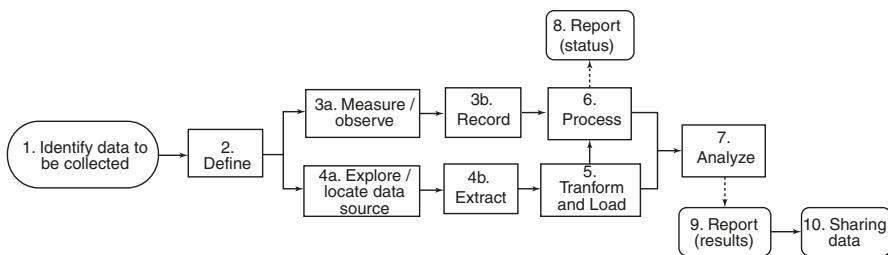


Fig. 11.2 Data-centric view of the research process. A set of general steps for choosing, defining, observing, measuring, recording, or otherwise obtaining, analyzing, and using data apply to almost all research. (Adapted from Data Gone Awry [12], with permission)

including new technology and data sources, (2) systematically evaluate that scenario, and (3) apply appropriate methods and processes to achieve the desired quality level.

A dimension-oriented data quality assessment approach helps assure that data will meet specified needs; however, data quality assessment alone is an incomplete solution. A systematic way to assess data sources and processes for a project is necessary. Figure 11.2 shows the set of steps comprising the data-related parts of the research process. These steps are described at a general level so that they can be applied to any project. From the data-oriented point of view, the steps include (1) identifying data to be collected; (2) defining data elements; (3a) observing and measuring values; (3b) recording those observations and measurements; (4a) locating and evaluating existing data for use in the study; (4b) extracting or otherwise obtaining the existing data; (5) transforming that data if necessary and importing, i.e., loading it into the study data system; (6) processing data to render them in electronic form and prepare them for analysis; and (7) analyzing data. While research is ongoing, data may be (8) reported for use managing or overseeing the project. After the analysis is completed, (9) results are reported, and (10) the data may be shared with others.

Identifying and Defining Data to Be Collected

Identifying and defining the data to be collected are critical aspects of clinical research. Data definition initially occurs as the protocol or research plan is developed. Too often, however, a clinical protocol reads more like a shopping list (with higher-level descriptions of things to be collected, such as *paper towels*) than a scientific document (with fully specified attributes such as *brand name, specific product, weight, size of package, and color of paper towels*). When writing a protocol, the investigator should be as specific as possible because in large studies, the research team will use the protocol to design the data collection forms. Stating in the protocol that a pregnancy test is to be done at baseline is not sufficient—the protocol writer should specify the type of sample on which the test is to be conducted (e.g., a urine dipstick pregnancy test is

to be performed on women of childbearing potential) and whether a local or central lab is to be used. A study procedures manual will then contain the specifics of sample collection and preparation.

As standards such as the Protocol Representation Standard [34] mature and supporting software becomes available, full specification of protocol elements will become the most efficient method for defining data, as metadata specified in the protocol will be immediately available for generation of data collection forms. (See Chap. 9). Lack of specificity in data definition is the mechanism by which data identification and definition can cause serious data quality problems, for example, two sites using different measurement methods, or otherwise not measuring the same thing. The information necessary to fully specify a clinical measurement, with context sufficient to remove ambiguity, differs based on the type of data. For example, specification of the specimen (and often, the method by which the specimen is obtained) is important for some tests. For blood pressure measurements, the position, resting period, location of measurement, and device used may be important. Without careful identification and specification of this context, data collectors at clinical sites may inadvertently introduce unwanted variability.

The principle of “Occam’s razor” applied to clinical research suggests collecting only the data needed to assure patient safety, answer the scientific question(s), and uniquely identify the collected data values. Jacobs and Studer report that for every dollar spent to produce a data collection form, \$20 to \$100 are required to fill one in, process it, and store the data, emphasizing that “the true cost of a form involves people not paper” [35]. When extensive data cleaning is required, this ratio becomes even more exaggerated. Eisenstein and colleagues report significant cost savings in clinical trials by decreasing the amount of data collected [36, 37]. At the time of this writing, the relationship between form length and data accuracy for online forms remains unprobed [38]. However, the evidence relating form length to decreased response rate while considered equivocal by some [38] has been demonstrated in controlled and replicated experiments [39, 40]. There is no question that collecting more data increases costs and places additional burden on clinical investigational sites and data centers [36, 37].

These two principles, parsimony in the number of data elements collected and full specification of those that are collected, are preventative data quality interventions. Parsimony, or lack thereof, may impact data accuracy and timeliness dimensions, while data definition impacts the specificity dimension and significantly impacts secondary data users.

Defining Data Collection Specifications

The previous section covered the definition and specification of data elements themselves. This section covers definition of the tools, often called data collection forms or case report forms, for acquiring data. The design of data collection forms, whether paper or electronic, directly affects data quality. Complete texts have been written on form design in clinical trials (see Data Collection Forms in Clinical Trials by Spilker

and Schoenfelder (1991) Raven Press NY). There are books on general form design principles, for example, Jacobs and Studer [35] *Forms Design II: The Complete Course for Electronic and Paper Forms*. In addition, the field of usability engineering and human-computer interaction has generated many publications on screen or user interface design. While this topic is too broad to discuss in depth here, two principles that are directly relevant to clinical research informatics, and for which application to clinical research is not covered in more general texts, warrant attention here. The first is the match between the type of data and data collection structure; the second is the *compatibility-proximity principle* [41]. The second is the general assumption which is that the more structured the data, the higher the degree of accuracy and ease of processing. However, this can be counterbalanced by considerations related to ease of use.

As a general principle, the data collection structure should match the type of data. Data elements can be classified according to Stevens' scales (nominal, ordinal, interval, and ratio) [42] or as categorical versus continuous or according to various other similar schemes. Likewise, classification can also be applied to data collection structures describing how the field is represented on a form, including verbatim text fill in the blank, drop-down lists, check boxes ("check all that apply"), radio buttons ("check one"), and image maps. Examples of data collection structures are shown in Fig. 11.3.

- a. **Write in** (the electronic equivalent of "fill in the blank")

Method of Birth Control: _____ Barrier method _____

- b. **Drop down list**

Method of Birth Control:

Barrier method	▼
Sterilization	
Abstinence	
Birth control pills	

- c. **Check lists** (the electronic equivalent of "check all that apply")

Method of Birth Control:

- Sterilization
- Barrier method
- Abstinence
- Birth control pills

- d. **Radio button** (the electronic equivalent of a "check")

Method of Birth Control:

- Sterilization
- Barrier method
- Abstinence
- Birth control pills

Fig. 11.3 Example data collection structures. For many data elements, more than one data collection structure exists

Mismatches between data type and collection structure can cause data quality problems. For example, collecting data at a more granular structure than exists or that can be discerned in reality, for example, 20 categories of hair color, invites variability in classification. Collecting data at a less granular structure, *data reduction*, that can be discerned in reality also invites variability and results in information loss. The original detail cannot be resolved once the data are lumped together into the categories. For example, if height is collected in three categories, short, medium, and tall, the data cannot be used to answer the question, “how many subjects are over 6 feet tall?” Another way to think about data reduction is in terms of Stevens’ scales [42]. Data are reduced through collection at a lower scale, for example, collecting a yes or no indicator for high cholesterol. When the definition of high cholesterol changed, data sets that collected the numerical test result continued to be useful, while the data sets that contained an indicator, yes or no to high cholesterol, became less so. There are many cases such as high-volume data collected through devices where reduction in the number of data values collected or retained or stored is necessary and desirable. The amount of information loss is dependent on the method employed. Reduction of CRF data occurs through both data collection at a lower scale than the actual data and through decision not to collect certain data values. Because data reduction results in information loss, it limits reuse of the data and should only be employed after careful deliberation.

Data collection structure can cause quality problems in capturing categorical data in other ways. When the desired response for a field is to mark a single item, the available choices should be exhaustive (i.e., comprehensive) and mutually exclusive [43–45]. Lack of comprehensiveness causes confusion when completing the form, leading to unwanted variability. Similarly, overlapping categories cause confusion and limit reuse of the data.

The *compatibility-proximity principle* was first recognized in the field of cognitive science. When applied to the design of data collection forms, it means that the representation on the form should match as closely as possible the cognitive task of the person completing the form. For example, if body mass index (BMI) is a required measurement, but the medical record captures height and weight, the form should capture height and weight. This matches the medical record abstractor’s task of finding the value and recording it on the form and keeps the operation one-to-one. For the same reason, values on the form should allow data to be captured using multiple units so that the person completing the form is not required to convert units. Importantly, the flow of the form should follow as closely as possible the flow of the source document where one exists [43–45]. An additional application of the compatibility-proximity principle is that all items needed by the person completing the form should be immediately apparent on the form itself (separate form completion instruction booklets are less effective) [44]. There is evidence that data elements with higher cognitive load on the abstractor or form completer also have higher error rates [45–57]. Adhering to the compatibility-proximity principle and keeping data collection and recording tasks “one-to-one” helps decrease cognitive load.

There are, however, four countervailing factors that must be weighed against the compatibility-proximity principle: (1) for projects involving multiple sites, matching aspects of each site’s medical record in the data collection form representation

may not be possible; (2) there may be reasons for using a more structured data collection form that outweighs the benefits of precisely matching the medical record; (3) in circumstances where a calculated or transformed value is necessary for immediate decision-making at the site, “one-to-one” data collection and recording should be maintained with addition of a real-time solution or tool to support the additional cognitive tasks is needed; such a tool would use the raw data as input; and (4) it may not be possible to design forms that match clinical source documents or workflow, for example, some electronic systems limit data collection structure to one question-answer pair per line, precluding collection of data using tabular formats.

Defining data collection is not limited to the data collection structure. It also includes the source and means by which the data will be obtained. For example, will data be abstracted from medical records, collected *de novo* from patients directly, or collected electronically through measuring devices? The identification of possibilities, selection of one over the alternatives, and deciding whether multiple mechanisms can be used without adverse impact is a design decision requiring knowledge of the advantages and disadvantages of each option and how they impact costs and the relevant dimensions of data quality. Thus, ability to characterize data sources and processes in these terms is a critical competency of clinical research informaticists.

Like parsimony, choice of and full specification of the data collection mechanism is a preventative data quality intervention. The chosen data sources and mechanisms of collection and processing may impact data accuracy, precision, and timeliness dimensions, while the definition itself may impact the specificity dimension and the utility of data for secondary uses.

Observing and Measuring Data

The different methods of measurement and observation used in clinical research are too many and too various to enumerate here. Clinical data may be reported by the patient, observed by a physician or other healthcare provider, or measured directly via instrumentation. These reflect three fundamentally different kinds of data [58]. Further, some measurements return a value that is used directly (e.g., temperature), while others require interpretation (e.g., the waveform output of an electrocardiogram).

It is difficult (and sometimes impossible) to correct values that are measured incorrectly, biased, or gathered or derived under problematic circumstances. Recorded data can be checked to ascertain whether they fall within valid values or ranges and can be compared with other values to assess consistency, but doing so after the data have been collected and recorded, and in the absence of an independent recording of the event of interest, eliminates the possibility to correct errors in measurement and observation [58]. For this reason, error-checking processes should be built into measurement and observation whenever feasible. This can be accomplished by building redundancy into data collection processes [59, 60]. Some examples include (1) measurement of more than one value (e.g., taking three serial blood pressures), (2) drawing an extra vial of blood and running

a redundant assay for important measurements, (3) asking a different question to measure the same construct, and (4) measuring the same parameter via two independent methods. Immediate independent measurement with immediate feedback can be used to identify and correct discrepancies at the point of measurement and is of course the most rigorous approach [58]. Independent measurement alone can also provide a replacement value if needed (e.g., the second vial of blood that saves the day when the first vial hemolyzes). Independent assessment with immediate feedback should be distinguished from discrepancy checking with immediate feedback. Discrepancy checking is a comparison of a recorded value against a known standard, for example, a valid range, or relative comparison to another value. While discrepancy checking can identify some errors, it will miss those within the valid value range or relationship. Errors within the range of valid values can only be identified through redundancy. Secondly, error checking may occur at the point of measurement or recording but is sometimes implemented after the fact. In the latter case with no upstream or other source of truth, discrepancy checking serves as an identification mechanism rather than as a correction mechanism [58]. In summary, measurement discrepancies can be mitigated through careful procedures and training; however, errors are nonetheless inevitable. While discrepancy checking near or after measurement can identify errors, immediate independent verification with contemporaneous feedback is the only scenario in which data quality will not suffer some degradation [58].

Another important aspect of measurement and observation, one that has a critical effect on data quality, is ensuring consistency between or among clinical investigational sites. The albuterol screening test clinical trial example given at the beginning of the chapter reflects an all-too-common problem rooted in the fact that clinical investigational sites each operationalize the practice of medicine and research differently and institutional policies vary from location to location. In addition, equipment may vary from site to site, and there is usually at least some degree of staff turnover during studies, meaning that levels of available skill, knowledge, and experience at a given site will fluctuate over time. These and other factors contribute to variations in procedures governing observation and measurement, adding unwanted variability to clinical data.

For these reasons, clear, unambiguous, and uniform procedures that all study personnel can follow are essential to maintaining data quality. Consistency can often be improved by providing sites with critical study-related equipment or devices (so that all study data are being gathered with the same devices), training site personnel in study procedures and the administration of tests and questionnaires, using central reading centers where rating or interpretation of data is required, and requiring all sites to follow equipment calibration schedules that offer preventative methods to improve data quality from measurement and observation.

Measurement and observation should also be subject to ongoing assessment and control. Some methods directly assess the measurement or observation; examples that include assessing interrater reliability, reviewing recorded interviews, and monitoring investigational sites for adherence to procedure are all ways of providing ongoing assessment and control. While other assessment and control methods are indirect, examples include counts of data inconsistencies, instances

of noncompliance to protocol specified time windows, and statistical methods of checking for aberrant data by site. These indirect methods may identify sites or study staff that may be performing aspects of the study differently from other sites. However, these indirect measures are only surrogates for data accuracy, i.e., measures of inconsistency, rather than direct assessment of accuracy. With such indirect assessments, care must be taken to respect natural variations (including those caused by variations in population) among sites. Assessment and control methods are usually targeted at the accuracy, timeliness, or completeness dimensions.

Recording Data

Recording data is the process of writing down (e.g., as from a visual readout or display) or directly capturing electronically data that have been measured, thereby creating a permanent record. The first time a data value is recorded—whether by electronic means or handwritten, on an official medical record form, or a piece of scratch paper, by a principal investigator or anyone else—is considered the source [7]. If questions about a study’s results arise, the researcher (and ultimately, the public) must rely upon the source to reconstruct the research results. Several key principles are applicable: (1) the source should always be clearly identified; (2) the source should be protected from untoward alteration, loss, and destruction; and (3) good documentation practices, as described by the US Food and Drug Administration regulations codified in 21 CFR Part 58 [61], should be followed. These practices include principles such as data should be legible, changes should not obscure the original value, the reason for change should be indicated, and changes should be attributable (to a particular person). While it seems obvious that the *source* is foundational, even sacred to the research process, cases where the source is not clearly identified or varies across sites have been reported and are common [62, 63]. Data quality is also affected at the recording step by differences such as the recorder’s degree of fidelity to procedures regarding number of significant figures and rounding; such issues can be checked on monitoring visits or subjected to assessment and control methods discussed in the previous section. Data recording usually impacts the accuracy, timeliness, or completeness dimensions. Where recording is not adequately specified, precision may also be impacted.

Processing Data

In a recent literature review and pooled analysis that characterized common data collection and processing methods with respect to accuracy, data quality was seen to vary widely according to the processing method used [64]. Further, it appears that the process most associated with accuracy-related quality problems, medical record abstraction, is the most ubiquitous, as well as the least likely to be measured and controlled within research projects [64]. In fact in a recent review, fewer than 9% of studies using medical record abstraction reported results of a quantitative quality

Table 11.2 Accuracy associated with common data processing methods

	Min.	Median	Mean	Max.	Std. dev.
Abstraction	70	647	960	5019	1018
Optical	2	81	207	1106	338
Single entry	4	26	80	650	150
Double entry	4	15	16	33	10
No batch data cleaning	2	270	648	5019	946
Batch data cleaning	2	36	306	1351	428

assessment [65]. While contemporaneous work called for reporting of data quality assessment results along with research results [32].

Although not as significant in terms of impact on quality as abstraction, the method of data entry and cleaning can also affect the accuracy of data. On average, double data entry is associated with the highest accuracy and lowest variability, followed by single data entry (Table 11.2). While optical scanning methods are associated with accuracy comparable to key-entry methods, they were also associated with higher variability. Other factors such as on-screen checks with single data entry, local versus centralized data entry and cleaning, and batch data cleaning checks may act as substantial mediators with the potential to mitigate differences between methods [64]. Additionally, other factors have been hypothesized in the literature, but an association has yet to be established, for example, staff experience [64], number of manual steps [66], and complexity of data [62]. For these reasons, measurement of data quality is listed as a minimum standard in the Good Clinical Data Management Practices document [66]. Because of the potentially significant impact that variations in data quality can have on the overall reliability and validity of conclusions drawn from research findings [67], publication of data accuracy with clinical research results should be required [32].

While our focus thus far has been on the accuracy dimension, data processing methods and execution can also impact timeliness and completeness dimensions. Impact on timeliness can be mitigated by using well-designed data status reports or otherwise actively managing data receipt and processing throughout the project or even prevented by designing processes that minimize delays. The impact of data processing on completeness can be mitigated in the design stages through collecting data that are likely to be captured in routine care or through providing special capture mechanisms, for example, measuring devices, capturing data directly from participants, or use of worksheets. Additionally, throughout the study, completeness rates for data elements can be measured and actively managed.

Analyzing Data, Reporting Status, and Reporting Results

Analyzing and reporting data differ fundamentally from other steps discussed in the preceding sections, as they lack the capacity to introduce error into the data values themselves. Errors in analysis and reporting programming or data presentation, while potentially costly, do not change underlying data. Analysis and reporting

programming is typically applied to a copy of the database. However, analysis and reporting do have the potential for error in themselves which may result in misrepresentation of the data and to present as data quality problems [68]. Assuring and controlling quality at the analysis and reporting stage are achieved through choice of appropriate methods, through validation of programming, and through applying good design principles to data analysis pipelines.

Planning for Data Quality

When starting a new project, the clinical data manager and/or clinical research informaticist is faced with a design task: match the data collection scenario for the project to the most appropriate data sources and processing methods. The first step is to group the data to be collected by data source, for example, medical history and medications may be manually abstracted from the medical record, blood pressures may come from a study provided device, lab values may be transferred electronically from a central lab, or the entire data set may be electronically extracted from an existing source, i.e., reused. Seemingly homogeneous data sets may in fact contain different data sources. Because different data sources are subject to different error sources and associated with varying extents of data processing, we recommend treatment by source.

An important distinction between data sources and available options for data quality assurance is the extent to which the initial observation or measurement of the data is within the control of the investigator [58]. Where initial observation or measurement is within the control of the investigator, prospective data quality assurance and control approaches such as those described earlier would be expected. Similarly, where the investigator does not control the initial observation or measurement but plans to undertake some data processing for a study, prospective data quality assurance and control approaches for those data processing activities would be expected. On the other hand, where the investigator does not control the initial observation or measurement, for example, with secondary use of EHR data or data from a completed study, and thus is not able to assert prospective assurance and control measures over initial observation or measurement or data processing, data quality assessment is still required to test capability of the data to support research conclusions. In a multicenter study, such assessment is necessarily performed by site.

To aid in planning, data sources and process by which the data are often diagrammed making it easier to see potential alternative sources, methods, and processes for consideration. For example, some data sources may have undesirable preprocessing steps or known higher variability that would exclude them from further consideration. Once the data sources have been chosen and the data gathering process has been specified, known error sources can be systematically reviewed to consider the possibility or necessity of error prevention or mitigation. At this point, data quality dimensions that are important to the research study can be assessed for each type of data and each processing step. The output of this process should be discussed with the research team and used to inform decision-making about the plan for data collection and management and documented in a data management plan for the study.

Assessing the Quality of Secondary Use Data

The fundamental difference between traditional clinical trial data and secondary use of healthcare data is that secondary use data are not collected for any specific or general research purpose. Rather, these data are a byproduct of complex healthcare systems and processes. The volume and variety of EHR data are tremendous, which is a benefit of using these data, but many of the assumptions and assurances one can make regarding prospectively collected research data do not hold for clinical data. First, patients only have contact with the healthcare system when there is a reason for them to do so and not always with the same provider or institution. Clinical data, therefore, are only collected episodically and are then often fragmented across multiple EHRs or other healthcare information systems. Second, when a patient does meet with a healthcare provider or service, usually only the clinical concepts relevant to that specific appointment will be captured in the EHR (exceptions to this are certain basic vitals or social history concepts). Third, the information that the patient conveys to the provider, or that the provider observes about the patient, must be entered into the EHR—a frequently manual process that is prone to inaccuracy and loss of information. And finally, once the data are in the EHR, most secondary use cases require some sort of data extraction and transformation in order to generate a usable data set; this process may also introduce data quality problems. It is no surprise, therefore, that the quality of EHR data is variable and often poor [69–71].

That said, there are steps that investigators can take to determine if the clinical data available to them are fit for their intended secondary use. Although investigators utilizing EHR data and other existing data sources do not have control over the prospective collection of these data, many of the quality assurance and assessment steps in secondary use are analogous to those in more traditional research paradigms. The bottom track of Fig. 11.2 above summarizes the common data-related steps in secondary use of clinical data for research and also shows the parallels between the two research approaches. It is important to note that the secondary use research process is often more iterative than is indicated by this figure, generally as a result of data quality problems.

Identification of Required Clinical Concepts

As in prospective research, the first research step in secondary use of clinical data—following the identification of a research question—is to identify the concepts that are required to answer that question. When reusing existing data, there may be a temptation to go on a fishing expedition to find significant associations or results. While there are certain cases where this approach is appropriate (e.g., in certain large-scale data mining efforts), most secondary use paradigms require the clear identification and description of research predictors, outcomes, and potential covariates. A research protocol should be no less clear in secondary use than in prospective research. Where there is likely to be a key difference, however, is in the fact that the concepts defined in this stage may later be determined to not be available in the clinical data source; in prospective research this is a less likely scenario.

Definition of Data Elements

Similarly, although these clinical data have already been collected according to clinical practice protocols and workflows, ensuring data quality during reuse requires defining and specifying data formats and, if necessary, abstraction tools. If, for example, data from an EHR are to eventually be loaded into a database, the fields in that database must be defined appropriately, e.g., binary variables, integers, floating point (decimal) values, date and time entries, etc. Ideally, this stage in the research process would include defining an entire data schema, with appropriate relational constraints and requirements. The category of data conformance from the harmonized data quality framework described above, which dictates accepted and appropriate data formats and standards, comes into play here [29].

Exploration and Availability Assessment of Clinical Data Source

At this stage in the research process secondary use truly departs from prospective research. When reusing existing clinical data, it is not possible to control how clinical phenomena are observed and measured. Rather, the clinical data already exist, and the researcher must determine which of the concepts required for their study are available and accessible. These clinical concepts have already been defined and formalized in the previous stages, so the next step requires that those concepts be mapped to existing fields in the clinical data source. In some cases this mapping is already at least partly established. For example, some institutions have adopted OHDSI's OMOP common data model [72], the PCORnet common data model [73], or some combination of the relevant data standards and terminologies (e.g., ICD10 or RxNorm). One benefit of these efforts is that sometimes the mapping between concepts and specific fields within the EHR has already been completed, thereby improving efficiency, reliability, and reproducibility (assuming the mapping has been done well). It is common, though, for investigators engaged in the secondary use of clinical data to have to perform at least some of this mapping (and sometimes all of it) manually. Mapping exercises usually result in information loss.

During this data exploration and mapping step, there are a few key data quality assurance and assessment methods. The most obvious is determining if the required clinical concepts are available at all. Some clinical research questions require very specific measurements and concepts that may not be recorded in the course of clinical care. Alternatively, a concept might be recorded, but not in a format that is accessible. Waveform data, for example, like those collected during an EKG or EEG, are frequently not available through an EHR or, if they are available, they may be included as attached images or PDFs, which cannot be extracted as computable data.

In secondary use, however, it is important to understand that data availability is rarely dichotomous. EHRs tend to be both fragmented and redundant—a single clinical concept may be recorded in multiple locations throughout the record. A diagnosis, for example, may be commonly found on a problem list but could also be extracted from billing data. In some cases a diagnosis might be mentioned in a

clinical note, but not in structured data. Other times a diagnosis can be inferred from relevant laboratory results, vital measurements, or medications. Therefore, the researcher must determine not only if they have found one field corresponding to their required concept but all corresponding fields.

Extraction of Relevant Data Elements

Once the relevant data fields within the EHR have been identified and mapped to the required clinical concepts for the secondary use case, the data must generally be extracted from the source using one or more queries. This is because health information systems rarely allow for direct analysis of data, beyond those simple aggregate statistics that can be calculated using queries. (It is worth noting that clinical data can rarely be extracted from “live” EHRs and are instead only available through back end databases, datamarts, or data warehouses, all of which have already been abstracted away from the live data to some extent.)

To ensure quality, the extraction process should be subject to various checks. This is also the stage where the concept mapping performed in the previous step can be assessed. The simplest checks involve comparing aggregate statistics like counts of data entries between the extracted data and the source data. For example, if the source data and extractions differ in numbers of patients, visits, lab results, or any other clinical concept, then there may have been an error in defining the scope of the query. It is also worth spot-checking a handful of representative records, ensuring that there is agreement in values between the source and extracted data. It is also beneficial to take advantage of the temporal nature of EHR data. For example, once data have been extracted, the investigator can plot simple aggregate statistics, especially counts, over time to identify potential failures in the extraction process. Most trends will be smooth. Significant leaps or dips in these trends generally indicate either extraction and mapping errors or notable changes in underlying EHR usage or care practices. The data extraction process (and the exploration process from the previous step) should be repeated as necessary.

Transformation and Curation of Extracted Clinical Data

Following the processes described in the previous steps, the extracted clinical data will often be stored in multiple files or data structures. At this point, the data must be transformed as necessary to allow loading into the previously defined data schema or definitions. This curation process will range from simple to complex for different concepts. The easiest data fields to transform and load into the schema are structured elements that exist only once for each patient (e.g., race or ethnicity) and are expected to remain consistent across clinical encounters. More commonly, each patient will have multiple instances of a single data element, as in the case of laboratory results. In such cases the decision must be made as to which value to select (e.g., most recent) or if some aggregate value (e.g., mean or median) should be used.

It may also be necessary at this stage to convert data from an entity, attribute, or value format to a “one column per data element” format. The situation becomes more complex for data that are documented in multiple places within the EHR. An even more complex scenario would be a situation where a single clinical concept might need to be inferred from multiple types of data. For example, a diagnosis that has low sensitivity in the EHR could be derived from a combination of problem list entries, laboratory results, and medications.

Each such transformation and curation of the extracted data introduces the opportunity for data quality problems. As above, spot-checking and comparison of aggregate statistics, like counts of records, are advised at this stage, following loading into the previously defined data schema. These comparisons should be made against the previously extracted data and/or against the source data.

Fitness-for-Use Assessment and Data Analysis

At this stage in the secondary use research process, once the definition, extraction, and curation of the research data set has been completed, it is time to determine if the data are in fact fit for the intended use. While the previous steps must be approached in such a way as to avoid the introduction of error, none of those data quality or assurance measures address underlying data quality problems in the source data. Prior to conducting the planned research analyses, the investigator must determine if the data are actually of sufficient quality to complete these analyses. Of the three major categories of data quality defined in the Kahn et al. data quality framework described above, conformance should have been addressed in previous research steps, leaving completeness and plausibility to be assessed at this stage.

To assess completeness, the investigator must consider their data from a number of dimensions. First, how many of the subjects in the sample have sufficient data for the intended analyses? Generally this means looking at how many of the expected or required clinical concepts are actually present for each patient. For longitudinal studies, though, the investigator must also consider the completeness of data at multiple time points for each subject. Second, for any variable that will be included in the analysis, are their sufficient data points available to power the analysis [27, 74, 75]?

Plausibility, as defined in the Kahn et al. framework, is analogous to what is commonly called accuracy or correctness in the data quality literature. True accuracy, however, can very rarely be assessed in the secondary use of clinical data. Instead, the investigator should at this stage determine if their data set is plausible when compared to external sources of knowledge (e.g., clinical expertise or medical literature) or sources of data (e.g., registry data or data from other institutions) or within the data set itself, either between related variables (e.g., diagnoses and medications are in agreement for a subject) or over time (e.g., temporal trends for a laboratory value appear plausible).

Infrastructure for Assuring Data Quality

Whenever organizations depend solely upon the skill, availability, and integrity of individuals to assure data quality, they place themselves at risk. Levels of skill, ability, and knowledge not only differ from one person to another but may even differ in the same person depending on circumstances (e.g., fatigue can degrade the performance of a skilled operator). Further, in the absence of clear and uniform procedures and standards, different persons will perform tasks in different ways; and while free expression is honored in artistic pursuits, it is not desirable when operationalizing research. A data quality assurance infrastructure provides crucial guidance and structure for humans who work with research data. Simply put, it assures that an organization will consistently produce the required level of data quality. The following criteria are commonly assessed in pre-award site visits and audits for clinical studies. It is no surprise that they comprise a system for assuring data quality.

1. *Organizational consensus regarding the required level of data quality, informed by an understanding of the cost of achieving it and the consequences of failing to achieve it.*

Because the leaders of organizations or clinical trials are not typically data quality professionals, informaticists, or statisticians, data quality-related information, i.e., needs and impacts of not meeting them, may need to be communicated to leadership in a manner that can be acted upon, for example, a draft policy for approval. Where organizations exhibit inadequate support data quality, it may be because this critical information has not been conveyed to leadership in a compelling way that demonstrates the need, the associated costs, and the benefits. Organizations be they companies performing data collection and management services, clinical research networks, or individual labs should work according to a data quality policy.

2. *Appropriate tools for supporting the collection and management of data.*

Although specialized devices and software are of themselves neither necessary nor sufficient for producing quality data, their presence is often perceived as representing rigor or important capability. Specialized tools often automate workflow and enforce controls on the collection and processing of data. Controls built into software are referred to as technical controls. These features can potentially increase efficiency, accuracy, and adherence to procedures by eliminating the variance associated with manual steps and options; for these reasons, data managed using automated systems are often perceived to be of higher quality. Where specialized software with these technical controls is not available, custom programming can be done to create them in available software. Other types of controls are managerial and procedural controls. These use policies, manuals of operations, and work procedures to assure consistency and quality. It is worth emphasizing that high-quality data can be achieved without specialized systems through the use of managerial and procedural controls; however, doing so is

subject to human variability and often entails more highly qualified staff and additional costly manual checking and review. Where specialized technical controls are not in place, depending on the quality needed, their function may need to be developed or addressed through procedural controls.

3. *Design of processes capable of assuring data quality.*

Likened to mass customization, in clinical research, scientific differences in studies and circumstances of management by independent research groups drive variation in data collection and processing. Because each study may use different data collection and management processes, the design and assessment of such processes is an important skill in applied clinical research informatics. The first step in matching a process to a project is to understand how the planned processes, including any facilitative software, perform with respect to data quality dimensions. For example, it is common practice for some companies to send a clinical trial monitor to sites to review data prior to data processing; thus, data may wait for a month or more prior to further processing. Where data are needed for interim safety monitoring, processes with such delays are most likely not capable of meeting timeliness requirements.

Designing and using capable processes is a main component of error prevention. For this reason, clinical research informaticists must be able to anticipate error sources and types and ascertain which errors are preventable, detectable, and correctable and the best methods for doing so. Processes should then be designed to include error mitigation, detection, and correction. Process control with respect to data quality involves ongoing measurement of data quality dimensions such as accuracy, completeness, and timeliness, plus taking corrective action when actionable issues are identified. A very good series of statistical process control books has been published by Donald Wheeler. Several articles have been published on SPC applications in clinical research [76–81].

4. *Documented standard operating procedures (SOPs) are required by FDA regulation and in most research contracts.*

The complete data collection and management process should be documented prior to system development and data collection. The importance of SOPs is underscored by the fact that documented work procedures are mandated by the International Standards Organization (ISO) quality system standards. Variations in approaches to documenting procedures are common, but the essential requirement is that each process through which data pass should be documented in such a way that the published data tables and listings can be traced back to the raw data [10]. Differences between the scientific and operational aspects of clinical research projects often necessitate multiple levels of documentation, for example, a standard procedure level that applies across studies, coupled with a project-specific level of procedural documentation that pertains to individual studies or groups of similar studies. Further, because organizations, regulations, and practices change, process documentation should be maintained in the context of a regular review and approval cycle.

5. *Personnel management infrastructure, job descriptions, review of and feedback on employee performance, and procedures for managing performance.*

Written job descriptions generally include minimum qualifications and experience, a detailed list of job responsibilities, and reporting structure. These descriptions help the candidate as well as the hiring manager(s) assess a person's suitability for a job. In addition, they help organizations communicate expectations and maintain performance standards for a given position. Appropriate data quality assurance infrastructure also includes regular review of employees' work and a means of providing meaningful and actionable feedback to employees. If management is nonexistent or incapable of reviewing employees' work and providing oversight and technical guidance, a key component of the quality assurance infrastructure is absent. Managers should also identify and define both good and inadequate performance, and there should be organizational procedures for encouraging the former and correcting the latter. While these concerns may sound more appropriate for a business office, personnel management infrastructure is crucial to data quality in clinical research because even with continuing technological development, humans still perform all of the design, and some of the data collection and processing and human performance directly affect data quality.

6. *Project management in clinical research informatics begins with understanding the basic data-related requirements of a study, i.e., the data deliverables, associated costs, the necessary levels of quality, and the amount of time required or available.*

Project management also includes planning to meet requirements as well as ongoing tracking, assessment, and reporting of status with respect to targets. Project management profoundly affects data quality; for example, good planning and forecasting make the necessary resources and time for a given project transparent. Keeping a project on schedule eliminates (or at least mitigates) pressure to rush or cut corners and often results in employees who feel less harassed or fatigued.

Together, these six structural components form a quality system for the collection and management of data in clinical research.

Data Governance

The organizational resources, processes, and policies that comprise a data quality program described in previous sections are often a component of a comprehensive data governance management structure. The need for and development of formal data governance management structures arose from the recognition by business leaders that enterprise data management was critical to the success of both strategic and operational objectives [82]. Formal data governance programs encompass more than just data quality oversight, such as enterprise metadata management, data infrastructures, and business analytics/business intelligence functions [83]. Strong

data governance ensures that the substantial investments made in collecting, managing, and using data are maximized and effective. In the research setting, the resulting data are the key products of what often involves millions of dollars, thousands of person-hours, and hundreds of research subjects' willingness to participate in generating new knowledge. When data are seen as perhaps the most expensive investment and the most critical lasting asset of a research project, data governance oversight structures become a central component of the research effort.

The vast majority of the data governance literature is written in the context of corporations setting policies, procedures, and infrastructures around data collected and used in the course of business. These are often long-term programs supporting organizational goals. Clinical research offers new challenges to data governance in that clinical research is project based and shorter term. The same needs exist and are met through study-specific and organizational SOPs and organizational infrastructure as articulated in the quality management system components above. Many of the same activities occur and similar infrastructure components exist. For example, controlled terminology for coding medications and adverse events exist in clinical research and in data governance parlance would be referred to as "reference data." Further, metadata critical to any study is variously managed on organizational and study-specific bases in clinical research. And the provenance articulated in data governance circles is managed at the data value and data element levels and referred to in clinical studies as traceability. In addition, studies based on secondary use of existing data rely on these and other features of data governance in the settings where they were originally collected. It is now expected that this critical information describing the collection and processing of the data whether from organizational data governance or study data management is brought forward and made available with data from clinical investigations. Thus, data governance whether in the context of a single study, a research quality management system, or institutional data governance is a key mechanism by which we assure that data are findable, accessible, interoperable, and reusable (FAIR).

Focusing specifically on data quality oversight features, a data governance program sets the core policies, procedures, metrics, and monitoring methods that will be used by the research team. Policies set the overall "rules of the road" that describe the data quality goals of the program and acceptable or expected means for achieving them. Procedures describe how policies are to be executed by all participating team members; if effective, they will achieve the desired policy goals. Metrics and monitoring methods provide quantitative insights at a frequency where deviations from desired goals can be detected as early as possible so that corrective or alternative procedures can be put into place. Since resources are usually constrained, an effective data governance structure aligns limited resources to those areas of data quality that are considered most impactful or data sources and processes presenting the most risk to human subject protection and study results. For example, data that can only be collected once or are collected as part of a high-risk intervention should be carefully scrutinized to detect any issues in data quality as quickly as possible, whereas data drawn from a pre-existing retrospective data source that could be re-queried may be subjected to less intensive data quality oversight (but not to less data quality assessment).

Several references [84–87] describe data quality or data governance maturity models based on the principles in the Software Engineering Institute (SEI) Capability Maturity Model (now the Capability Maturity Model Integration, CMMITM, including quality management of both products and services). There are many such models in the grey literature from practitioners, consultants, and software vendors. The purpose of maturity models is to highlight the evolution of data quality governance from an initial “ad hoc” set of uncoordinated activities, to a state of highly optimized, fully integrated embedded routine processes. Such models are applicable to all organizations reliant upon data and especially those responsible for data collection and management for clinical research. While corporate and government organizations extend their authority to all conducted studies, academic institutions tend to differ to individual investigators running studies, thus pushing the responsibility and burden of quality management to individual investigators. As documented in a recent review of data management plan requirements, funders only sporadically state requirements for data quality management [88].

For a specific research project or research data network, data quality governance structures should be incorporated into the original project charter or plan. Putting oversight structures in place at the very beginning ensures that the data technologies and data monitoring programs are put into place at the time data management systems are put into production. Once data collection has started, it is difficult to make changes due to concerns about the impact of any data management changes on the comparability of data elements collected or managed under different circumstances. These concerns notwithstanding, it is both a scientific and ethical imperative to ensure that the investments made in executing the study will yield maximum value that only comes with data that have been collected and managed under well-managed data governance.

Impact of Data Quality on Research Results

In most clinical research, the goal is to answer a scientific question. This is often done through inferential statistics. Unfortunately, a “one size fits all” data quality acceptance criterion is not possible because statistical tests vary in their robustness to data errors. Further, the impact on the statistical test depends on the variable in which the errors occur and the frequency and extent of the errors. Further still, data that are of acceptable quality for one use may not be acceptable for another, i.e., the “fitness for use” aspect addressed earlier. It is for these reasons that regulators and often even sponsors cannot set a data quality minimum standard or an “error rate threshold.”

What we can say is that data errors, measurement variability, incompleteness, and delays directly impact the statistical tests when they increase variability, potentially decreasing power. The undesirable scenario of data error increasing variability is shown conceptually in Fig. 11.4; added variability makes it more difficult to tell if two distributions (i.e., a treatment and a control group) are different. Data error rates reported in the literature are well within ranges shown to cause power drops or necessitate increases in sample size in order to preserve statistical power [89, 90].

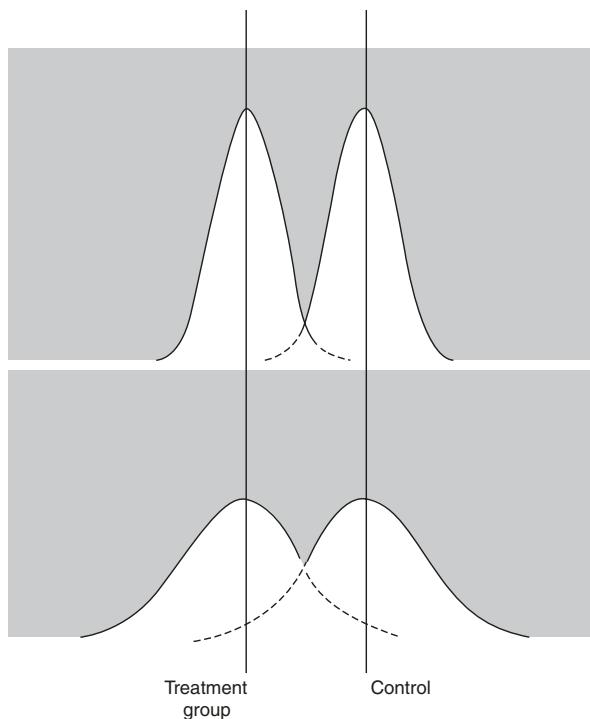


Fig. 11.4 Effect of adding variability. The top two distributions have less variability (are narrower) than the bottom two, making it easier to tell them apart both visually and statistically

While it is true that sample size estimates are based on data that also have errors, i.e., the sample size accounts for some base level of variability, data errors have been shown to change p -values [36] and attenuate correlation coefficients to the null [91–93] (i.e., for trials that fail to reject the null hypothesis, data errors rather than a true lack of effect could be responsible) [94]. However, data errors do not always cause these. Thus, the National Academy of Sciences definition of quality data is data that support the same conclusions as error-free data [1].

In the context of large data error rates adding variability, a researcher must choose either to (1) accept power loss, risking an incorrect indication toward the null hypothesis due to data error, or (2) undertake the expense of measuring the error rate and possibly also the expense of increasing the sample size accordingly to maintain the original desired statistical power [67, 90, 93]. The adverse impact of data errors has also been demonstrated in other secondary data uses such as registries and performance measures [95–101]. Data error can also indicate or be a source of bias in a clinical study. Thus, whether or not data are of acceptable quality for a given analysis is a question to be assessed by the study statistician according to potential impact on the analysis. The assessment should be based on measured error and completeness rates and include description and categorization of root causes so that randomness of the errors can be assessed.

Summary

The following important points apply to data and information collected and managed in clinical research: (1) errors occur naturally, (2) sources of error are numerous and often too numerous to prospectively enumerate and prevent (thus data quality assessment and control are usually required), (3) some errors can be prevented, (4) some errors can be detected, and (5) some errors can be corrected. The sets in three to five do not completely overlap. At the same time, there are errors that cannot be prevented, detected, or corrected (e.g., a study subject who deliberately provides an inaccurate answer on a questionnaire). Errors exist in all data sets, and it is foolish to assume that any collection of data is error-free. While higher quality data are often associated with overall savings, preventing, detecting, and correcting errors are associated with additional or redistributed costs.

The skilled practitioner possesses knowledge of error sources and ability to identify, design, implement, and evaluate methods for error prevention, mitigation, detection, and correction to clinical studies. Further, the skilled practitioner applies this knowledge to design clinical research data collection and management processes to provide the needed quality at an acceptable cost or to identify cases where doing so is not possible. Achieving and maintaining data quality in clinical research is a complex undertaking. If data quality is to be maintained, it must also be measured and acted upon throughout the course of the research project.

There is widespread agreement that the validity of clinical research rests on a foundation of data. However, there is limited research to guide data collection and processing practice. The many unanswered questions, if thoughtfully addressed, can help investigators and research teams balance costs, time, and quality while assuring scientific validity.

References

1. Davis JR, Nolan VP, Woodcock J, Estabrook EW, editors. Assuring data quality and validity in clinical trials for regulatory decision making, Institute of Medicine Workshop report. Roundtable on research and development of drugs, biologics, and medical devices. Washington, DC: National Academy Press; 1999. http://books.nap.edu/openbook.php?record_id=9623&page=R1. Accessed 6 July 2009.
2. Deming WE, Geoffrey L. On sample inspection in the processing of census returns. *J Am Stat Assoc.* 1941;36:351–60.
3. Deming WE, Tepping BJ, Geoffrey L. Errors in card punching. *J Am Stat Assoc.* 1942;37:525–36.
4. Donabedian A. A guide to medical care administration, Medical care appraisal – quality and utilization, vol. 2. New York: American Public Health Association; 1969. p. 176.
5. Arndt S, Tyrell G, Woolson RF, Flaum M, Andreasen NC. Effects of errors in a multicenter medical study: preventing misinterpreted data. *J Psychiatr Res.* 1994;28:447–59.
6. Lee YW, Pipino LL, Wang RY, Funk JD. Journey to data quality. Reprint ed. Cambridge, MA: MIT Press; 2009.
7. Weber GM, Mandl KD, Kohane IS. Finding the missing link for big biomedical data. *JAMA.* 2014;311(24):2479–80.

8. Steinhubl SR, Muse ED, Topol EJ. The emerging field of mobile health. *Sci Transl Med.* 2015;7(283):283rv3.
9. Friedman CP. A “fundamental theorem” of biomedical informatics. *J Am Med Inform Assoc.* 2009;16(2):169–70. <https://doi.org/10.1197/jamia.M3092>. Epub 2008 Dec 11.
10. United States Department of Health and Human Services (HHS), E6(R2) Good Clinical Practice: Integrated Addendum to ICH E6(R1) Guidance for Industry, OMB Control No. 0910-0843 March 2018. Available from: <https://www.fda.gov/downloads/Drugs/Guidances/UCM464506.pdf>.
11. International Organization for Standardization (ISO). Data quality – Part 2: Vocabulary ISO 8000-2:2017.
12. Reprinted with permission from Data Gone Awry, DataBasics, vol 13, no 3, Fall. 2007. Society for Clinical Data Management. Available from <http://www.scdm.org>.
13. Nagurney JT, Brown DF, Sane S, Weiner JB, Wang AC, Chang Y. The accuracy and completeness of data collected by prospective and retrospective methods. *Acad Emerg Med.* 2005;12:884–95.
14. Feinstein AR, Pritchett JA, Schimpff CR. The epidemiology of cancer therapy. 3. The management of imperfect data. *Arch Intern Med.* 1969;123:448–61.
15. Reason J. Human error. Cambridge, UK: Cambridge University Press; 1990.
16. Nahm M, Dziem G, Fendt K, Freeman L, Masi J, Ponce Z. Data quality survey results. *Data Basics.* 2004;10:7.
17. Schuyt ML, Engel T. A review of the source document verification process in clinical trials. *Drug Info J.* 1999;33:789–97.
18. Batini C, Catarci T, Scannapieco M. A survey of data quality issues in cooperative information systems. In: 23rd international conference on conceptual modeling (ER 2004), Shanghai; 2004.
19. Tayi GK, Ballou DP. Examining data quality. *Commun ACM.* 1998;41:4.
20. Redman TC. Data quality for the information age. Boston: Artech House; 1996.
21. Wand Y, Wang R. Anchoring data quality dimensions in ontological foundations. *Commun ACM.* 1996;39:10.
22. Wang R, Strong D. Beyond accuracy: what data quality means to data consumers. *J Manag Inf Syst.* 1996;12:30.
23. Batini C, Scannapieco M. Data quality concepts, methodologies and techniques. Berlin: Springer; 2006.
24. Wyatt J. Acquisition and use of clinical data for audit and research. *J Eval Clin Pract.* 1995;1:15–27.
25. U.S. Food and Drug Administration. In: Services DoHaH, editor. Guidance for industry. Computerized systems used in clinical trials. Rockville: U.S. Food and Drug Administration; 2007.
26. Arts DG, De Keizer NF, Scheffer GJ. Defining and improving data quality in medical registries: a literature review, case study, and generic framework. *J Am Med Inform Assoc.* 2002;9:600–11.
27. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *J Am Med Inform Assoc.* 2013;20:144–51. <https://doi.org/10.1136/amiajnl-2011-000681>.
28. GCP Inspectors Working Group European Medicines Agency (EMA). Reflection paper on expectations for electronic source data and data transcribed to electronic data collection tools in clinical trials. EMA/INS/GCP 454280/2010, 9 June 2010.
29. Kahn MG, Callahan TJ, Barnard J, Bauck AE, Brown J, Davidson BN, Estiri H, Goerg C, Holte E, Johnson SG, Liaw S-T, Hamilton-Lopez M, Meeker D, Ong TC, Ryan P, Shang N, Weiskopf NG, Weng C, Zozus MN, Schilling L. A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. eGEMS (Generating Evid Methods Improve Patient Outcomes) [Internet]. 2016;4(1):1244. Sep 11 [cited 2016 Sep 12]. Available from: <http://repository.edm-forum.org/egems/vol4/iss1/18>.

30. Callahan TJ, Bauck AE, Bertoch D, Brown J, Khare R, Ryan PB, Staab J, Zozus MN, Kahn MG. A comparison of data quality assessment checks in six data sharing networks. eGEMS (Generating Evid Methods Improve Patient Outcomes) [Internet]. 2017;5(1):8. Jun 12 [cited 2017 Jun 15]. Available from: <http://repository.edm-forum.org/egems/vol5/iss1/8>.
31. Estiri H, Stephens K. DQe-v: a database-agnostic framework for exploring variability in electronic health record data across time and site location. eGEMS (Generating Evid Methods Improve Patient Outcomes) [Internet]. 2017;5(1):3. May 10 [cited 2017 Jul 30]. Available from: <http://repository.edm-forum.org/egems/vol5/iss1/3>.
32. Kahn MG, Brown JS, Chun AT, Davidson BN, Meeker D, Ryan PB, Schilling LM, Weiskopf NG, Williams AE, Zozus MN. Transparent reporting of data quality in distributed data networks. eGEMS (Generating Evid Methods Improve Patient Outcomes). 2015;3(1):7. <https://doi.org/10.13063/2327-9214.1052>. Available at: <http://repository.academyhealth.org/egems/vol3/iss1/7>.
33. Zozus MN, Lazarov A, Smith L, Breen T, Krikorian S, Zbyszewski P, Knoll K, Jendrasek D, Perrin D, Zambas D, Williams T, Pieper C. Analysis of professional competencies for the clinical research data management profession: implications for training and professional certification. *JAMIA*. 2017;24:737–45.
34. (CDISC) CDISC. The protocol representation model version 1.0 draft for public comment: CDISC; 2009. p. 96. Available from <http://www.cdisc.org>.
35. Jacobs M, Studer L. Forms design II: the course for paper and electronic forms. Cleveland: Ameritype & Art; 1991.
36. Eisenstein EL, Lemons PW, Tardiff BE, Schulman KA, Jolly MK, Califf RM. Reducing the costs of phase III cardiovascular clinical trials. *Am Heart J*. 2005;9:482–8.
37. Eisenstein EL, Collins R, Cracknell BS, et al. Sensible approaches for reducing clinical trial costs. *Clin Trials*. 2008;5:75–84.
38. Galešić M. Effects of questionnaire length on response rates: review of findings and guidelines for future research. 2002. http://mrav.ffzg.hr/mirta/Galesic_handout_GOR2002.pdf. Accessed 29 Dec 2009.
39. Roszkowski MJ, Bean AG. Believe it or not! Longer questionnaires have lower response rates. *J Bus Psychol*. 1990;4:495–509.
40. Edwards P, Roberts I, Clarke M, DiGuiseppi C, Pratap S, Wentz R, Kwan I. Increasing response rates to postal questionnaires systematic review. *Br Med J*. 2002;324:1183.
41. Wickens CD, Hollands JG, Parasuraman R. Engineering psychology and human performance. 4th ed. New York: Routledge; 2016.
42. Stevens SS. On the theory of scales of measurement. *Science*. 1946;103:677–80.
43. Allison JJ, Wall TC, Spettell CM, et al. The art and science of chart review. *Jt Comm J Qual Improv*. 2000;26:115–36.
44. Banks NJ. Designing medical record abstraction forms. *Int J Qual Health Care*. 1998;10:163–7.
45. Engel L, Henderson C, Fergenbaum J, Interrater A. Reliability of abstracting medical-related information medical record review conduction model for improving. *Eval Health Prof*. 2009;32:281.
46. Cunningham R, Sarfati D, Hill S, Kenwright D. An audit of colon cancer data on the New Zealand cancer registry. *N Z Med J*. 2008;121(1279):46–56.
47. Fritz A. The SEER program's commitment to data quality. *J Registry Manag*. 2001;28(1):35–40.
48. German RR, Wike JM, Wolf HJ, et al. Quality of cancer registry data: findings from CDC-NPCR's breast, colon, and prostate cancer data quality and patterns of care study. *J Registry Manag*. 2008;35(2):67–74.
49. Herrmann N, Cayten CG, Senior J, Staroscik R, Walsh S, Woll M. Interobserver and intraobserver reliability in the collection of emergency medical services data. *Health Serv Res*. 1980;15(2):127–43.
50. Pan L, Fergusson D, Schweitzer I, Hebert PC. Ensuring high accuracy of data abstracted from patient charts: the use of a standardized medical record as a training tool. *J Clin Epidemiol*. 2005;58(9):918–23.

51. Reeves MJ, Mullard AJ, Wehner S. Inter-rater reliability of data elements from a prototype of the Paul Coverdell National Acute Stroke Registry. *BMC Neurol.* 2008;8:19.
52. Scherer R, Zhu Q, Langenberg P, Feldon S, Kelman S, Dickersin K. Comparison of information obtained by operative note abstraction with that recorded on a standardized data collection form. *Surgery.* 2003;133(3):324–30.
53. Stange KC, Zyzanski SJ, Smith TF, et al. How valid are medical records and patient questionnaires for physician profiling and health services research? A comparison with direct observation of patients visits. *Med Care.* 1998;36(6):851–67.
54. Thoburn KK, German RR, Lewis M, Nichols PJ, Ahmed F, Jackson-Thompson J. Case completeness and data accuracy in the centers for disease control and prevention's national program of cancer registries. *Cancer.* 2007;109(8):1607–16.
55. To T, Estrabillo E, Wang C, Cicuttu L. Examining intra-rater and inter-rater response agreement: a medical chart abstraction study of a community-based asthma care program. *BMC Med Res Methodol.* 2008;8:29.
56. Yawn BP, Wollan P. Interrater reliability: completing the methods description in medical records review studies. *Am J Epidemiol.* 2005;161(10):974–7.
57. La France BH, Heisel AD, Beatty MJ. A test of the cognitive load hypothesis: investigating the impact of number of nonverbal cues coded and length of coding session on observer accuracy. *Commun Rep.* 2007;20:11–23.
58. Zozus MN. The data book: collection and management of research data. Taylor & Francis/CRC Press Catalog #: K26788, ISBN: 978-1-4987-4224-5.
59. Helms R. Redundancy: an important data forms/design data collection principle. In: Proceedings Stat computing section, Alexandria; 1981. p. 233–7.
60. Helms R. Data quality issues in electronic data capture. *Drug Inf J.* 2001;35:827–37.
61. U.S. Food and Drug Administration regulations. Title 21 CFR Part 58. 2011. Available from <http://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfrsearch.cfm?cfrpart=58>. Accessed Aug 2011.
62. Nahm ML, Pieper CF, Cunningham MM. Quantifying data quality for clinical trials using electronic data capture. *PLoS One.* 2008;3(8):e3049.
63. Winchell T. The mystery of source documentation. *SOCRA Source* 62. 2009. Available from <http://www.socra.org/>.
64. Nahm M. Data accuracy in medical record abstraction. Doctoral Dissertation, University of Texas at Houston, School of Biomedical Informatics, Houston, May 6, 2010.
65. Zozus MN, Pieper C, Johnson CM, Johnson TR, Franklin A, Smith J, et al. Factors affecting accuracy of data abstracted from medical records. *PLoS One.* 2015;10(10):e0138649.
66. SCDM. Good clinical data management practices. <http://www.scdm.org>. Society for Clinical Data Management; 2010. Available from <http://www.scdm.org>.
67. Rostami R, Nahm M, Pieper CF. What can we learn from a decade of database audits? The Duke Clinical Research Institute experience, 1997–2006. *Clin Trials.* 2009;6(2):141–50.
68. Stellman SD. The case of the missing eights an object lesson in data quality assurance. *Am J Epidemiol.* 1989;129(4):857–60. <https://doi.org/10.1093/oxfordjournals.aje.a115200>.
69. Hogan WR1, Wagner MM. Accuracy of data in computer-based patient records. *J Am Med Inform Assoc.* 1997;4(5):342–55.
70. Thiru K, Hassey A, Sullivan F. Systematic review of scope and quality of electronic patient record data in primary care. *BMJ.* 2003;326(7398):1070. Review.
71. Chan KS, Fowles JB, Weiner JP. Review: electronic health records and the reliability and validity of quality measures: a review of the literature. *Med Care Res Rev.* 2010;67(5):503–27. <https://doi.org/10.1177/1077558709359007>.
72. Observational Health Data Sciences and Informatics. OHDSI Observational Medical Outcomes Partnership (OMOP) Common Data Model. <https://www.ohdsi.org/>. Accessed 29 May 2018.
73. The National Patient-Centered Clinical Research Network (PCORnet). Common data model v3.0. https://pcornetcommons.org/resource_item/pcornet-common-data-model-cdm-specification-version-3-0/. Accessed 1 Feb 2016.

74. Kahn MG, Raebel MA, Glanz JM, Riedlinger K, Steiner JF. A pragmatic framework for single-site and multisite data quality assessment in electronic health record-based clinical research. *Med Care.* 2012;50(suppl):S21–9. <https://doi.org/10.1097/MLR.0b013e318257dd67>.
75. Weiskopf NG, Hripcak G, Swaminathan S, Weng C. Defining and measuring completeness of electronic health records for secondary use. *J Biomed Inform.* 2013;46:830–6. <https://doi.org/10.1016/j.jbi.2013.06.010>.
76. Svolba G, Bauer P. Statistical quality control in clinical trials. *Control Clin Trials.* 1999;20(6):519–30.
77. Chilappagari S, Kulkarni A, Bolick-Aldrich S, Huang Y, Aldrich TE. A statistical process control method to monitor completeness of central cancer registry reporting data. *J Registry Manag.* 2002;29(4):121–7.
78. Chiu D, Guillaud M, Cox D, Follen M, MacAulay C. Quality assurance system using statistical process control: an implementation for image cytometry. *Cell Oncol.* 2004;26(3):101–17.
79. McNees P, Dow KH, Loerzel VW. Application of the CuSum technique to evaluate changes in recruitment strategies. *Nurs Res.* 2005;54(6):399–405.
80. Baigent C, Harrell FE, Buyse M, Emberson JR, Altman DG. Ensuring trial validity by data quality assurance and diversification of monitoring methods. *Clin Trials.* 2008;5(1):49–55.
81. Matheny ME, Morrow DA, Ohno-Machado L, Cannon CP, Sabatine MS, Resnic FS. Validation of an automated safety surveillance system with prospective, randomized trial data. *Med Decis Mak.* 2009;29(2):247–56.
82. McGilvray D. Executing data quality projects: ten steps to quality data and trusted information. 1st ed. Amsterdam: Morgan Kaufmann; 2008. 352 p.
83. Ladley J. Data governance: how to design, deploy and sustain an effective data governance program. 1st ed. Waltham: Morgan Kaufmann; 2012. 264 p.
84. Loshin D. The practitioner's guide to data quality improvement. 1st ed. Burlington: Morgan Kaufmann; 2010. 432 p.
85. Baskarada S. IQM-CMM: information quality management capability maturity model. Germany: Vieweg and Teubner; 2010.
86. Capability Maturity Model Integration (CMMITM) Institute, Data Management maturity model, CMMI Institute 2014.
87. Stanford University. Stanford data governance maturity model. Accessed 12 May 2018. Available from <http://web.stanford.edu/dept/pres-provost/irds/dg/files/StanfordDataGovernanceMaturityModel.pdf>.
88. Williams M, Bagwell J, Zozus M. Data management plans, the missing perspective. *J Biomed Inform.* 2017;71:130–42.
89. Freedman LS, Schatzkin A, Wax Y. The impact of dietary measurement error on planning sample size required in a cohort study. *Am J Epidemiol.* 1990;132:1185–95.
90. Perkins DO, Wyatt RJ, Bartko JJ. Penny-wise and pound-foolish: the impact of measurement error on sample size requirements in clinical trials. *Biol Psychiatry.* 2007;47:762–6.
91. Mullooly JP. The effects of data entry error: an analysis of partial verification. *Comput Biomed Res.* 1990;23:259–67.
92. Liu K. Measurement error and its impact on partial correlation and multiple linear regression analyses. *Am J Epidemiol.* 1988;127:864–74.
93. Stepnowsky CJ Jr, Berry C, Dimsdale JE. The effect of measurement unreliability on sleep and respiratory variables. *Sleep.* 2004;27:990–5.
94. Myer L, Morroni C, Link BG. Impact of measurement error in the study of sexually transmitted infections. *Sex Transm Infect.* 2004;80(318–323):328.
95. Williams SC, Watt A, Schmaltz SP, Koss RG, Loeb JM. Assessing the reliability of standardized performance indicators. *Int J Qual Health Care.* 2006;18:246–55.
96. Watt A, Williams S, Lee K, Robertson J, Koss RG, Loeb JM. Keen eye on core measures. Joint commission data quality study offers insights into data collection, abstracting processes. *J AHIMA.* 2003;74:20–5; quiz 27–8.

97. US Government Accountability Office. Hospital quality data: CMS needs more rigorous methods to ensure reliability of publicly released data. In: Office UGA, editor. Washington, DC; 2006. www.gao.gov/new.items/d0654.pdf.
98. Braun BI, Kritchevsky SB, Kusek L, et al. Comparing bloodstream infection rates: the effect of indicator specifications in the evaluation of processes and indicators in infection control (EPIC) study. *Infect Control Hosp Epidemiol*. 2006;27:14–22.
99. Jacobs R, Goddard M, Smith PC. How robust are hospital ranks based on composite performance measures? *Med Care*. 2005;43:1177–84.
100. Pagel C, Gallivan S. Exploring consequences on mortality estimates of errors in clinical databases. *IMA J Manag Math*. 2008;20(4):385–93. <http://imaman.oxfordjournals.org/content/20/4/385.abstract>.
101. Goldhill DR, Sumner A. APACHE II, data accuracy and outcome prediction. *Anaesthesia*. 1998;53:937–43.



Patient-Reported Outcome Data

12

Robert O. Morgan, Kavita R. Sail, and Laura E. Witte

Abstract

This chapter provides a brief introduction to patient-reported outcome measures (PROs), with an emphasis on measure characteristics and the implications for informatics of the use of PROs in clinical research. Because of increased appreciation on behalf of health-care funders and regulatory agencies for actual patient experience, PROs have become recognized as legitimate and attractive endpoints for clinical studies and for comparative effectiveness research. “Patient-reported outcomes” is an internationally recognized umbrella term that includes both single dimension and multidimension measures of symptoms, with the defining characteristic that all information is provided directly by the patient. PROs can be administered in a variety of formats and settings, ranging from face-to-face interaction in clinics to web interfaces to mobile devices (e.g., smart phones). PRO instruments measure one or more aspects of patients’ health status and are especially important when more objective measures of disease outcome are not available. PROs can be used to measure a broad array of health status indicators within the context of widely varying study designs exploring a multitude of diseases. As a result, they need to be well characterized so that they can be identified and used appropriately. The standardization, indexing, access, and implementation of PROs are issues that are particularly relevant to clinical research informatics. In this chapter, we discuss design characteristics of PROs, measurement issues relating to the use of PROs, modes of administration, item and scale development, scale repositories, and item banking.

R. O. Morgan, PhD (✉) · L. E. Witte, MPH

Department of Management, Policy and Community Health, University of Texas Health

Science Center at Houston School of Public Health, Houston, TX, USA

e-mail: Robert.O.Morgan@uth.tmc.edu; Laura.E.Witte@uth.tmc.edu

K. R. Sail, PhD

Health Economics and Outcomes Research, AbbVie Pharmaceuticals,
North Chicago, IL, USA

Keywords

Patient-reported outcome data · Outcome data by patient report · Scales
Assessment methods · Reliability · Validity · Electronic data collection devices
The patient-reported outcome measurement information system

The term *patient-reported outcomes* (PRO) is an umbrella term that includes both single dimension and multidimension measures of symptoms. While there is no standard definition of a PRO, most commonly used definitions are in close agreement. In general, PROs include “...any report of the status of a patient’s health condition that comes directly from the patient, without interpretation of the patient’s response by a clinician or anyone else. The outcome can be measured in absolute terms (e.g., severity of a symptom, sign, or state of a disease) or as a change from a previous measure” [1].

PROs provide information on the patient’s perspective of a disease and its treatment [1] and are especially important when more objective measures of disease outcome are not available. PRO instruments measure one or more aspects of patients’ health status. These can range from purely symptomatic (e.g., pain magnitude) to behaviors (e.g., ability to carry out activities of daily living), to much more complex concepts such as quality of life (QoL), which is considered as a multidomain attribute with physical, psychological, and social components. Consequently, PROs are a large set of patient-assessed measures ranging from single-item (e.g., pain visual analog scale [VAS], global health status) to multi-item tools. In turn, multi-item tools can be monodimensional (e.g., measuring a single dimension such as physical functioning, fatigue, or sexual function) or multidimensional questionnaires. This chapter is intended to provide an overview of patient-reported outcomes measurement. We touch on five main topics in this chapter: design characteristics of PROs, measurement issues, modes of administration, item and scale development, and banking and retrieval of PROs.

Characteristics of Patient-Reported Outcomes

PROs can be classified along multiple dimensions, including the generality of symptoms, specificity of disease or population, and whether patients are reporting experiences or attitudes [1]. The more *specific* a PRO is, the more responsive it is likely to be to changes in health status for the health problem being investigated [2]. In contrast, PROs that assess more *general* states or conditions provide broader information on health and quality of life and are frequently more usable in economic evaluations [3]. They have a greater potential to measure unforeseen effects or side effects of health care, and the results can usually be compared with those for other patient populations. Selection of the appropriate PRO is clearly dependent on the purpose for which it is intended [4, 5]; however, it is generally considered good practice to use both types of PROs where possible [3].

There has been considerable work on the development and assessment of PROs [6, 7]. PROs have been endorsed by the NIH and FDA as credible endpoints for clinical research studies and comparative effectiveness studies. The FDA has

outlined 14 design characteristics for PROs used in clinical trials [1]. These serve as an excellent guide for PROs in general and for choosing the appropriate instruments. A good summary of this guidance is provided by Shields et al. [8]. The recommended FDA design characteristics address:

1. *Concepts being measured*: Any use of a PRO is predicated on clearly understanding what trait or characteristic that measure is designed to capture and whether the PRO is appropriate for the disease and population under study. A list of example traits or characteristics might include overall health status, symptoms and signs, functional status, health perceptions, satisfaction, preference, and adherence.
2. *Number of items*: This is important in terms of response burden and data completeness. PROs can be constructed as single-item measures, as indices with several individual item measures, with multiple items measuring a single construct (e.g., a scale), or as a collection of multiple scales.
3. *Conceptual framework of the instrument*: This represents the conceptual context of the information being gathered, the related concepts, and the relationship of those concepts to a population, treatment, condition, or knowledge domain. Understanding the context is key to assessing the appropriateness of the PRO in a given application.
4. *Medical condition for intended use*: Is the PRO intended for use as a generic measure, or is it disease specific? Along with the target population (characteristic 5), the medical condition being targeted affects the specificity of the population that the data relate to and the utility of the data for making comparisons to other patient populations.
5. *Population for intended use*: Is the PRO intended for use with any individuals, or is it age or gender specific, specific to the patient or caregiver, etc.?
6. *Data collection method*: What mechanism is being used to collect the data? Paper and pencil, a computer, a tablet PC, using web-based systems, interactive voice response, or some other method? This affects the ease and effectiveness of administration within a given situation. Note that “ease” and “effectiveness” are not the same.
7. *Administration mode*: Is the PRO self-administered, interviewer-administered, or administered in another way? As with the data collection method (characteristic 6), this affects the ease and effectiveness of administration. It also affects the scope of the data that can be collected and strongly influences the completeness of the data gathered.
8. *Response options*: This is the way responses are enumerated (e.g., Likert type, true/false, visual analog, etc.). This affects the sensitivity of the PRO questions, that is, will the questions capture the information desired?
9. *Recall period*: Do the PRO questions relate to the patient’s current status, or do they require recall of prior states or experiences? If prior status, the time period over which recall is requested can significantly impact the accuracy of the data, particularly if long recall periods are used.
10. *Scoring*: Does the PRO measure yield a single rating, an index score combining multiple ratings, a profile – multiple uncombined scores, a composite – an

- index, profile or battery, free text information, or some other type of summarization? This will affect the specificity and reliability (reproducibility) of the information collected by the PRO.
11. *Weighting of items or domains*: Do summary scores use equal or variable weighting of items and/or scales? This will reflect the relative importance of the individual items (or scales) on the PRO measure and will affect the sensitivity of the measure to information from items with different weights.
 12. *Format*: What is the text layout, and are there skip patterns, drop-down lists, interactive scales, and so on? As with characteristics 6 and 7, this can affect the ease and effectiveness of administration, as well as the scope and completeness of the data collected.
 13. *Respondent burden*: Are the PRO items cognitively complex? What are the time or effort demands? This directly affects the ability of respondents to provide effective responses to the PRO items or even to complete the PRO measure.
 14. *Translation or cultural adaptation availability*: Are validated, alternative versions for specific patient subgroups available? As with estimates of response burden (characteristic 13), this affects the ability of respondents to provide effective and accurate responses to the PRO items.

Valderas and Alonso [9] provide an alternative classification system for PROs that incorporates many of the same elements presented above. Calvert, Bazeby, et al. [10] provide guidance for CONSORT reporting guidelines for PROs in clinical trials.

Measurement Issues

Comparability of PROs Across Studies and Time

Data that are unreliable or have poor validity can lead to erroneous and nongeneralizable study results through a combination of low statistical power and lack of sensitivity in data analyses, biases in statistical conclusions, and biases in estimates of prevalence and risk [11]. These errors can affect our understanding of therapeutic effectiveness by restricting our ability to detect an intervention's effect and distort our assessments of the epidemiology of medical conditions by biasing our assessment of different subpopulations of patients.

It is widely recognized that measurement properties such as reliability and validity are both sample and purpose dependent [12]. That is, they vary across the populations and purposes for which measures are used. Researchers are most familiar with these issues in the context of measurement with self-report instruments, surveys, or scales. On scales, for example, individual items may differ across populations in terms of how they relate to the underlying constructs being measured, and the constructs themselves may shift across populations. Measures may be affected by differences in demographic characteristics (e.g., age, socioeconomic status, location), illness burden, psychological health, or cultural identity. Consequently, a

scale developed to assess communication ability in Anglo-Americans may not be as effective when used with African- or Hispanic-Americans; a scale may not work as well with individuals raised in a rural setting as with those raised in an urban one, or the properties of a scale developed in a sample of young female patients may not generalize when the scale is used with older males. Similarly, the measurement properties of scales may vary according to how they are used. For example, a measure developed for assessing cross-sectional group differences in health status may be inadequate as an instrument for measuring change over time for a particular individual. When measurement is conducted via survey methodology, these vulnerabilities may be compounded by biased nonresponse to the survey or partial completion of survey items [13].

The need to verify measurement properties extends beyond “traditional” psychometric applications (e.g., reliability or validity of survey or other self-report measures) and beyond the characteristics of the population we are attempting to study. For the US population in general, there are substantial differences among the health-care systems in which individuals seek care. These differences may affect entry into the system (e.g., access), therapeutic decisions (e.g., quality), and availability of endpoints (e.g., outcomes). Thus, measurement and the resulting findings are influenced by features of the health-care system. Attention to measurement quality necessarily includes design issues (e.g., formatting and administration of measurement instruments), settings in which measurement is conducted (e.g., at a physician’s office versus a hospital setting or at home), and the source from which the measures are obtained [13]. Frost et al. provide a good description of evidence required for establishing the reliability and validity of PROs used in randomized clinical trial [14].

Reliability

The reliability of a measure refers to the *stability* or *equivalence* of repeated measurements of the same phenomena within the same patient [15]. In this context, *stability* refers to the consistency of information collected at different points in time, assuming no real changes have occurred. *Equivalence* refers to the consistency of observations or responses given to different observers. One way to visualize reliability is as a “signal-to-noise” ratio. High reliability would be equivalent to a high signal-to-noise ratio (more signal, less noise). Low reliability would be equivalent to a low signal-to-noise ratio (less signal, more noise).

Reliability is generally expressed as a correlation coefficient or a close statistical relative (e.g., kappa coefficients, Cronbach’s alpha, intraclass correlations [ICC]) [16] and is on a scale of 0.00–1.00, where 0.00 reflects the lowest possible reliability (i.e., none) and 1.00 reflects perfect reproducibility or correspondence. In practice, low reliability equates to high variability in measurement. Consequently, measures with low reliability are minimally useful. From a research perspective, highly reliable measures increase the statistical power for a given sample size, enabling statistical significance to be achieved with a smaller sample (i.e., more signal, less noise).

Since the reliability of a measure depends both on the characteristics of the measure and on how it is being used, there is no single way to assess reliability. The most common types of reliability assessments are *test-retest*, *internal consistency*, and *interrater* reliability [15, 17].

Test-retest reliability is estimated by the correlation between responses to the same measure by the same respondent at two different points in time. The presumption is that the correlation between the two measures represents a *lower-bound estimate* on the stability or consistency of the measuring instrument. Clearly, the more transient the construct that is being measured is, the less effective test-retest correlations are as a measure of reliability. Transient personal characteristics, such as physical or mental states, and situational factors, such as changes in the measurement context (e.g., clinic versus home environments or mailed administration versus in-person administration), can have a significant impact on test-retest reliability estimates.

Internal consistency reliability is a variant on test-retest methodology. Internal consistency is used to estimate the level of association among responses by the same respondent to individual items on a multi-item scale assessing a single construct [15]. Under classical test theory, the individual scale items can be presumed to be approximately equivalent measures of the same construct. As such, correlations among items are a form of test-retest reliability, with the correlation among scale items representing an estimate of the reliability of the overall scale. The two most widely used internal consistency estimators are split-half reliability and Cronbach's alpha [17]. Split-half reliability is self-explanatory. Since items are presumed to be interchangeable, the scale items are randomly split into two equal groups, and the subgroup totals are correlated. This correlation, once adjusted for the length of the full scale, is an estimate of the scale's reliability [17]. The more widely used Cronbach's alpha is an extension of this approach.

Internal consistency estimates are fundamentally driven by the number of questions asked to capture the underlying construct (more questions = higher consistency estimates) and the average correlation between the individual questions (higher average correlation = higher consistency estimates).

Interrater reliability is important in situations where multiple interviewers are needed to collect information from a large group of patients, patients in multiple locations, or across multiple staffing shifts. Interrater reliability is estimated by the correlation between measurements on the same respondent obtained by different observers at the same point in time and is used to test the presumption that the interviewers are collecting equivalent data, that is, that the interviewers are interchangeable. For continuous measures, interrater reliability is estimated by a Pearson r (or an intraclass correlation coefficient for more than two interviewers). For categorical measures, interrater reliability is estimated by a kappa coefficient [16, 17].

Validity

The validity of a measure represents the degree of systematic differences between responses to PROs relative to (1) the concept they were intended to assess (*content*

validity), (2) related assessments of the same concept (*criterion validity*), and (3) hypotheses about relationships to other concepts (*construct validity*) [15, 17].

Content validity (or face validity) is the extent to which a measure *adequately represents* the concept of interest. Content validity primarily relies on judgments about whether the measure (or the individual items of a scale) represents the concept that it was chosen to represent (Table 12.1) [16]. Content validity is directly affected by any lack of clarity regarding the domain in the concept being evaluated. Even when the concept being evaluated is clearly defined, failure to thoroughly conduct background research on the concept's definition and measurement may reduce validity.

Criterion validity is the extent to which a PRO predicts or agrees with a criterion indicator of the “*true*” value (gold standard) of the concept of interest [15, 16]. The two principal types of criterion validity are *predictive validity*, where the criterion indicator or indicators are predicted by a PRO measure, and *concurrent validity*, where the PRO measure corresponds to (correlates with) criterion measures of the concept of interest (Table 12.1). Criterion validity is adversely affected by lack of clarity in the measures (either low content or low construct validity) and by response bias, particularly under- or overreporting events due to frequency and/or particularly high or low salience. Criterion validity is also negatively impacted by low reliability (low signal-to-noise ratio), which makes validity difficult to demonstrate.

Construct validity is the extent to which relationships between a PRO and other measures agree with relationships predicted by existing theories or hypotheses (Table 12.1) [15, 17]. Construct validity can be separated into *convergent validity*, where the PRO measure shows *positive* associations with measures of constructs it should be positively related to (i.e., converging with), and *discriminant validity*, where the PRO measure shows *negative* associations with measures of constructs it should be negatively related to (i.e., discriminating from). Construct validity is particularly useful when there are no good criterion measures or gold standards for establishing criterion validity, for example, when the construct measured is abstract (e.g., “pain”). Construct validity is negatively affected by the same things as criterion validity, including low reliability, lack of clarity in defining the construct, and response bias. The ability to demonstrate construct validity can also be hampered by inadequate theory for guiding the specification of hypothesized relationships.

Table 12.1 Methods of computing validity

Methods	Types of validity		
	Content	Criterion	Construct
Literature review	X		
Expert judgment	X		
Sensitivity-specificity analysis		X	
Correlation coefficients		X	X
Known-groups validity			X
Factor analysis			X
Multitrait multimethod			X

From Aday and Cornelius [17], Table 3.3 (p. 64). Reprinted with permission of John Wiley & Sons, Inc

Responsiveness is the extent to which a PRO is sensitive to change in the health construct being measured. That is, does the PRO reflect a change that has occurred, and does it remain stable if there has been no true change? As noted above, PROs that are more specific to a disease, condition, or population or that have a more fine-grained measurement resolution are generally more sensitive to change than are more generic PROs [18]. Although the general concept of responsiveness is straightforward, there is no consensus on the best way to measure it. Revicki, Hays, et al. provide guidance on assessing responsiveness [19], and McDowell provides a summary of different approaches, all of which reflect some form of standardizing the change score [18].

Modes of Administration

Researchers need to consider many factors in deciding the appropriate mode for data collection, including the burden (time, effort, stress, etc.) on the respondent and the cost of administration. Also, researchers need to be aware of the impact of changes in mode of administration on the overall reliability and validity of the resulting data. Common administration modes are presented below.

Personal (Face-to-Face) Administration

Personal or face-to-face administration is recognized as the gold standard among data collection methodologies [20]. Instruments are completed by the interviewer based on what the respondent says, and the interviewer has the opportunity to probe or ask follow-up questions to the respondent. This type of administration is credited for achieving high response rates and better quality of data. Once the administration is initiated, the interviewer builds a rapport or trust with the respondent which generally leads to more accurate responses. This method allows direct observation of the respondents and hence allows for flexibility in the way questions are asked. A skilled interviewer can read people, assess moods, and probe, clarify, rephrase, or restate the question in an alternative manner to the participant. Personal administration can vary from a highly structured set of questions to an unstructured conversation. This type of administration generally yields the highest levels of cooperation and lowest refusal rates; it allows for longer, more complex interviews; the responses are generally of high quality; the administration can be designed to take advantage of the interviewer's presence; and it allows for the use of multiple methodologies in the data collection process [17].

Face-to-face administration has several disadvantages as well. It is resource intensive and usually more costly than other modes of administration, it typically requires a longer data collection period, the interviewer(s) require significant training, and when multiple interviewers are used, correspondence among interviewers needs to be demonstrated and maintained over the data collection period (i.e., interrater reliability) [17, 20].

Telephone Administration

Telephone administration allows more rapid collection of information than face-to-face administration. Like face-to-face administration, it allows for significant personal contact between the respondent and the interviewer. The steps followed for telephone administration are essentially the same as those for face-to-face administration above. Since telephone administration typically does not require in-person interaction, it is usually less expensive than face-to-face administration with a shorter data collection period. It offers many of the same advantages of face-to-face administration listed above while also allowing better control and supervision of interviewers.

Telephone administration carries some of the same disadvantages as well. For example, telephone data collection is usually less expensive than for face-to-face administration but remains more expensive, per completed PRO battery, than for mailed surveys. Further, interviewer training and correspondence remain issues, and telephone administration can be biased against households without telephones, households with unlisted numbers, or households that rely exclusively on cell phones, although methodologies mitigating these biases are becoming more widespread [17, 20]. Since administration is conducted over the phone, it typically does not (or cannot) last as long as face-to-face administration, restricting the number of PRO measures that can be collected. It can also be difficult to administer PRO instruments on sensitive or complex topics.

Mailed Surveys

Mailed surveys are self-administered instruments sent via mail to recipients. This mode of administration is generally lower in cost, per completed PRO instrument, than either face-to-face or telephone administration. Surveys can be administered by a smaller team since no field staff is required and can be effective with populations that are difficult to reach by phone or in person. Mailed surveys also offer respondents flexibility in when and how they choose to complete the instruments.

However, since there is typically little individualized contact with the recipients, at least until late in the data collection process, it can be more difficult to obtain cooperation from the individuals receiving the survey. Since the survey instruments are intended to be self-administered, they typically must be more rigidly structured than in either face-to-face or telephone administration, restricting both the content and the length of the PRO instruments. Further, wording of individual items must be straightforward and easily interpreted, which in turn can increase the time it takes to develop and refine the mailed survey.

According to Dillman [21], the steps needed for achieving acceptable response rates in mailed surveys are:

- A prenotice letter informing the respondent about the survey sent to the respondent prior to sending the actual questionnaire.

- The actual survey packet is sent, including a detailed cover letter explaining the survey and the importance of the respondent participation, as well as any incentive offered to prospective respondents.
- A thank you postcard sent a few weeks later indicating appreciation if response has been sent or hoping that the questionnaire would be completed soon.
- A replacement questionnaire sent to nonrespondents, usually 2 weeks after the reminder postcard, including a second cover letter urging the recipients to respond to the survey.
- A final reminder, sometimes made by telephone (if the telephone numbers are available) or sent through priority mail.

Web Surveys and Email Communication

Web surveys are self-administered surveys accessed through the Internet. Links to the secure survey URLs are often sent to respondents through electronic mail. They are constructed on a website, and the respondent must access the particular website to be able to respond to the survey. The questions are constructed in a fixed format, and there are different programming languages and styles that can be utilized for building a web survey. Web surveys provide the possibility for dynamic interaction between the respondent and the questionnaire [21]. The difficult structural features of questionnaires, such as skip patterns, drop-down boxes for answer choices, instructions for individual questions, and so on, can be easily incorporated in a web survey. Pictures, animations, and video clips can be added to the survey to aid the respondent.

Electronic mail is useful for sending links to web-based self-administered PRO instruments and reminder communications to respondents. The guidelines for survey email communications are [21]:

- As with mailed surveys, it is important to send the respondent a prenotice e-mail message informing the respondent of the survey. The objective of sending a prenotice letter is to leave a positive impression of importance of the survey so that the recipient does not discard the questionnaire upon arrival.
- To help preserve confidentiality and promote a higher response rate, e-mail contacts should be personalized, and the recipient should receive a personalized e-mail, rather than be a part of a list serve.
- When the survey is sent, the cover e-mail should be kept as brief as possible since respondents usually have less attentive reading while reading electronic mail.
- Within the cover e-mail, the participants should be informed of alternative ways to respond such as printing the survey and sending it back.
- Follow-up contacts should follow the same timeline as for mailed surveys. The survey link should be included with any follow-up contact.

E-mail communications and web surveys offer several advantages over mailed surveys. They are usually of lower cost (no paper, postage, mailing, data entry

costs); the time required for implementation is reduced; because of the minimal distribution costs, sample sizes can be much greater and the scope of distribution can be worldwide; and the formatting of the surveys can be complex and interactive, for example, skip patterns and alternative question pathways can be programmed in [21]. New technology and software have made implementation of e-mail and web-based PROs relatively straightforward, including features such as sending patients email reminders to complete PRO questionnaires at predetermined intervals [22].

However, there are significant limitations as well. Not all homes have a computer or e-mail access. Consequently, representative (unbiased) samples are difficult to obtain, and sampling weights are hard to determine. There are also differences in the capabilities of people's computers and software for accessing web surveys and the speed of Internet service providers and line speeds, further limiting the representativeness of samples [21].

Electronic Data Collection Devices/Systems (ePRO)

The emergence of telephone- and web-based data collection has gone hand in hand with the development of interactive devices. There are two main categories of ePRO administration platforms: *voice/auditory devices* and *screen text devices* [23].

Voice auditory systems These systems are often referred to as interactive voice response (IVR) and are usually telephone-based, although Voice over Internet Protocols (VOIP) are increasingly being incorporated into their designs [24, 25]. With these devices, an audio version of the questions and response choices is provided to the respondent. Typically, IVR systems interact with callers via a prerecorded voice question and response system. The advantages of an IVR system include [23]: no additional hardware is required for the respondent, minimum training is necessary for respondent, data are stored directly to the central database, the voice responses can be recorded, low literacy requirements exist for respondents, a combination of voice input and touch-tone keypad selection is accepted to assist the questionnaire completion, and it allows both call generation and call receipt.

Screen text devices Numerous screen text devices exist, including desktop and laptop computers, tablet or touch-screen notebook (and netbook) computers, handheld/palm computers, web-based systems, audiovisual computer-assisted self-interviewing (A-CASI) systems, and mobile devices, including cell phones.

These devices have a number of advantages over more traditional, hard-copy data collection systems. The collection of PROs is fast, accurate, and reliable; time to analysis is reduced; remote monitoring is possible, including access to individual participant-reported information and biometric data from devices such as glucometers, scales, BP monitors, and spirometers; and researchers and staff can communicate securely with study subjects and patients through encrypted messaging systems [23]. These devices can also be developed for specific respondent

populations. As an example, an ePRO platform designed to support patients with visual impairments is available via web entry and mobile app from IBM Clinical Development [26]. However, access to these devices is neither universal nor necessarily representative of the populations of interest [21].

Desktop, laptop, and touch-screen tablet computers These systems are usually fully functional computers, and they offer more screen space than other screen-based options. Consequently, a major advantage of such systems is that the question and the response text can be presented in varying font sizes and languages. Stand-alone desktop systems may be limited in mobility. Touch-screen systems have a touch-sensitive monitor screen and may be used with or without a keyboard or a mouse [23]. Many ePRO systems are compatible with multiple technologies; for example, VitalHealth's QuestLink and Acceliant's ePRO platform are compatible with web, mobile, smartphone, and tablets [27, 28].

Audiovisual computer-assisted self-interviewing (A-CASI) systems This system combines IVR and screen text. The questionnaire is presented on a computer monitor and is accompanied by an audible presentation of questions and responses. These devices may be helpful for evaluating special populations [23].

Mobile devices Another method of obtaining patient-reported data is through the use of mobile devices or cell phones (MPRO, mobile patient-reported outcomes). This technique utilizes web and mobile technology to enhance the collection and management of patient-reported data. For example, Medidata offers a mobile app that allows patients to enter PRO information from their mobile devices into the Medidata data cloud [29]. The rapid growth of “smartphones” with sophisticated web interfaces, or tablet PCs with cellular interfaces, is blurring the lines between tablet, handheld, and voice-operated systems. Newly developed digital pen and paper technologies use tiny cameras in the tips of pens along with special paper with unique dot patterns to create electronic replicas of handwritten pages. Researchers use the forms in the same way they would an ordinary paper form and then upload the information to a study database. While the actual integration of these devices as data collection instruments is still in progress, this class of mobile devices holds substantial potential for broad application [30]. Notably, use of cellular networks also permits real-time geotracking of the devices, allowing PROs to be combined with specific location information. These data are particularly useful in social network analysis and the evaluation of lifestyle interventions [31–33].

Item and Scale Development

Although technology can significantly ease implementation of PRO measures, actually developing items and scales from scratch can be a laborious and time-consuming activity, with no guarantee of a well-performing scale when finished. It is frequently better (and easier) to use an existing, validated scale, assuming that it adequately

meets the needs of the research study. Next best is an existing scale that comes close to meeting the requirements of the study but needs some modification. Note that modifying an instrument, or using an existing instrument in a modified context, may still necessitate a reevaluation of the instrument's properties. Steps for modifying a scale are described below, after the *guidelines for item and scale development*.

Although the work required to develop a new scale is significant (and almost always underestimated), there is plenty of guidance available. An extensive literature documents methods for developing and modifying scales and scale items [1, 15, 17, 21]. The following is a summary of the key guidelines presented by DeVellis [15]:

1. *Determine clearly what it is to be measured:* Scale development needs to be based in a clear conceptual framework. *This is the most important step in developing a scale.* Everything follows from this, so it is crucial to spend the time necessary for clarifying the constructs to measure. This includes clearly identifying the scope of the content, the target population, the desired measurement setting, the method(s) for administration, and the period of recall over which subjects will report.
2. *Generate an item pool:* Following from step 1, items must reflect the constructs to be measured. Create a large number of items to reflect the concept. This is the *item pool*. At this stage, emphasize quantity over quality; redundancy is fine. Then, eliminate the poorly worded or less clear items. These would include those that are cognitively complex (too long, hard to read or interpret), double-barreled items (two items masquerading as one), and items with ambiguous wording.
3. *Determine the format for measurement:* What type of responses are desired? Do they include binary (e.g., yes/no) or ordinal categories (e.g., Likert scale-type responses) or a response on a continuous scale (e.g., a visual analog scale)? Next, assign descriptors for the response options; these are also called *item anchors* (e.g., “strongly disagree” to “strongly agree”). These provide the framing for the responses. They need to be clear and to match the item wording. For example, anchors for attitudinal items would clearly be different from anchors for frequency items. *Take the time to be sure that the response format is likely to provide the variability desired. Will the targeted respondents be able to distinguish among the response options?* Do not “reinvent the wheel”; wherever possible, look for examples. A nonexhaustive list of possible response options are shown in Table 12.2.
4. *Have the item pool reviewed by experts:* These should include individuals with expertise in the content area and whenever possible representatives of the target population and someone with experience in developing scales. Make sure to provide the expert panel with a conceptual guide to what you are measuring. The panel should review the items for clarity, readability, and completeness. Are there items that are too similar? Are there aspects of the content area that are not represented in the item pool?
5. *Consider inclusion of validity items:* Consider including items for assessing response bias (e.g., socially desirable answers), as well as items for establishing scale validity, for example, previously validated items measuring related constructs.

Table 12.2 Types of response options

Type	Description
Visual analog scale (VAS)	A fixed length line that has words that anchor the scale at extreme ends and no other words in between. Patients are required to place a mark on the line that corresponds to their perceived state. These scales are not usually very accurate
Anchored or categorized VAS	It has the addition of one or more intermediate marks with reference terms that help the patient to locate in between the ends of the scale
Likert scale	An ordered scale that requires the patient to choose the response that best describes their state or experience
Rating scale	A scale with numerical categories without labels and the ends of the rating scales are anchored with words. Patients are asked to choose the category which best describes their state or experience
Frequency scale	A scale with ordinal categories representing ordered categories of frequencies, for example, income categories or frequencies of occurrence
Event log	A patient diary or a reporting system in which the specific events are recorded as they occur
Pictorial scale	A set of pictures are applied to the other types of response options. Specially used for pediatric patients or for patients with cognitive impairments
Checklist	Patients are provided a simple choice between a fixed number of options such as yes, no, and don't know. They are reviewed for completeness and nonredundancy

From Ref. FDA [1]

6. *Administer items to a development sample:* This is often done in stages: item review and development and psychometric assessment. Both stages require careful consideration of the purposes of the assessment and the sample composition. *Item review and development* focuses on readability, format, administration, and identification of missing content. As such, it is easier with smaller samples. *Psychometric assessment* is used to help establish the measurement properties of the items and scale (see below). Since these are based on summaries of data, larger samples are better.
7. *Evaluate the items:* Using data from step 6, examine the item properties: scoring ranges, item variance, item-scale correlations, and item means. Assess the dimensionality of the draft scale(s): examine the underlying latent constructs (e.g., using exploratory or confirmatory factor analysis [EFA or CFA], as appropriate and if your sample is of sufficient size) and the internal consistency (e.g., using Cronbach's alpha). For well-developed scales, where basic item-scale properties have been established, consider examining differential item functioning (DIF), that is, whether items function differently among subgroups of respondents [17].
8. *Optimize scale length:* There is no magic length for a scale. Longer scales usually have better internal consistency, but having more items increases respondent burden. Fewer than four items is a pretty short scale but certainly not unknown. Items with a low (or worse, negative) contribution to alpha, a low item-total correlation, or a very high correlation with other items should be targeted for exclu-

sion; but be careful, dropping items changes the scale, and item statistics are sample estimates and therefore dependent on who is in the development sample. Being a little conservative is probably prudent.

Modification of Existing PROs

Modification of existing PROs may involve any or all of the same steps as developing a new instrument. Clearly, some modifications, such as changing the number of response categories on a few items, involve less effort than others, such as translating a PRO to a new language. However, any of these changes may necessitate reevaluation of the instrument's psychometric properties. The FDA recommends validation of revised instruments when any of the following occur [1].

- *Revision of the measurement concept:* For example, administering a single subscale from a multisubscale instrument or use of items from an existing instrument in order to create a new instrument
- *Application of the PRO to a new population or condition:* For example, use of a PRO validated in a healthy population for a population of patients with chronic illness
- *Changes in item content or format:* For example, changes in wording or scaling, changes in the recall period, or changes in formatting or instruction
- *Changes in mode of administration:* For example, adapting a PRO designed for face-to-face administration for use in a web-based battery
- *Changes in the culture or language of application:* For example, translations to another language from the language used in validation or use of an instrument in a culture it has not been validated in (even if left in the original language)

Thankfully, a recent systematic review by Muehlhausen et al. strongly suggests that electronic and paper and pencil versions of the same instruments provide data of equivalent reliability and validity [34]. The International Society for Pharmacoeconomics Outcomes Research (ISPOR) has also provided guidance on best practices in the assessment of validity in newly developed or modified PROs [35–37].

Instrument Repositories

Collections of instruments are available both in hard copy and in electronic form. McDowell provides one of the most comprehensive print compendiums of health measures available, with over 100 separate measures reviewed [13]. The purpose, conceptual basis, administration information, known psychometric properties, and copies of the items are provided for each instrument. The health domains covered include physical disability and handicap, social health, psychological well-being and affect (anxiety and depression), mental status, pain, and general health status

and quality of life. This compendium also includes an introduction to the theoretical and technical foundations of health measurement.

Online repositories are becoming increasingly available and can be significantly more expansive than print compendiums. The TREAT-NMD: Neuromuscular Network maintains the Registry of Outcome Measures (<http://www.researchrom.com/>), a searchable registry with descriptive, psychometric, availability, and contact information for each measure. Similarly, the Patient-Reported Outcome and Quality of Life Instruments Database (PROQOLID, maintained by ePROVIDE: <https://eprovide.mapi-trust.org/>) was developed by the Mapi Research Institute and managed by the Mapi Research Trust in Lyon, France, to "...identify and describe PRO and QOL instruments...." As of February, 2018, the PROQOLID site provided information on over 1500 PRO and QOL instruments and varying levels of details (basic versus detailed) depending on subscriber status.

Item Banks

The Patient-Reported Outcome Measurement Information System (PROMIS) provides a different approach to PRO measurement. PROMIS was formed by collaboration of outcomes researchers from seven institutions and the National Institutes of Health (NIH) in 2004. This cooperative group is funded under the NIH Roadmap for Medical Research Initiative to reengineer the clinical research enterprise by developing, validating, and standardizing item banks to measure PROs relevant across common chronic medical conditions, for example, cancer, congestive heart failure, depression, arthritis, multiple sclerosis, and chronic pain conditions. The main objectives of the PROMIS initiative are (adapted from the PROMIS website; <http://www.nihpromis.org/default.aspx>):

- Create item pools and core questionnaires measuring health outcome domains relevant to a variety of chronic diseases. The item pools consist of new items, as well as existing items from established questionnaires. These new items undergo rigorous qualitative, cognitive, and quantitative review before approval.
- Establish and administer the PROMIS core questionnaire in paper and electronic forms to patients suffering from a variety of chronic diseases. The collected data will then be analyzed and utilized to calibrate the item sets for building the PROMIS item banks.
- Develop a national resource for precise and efficient measurement of PROs and other health outcomes in clinical practice.
- Build an electronic web-based resource for administering computerized adaptive tests, collecting self-report data, and reporting instant health assessments.
- Conduct feasibility studies to assess the utility of PROMIS and promote extensive use of the instrument for clinical research and clinical care.

The PROMIS item library is a large relational database of items gathered from existing PROs. The library was created with an intention of supporting the

identification, classification, improvement, and writing of items that serve as candidate items for upcoming PROMIS item banks.

During the first phase of the initiative (2004 to present), the PROMIS network of researchers have developed questions or “items” for assessing patient outcomes (e.g., pain, fatigue, physical functioning, emotional distress, and social role) [38–40]. PROMIS is creating a computer adaptive testing (CAT) system, based on item response theory (IRT), to administer these items, and is developing a web-based system to give clinical researchers access to the item banks and the CAT system [41]. Using these approaches, PROMIS has demonstrated improved item performance relative to existing PROs.

Conclusion

Well-developed PRO instruments are the best and perhaps only way to gather valid data from the patient perspective. PROs are now accepted as providing a necessary adjunct to more traditional clinical and laboratory outcome measures; for example, a patient’s perception of their overall health status is increasingly used in conjunction with clinical measures of disease burden. PRO measures may also provide primary outcome data when clinical and/or laboratory measures are not appropriate or available, for example, when a patient’s assessment of pain or quality of life is needed.

The increased emphasis on the patient’s experience as a therapeutic outcome and a health-care priority is necessitating the development and use of PRO measures that are appropriate for a variety of diseases and patient populations. A large literature on PRO measures and their application already exists. The development of instrument compendia and repositories, such as the Registry of Outcome Measures and the PROQOLID, and item banks, such as the PROMIS database and their related technologies, is providing valuable tools for expanding the implementation of PRO measures. However, with thousands of identified diseases, and with instruments having demonstrated utility needing adaptation and validation across languages and cultures, a considerable amount of work remains to be done.

Along the same lines, the evolution of the clinical information infrastructure is revolutionizing the way medical information can be organized, accessed, and used. Collection and use of PROs is a key piece of that revolution. Technological development has made the implementation of PRO measures much easier. However, the evaluation of the impact of new technologies on the validity and usability of the information collected remains, and will likely always remain, ongoing. It is crucial that health information professionals have a thorough understanding of the design principles outlined here and their potential impact on the reliability and validity of PRO measures. These principles should be the foundation of any PRO development effort.

References

1. FDA. Guidance for industry: patient-reported outcome measures; use in medical product development to support labeling claims. Silver Spring: U. S. D. o. H. a. H. Services; 2009.
2. McKenna P, Doward L. Integrating patient reported outcomes. *Value Health*. 2004;7:S9–12.
3. Garratt A. Patient reported outcome measures in trials. *BMJ*. 2009;338:a2597.
4. Wiklund I. Assessment of patient-reported outcomes in clinical trials: the example of health-related quality of life. *Fundam Clin Pharmacol*. 2004;18:351–63.
5. Fayers PM, Machin D. Quality of life: the assessment, analysis and interpretation of patient-reported outcomes. Chichester: Wiley; 2013.
6. Atkinson MJ, Lennox RD. Extending basic principles of measurement models to the design and validation of patient reported outcomes. *Health Qual Life Outcomes*. 2006;4(1):65.
7. Frost MH, Reeve BB, Liepa AM, Stauffer JW, Hays RD, Mayo/FDA Patient-Reported Outcomes Consensus Meeting Group. What is sufficient evidence for the reliability and validity of patient-reported outcome measures? *Value Health*. 2007a;10:S94–S105.
8. Shields A, Gwaltney C, Tiplady B, et al. Grasping the FDA's PRO guidance: what the agency requires to support the selection of patient reported outcome instruments. *Appl Clin Trials*. 2006;15:69–83.
9. Valderas J, Alonso J. Patient reported outcome measures: a model-based classification system for research and clinical practice. *Qual Life Res*. 2008;17:1125–35.
10. Calvert M, Blazeby J, Altman DG, Revicki DA, Moher D, Brundage MD, CONSORT PRO Group. Reporting of patient-reported outcomes in randomized trials: the CONSORT PRO extension. *JAMA*. 2013;309(8):814–22.
11. Skinner J, Teresi J, et al. Measurement in older ethnically diverse populations: overview of the volume. *J Ment Health Aging*. 2001;7:5–8.
12. Anastasi A. Psychological testing. 6th ed. New York: Macmillan Publishing Company; 1998.
13. Morgan R, Teal C, et al. Measurement in VA health services research: veterans as a special population. *Health Serv Res*. 2005;40:1573–83.
14. Frost MH, Reeve BB, Liepa AM, Stauffer JW, Hays RD, the Mayo/FDA Patient-Reported Outcomes Consensus Meeting Group. What is sufficient evidence for the reliability and validity of patient-reported outcome measures? *Value Health*. 2007b;10(S2):S94–S105.
15. DeVellis RF. Scale development: theory and applications. 3rd ed. Thousand Oaks: Sage; 2012.
16. Vogt W. Dictionary of statistics and methodology: a nontechnical guide for the social sciences. 2nd ed. Thousand Oaks: Sage Publications; 1999.
17. Aday L, Cornelius L. Designing and conducting health surveys: a comprehensive guide. 3rd ed. San Francisco: Jossey-Bass; 2006.
18. McDowell I. Measuring health: a guide to rating scales and questionnaires. 3rd ed. New York: Oxford University Press; 2006.
19. Revicki D, Hays RD, Celli D, Sloan J. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. *J Clin Epidemiol*. 2008;61(2):102–9.
20. Bowling A. Mode of questionnaire administration can have serious effects on data quality. *J Public Health*. 2005;27:281–91.
21. Dillman DA. Internet, mail and mixed-mode surveys: the tailored design method. 4th ed. New York: Wiley; 2014.
22. Snyder CF, Blackford AL, Wolff AC, Carducci MA, Herman JM, Wu AW, the PatientViewpoint Scientific Advisory Board. Feasibility and value of PatientViewpoint: a web system for patient-reported outcomes assessment in clinical practice. *Psycho-Oncology*. 2013;22(4):895–901.
23. Coons S, Gwaltney C, et al. Recommendations on evidence needed to support measurement equivalence between electronic and paper-based patient-reported outcome (PRO) measures: ISPOR ePRO good research practices task force report. *Value Health*. 2009;12:419–29.
24. Electronic Patient Reported Outcomes. PAREXEL. <https://www.parexel.com/solutions/informatics/clinical-outcome-assessments/epro>. Accessed 2 Feb 2018.

25. Electronic Patient Reported Outcomes (ePRO). ICON plc. <http://www.iconplc.com/jp/technology/application-areas/electronic-patient-report/>. Accessed 2 Feb 2018.
26. Patient EngagementlePRO. IBM clinical development. <https://www.ibmclinicaldevelopment.com/en/ibm-clinical-development-epro>. Accessed 2 Feb 2018.
27. Patient Reported Outcomes. VitalHealth Software. <https://www.vitalhealthsoftware.com/products/patient-health-questionnaires/patient-reported-outcomes>. Accessed 2 Feb 2018.
28. e-Patient Reported Outcomes.Acceliant. <http://www.acceliant.com/products/e-patient-reported-outcomes>. Accessed 2 Feb 2018.
29. Rave eCOA/ePRO. Medidata. <https://www.mdsol.com/en/products/rave/ecoapro>. Accessed 2 Feb 2018.
30. Cole E, Pisano ED, Clary GJ, Zeng D, Koomen M, Kuzniak CM, Seo BK, Lee Y, Pavic D. A comparative study of mobile electronic data entry systems for clinical trials data collection. *Int J Med Inform.* 2006;75:722–9.
31. Collins R, Kashdan T, et al. The feasibility of using cellular phones to collect ecological momentary assessment data: application to alcohol consumption. *Exp Clin Psychopharmacol.* 2003;11:73–8.
32. Freedman M, Lester K, et al. Cell phones for ecological momentary assessment with cocaine-addicted homeless patients in treatment. *J Subst Abus Treat.* 2006;30:105–11.
33. Reid S, Kauer S, et al. A mobile phone program to track young people's experiences of mood, stress and coping. *Soc Psychiatry Psychiatr Epidemiol.* 2009;44(6):501–7.
34. Muehlhausen W, Doll H, Quadri N, Fordham B, O'Donohoe P, Dogar N, Wild DJ. Equivalence of electronic and paper administration of patient-reported outcome measures: a systematic review and meta-analysis of studies conducted between 2007 and 2013. *Health Qual Life Outcomes.* 2015;13:167. <https://doi.org/10.1186/s12955-015-0362-x>.
35. Rothman M, Burke L, Erickson P, Leidy NK, Patrick DL, Petrie CD. Use of existing patient-reported outcome (PRO) instruments and their modification: the ISPOR good research practices for evaluating and documenting content validity for the use of existing instruments and their modification PRO task force report. *Value Health.* 2009;12(8):1075–83.
36. Patrick DL, Burke LB, Gwaltney CJ, Leidy NK, Martin ML, Molsen E, Ring L. Content validity—establishing and reporting the evidence in newly developed patient-reported outcomes (PRO) instruments for medical product evaluation: ISPOR PRO good research practices task force report: part 1—eliciting concepts for a new PRO instrument. *Value Health.* 2011;14(8):967–77.
37. Patrick DL, Burke LB, Gwaltney CJ, Leidy NK, Martin ML, Molsen E, Ring L. Content validity—establishing and reporting the evidence in newly developed patient-reported outcomes (PRO) instruments for medical product evaluation: ISPOR PRO Good Research Practices Task Force report: part 2—assessing respondent understanding. *Value Health.* 2011;14(8):978–88.
38. Reeve BB, Hays RD, Bjorner JB, Cook KF, Crane PK, Teresi JA, et al. Psychometric evaluation and calibration of health-related quality of life item banks: plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). *Med Care.* 2007;45(5):S22–31.
39. Celli D, Yount S, Rothrock N, Gershon R, Cook K, Reeve B, et al. The Patient-Reported Outcomes Measurement Information System (PROMIS): progress of an NIH Roadmap cooperative group during its first two years. *Med Care.* 2007;45(5 Suppl 1):S3.
40. Celli D, Riley W, Stone A, Rothrock N, Reeve B, Yount S, et al. The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005–2008. *J Clin Epidemiol.* 2010;63(11):1179–94.
41. Harniss M, Amtmann D, et al. Considerations for developing interfaces for collecting patient-reported outcomes that allow the inclusion of individuals with disabilities. *Med Care.* 2007;45:S48–54.



Patient Registries for Clinical Research

13

Rachel L. Richesson, Leon Rozenblit, Kendra Vehik,
and James E. Tcheng

Abstract

Patient registries are fundamental to biomedical research. Registries provide consistent data for defined populations and can be used to support the study of the determinants and manifestations of disease and provide a picture of the natural history, outcomes of treatment, and experiences of individuals with a given condition or exposure. It is anticipated that electronic health record (EHR) systems will evolve to ubiquitously capture detailed clinical data that supports observational, and ultimately interventional, research. Emerging data representation and exchange standards can enable the interoperability required for automated transmission of clinical data into patient registries. This chapter describes informatics principles and approaches relevant to the design and implementation of patient registries, with emphasis on the ingestion of clinical data and the role of patient registries in research and learning health activities.

R. L. Richesson, PhD, MPH, FACMI (✉)
Duke University School of Nursing, Durham, NC, USA
e-mail: rachel.richesson@dm.duke.edu

L. Rozenblit, JD, PhD
Prometheus Research, LLC, New Haven, CT, USA
e-mail: leon@prometheusresearch.com

K. Vehik, PhD, MPH
University of South Florida, Health Informatics Institute, Tampa, FL, USA
e-mail: kendra.vehik@epi.usf.edu

J. E. Tcheng, MD
Duke University School of Medicine, Durham, NC, USA
e-mail: james.tcheng@dm.duke.edu

Keywords

Registries · Clinical research · Secondary data use · Observational research methods · Data standards · Interoperability · Outcomes measurement · Learning health systems

Definitions and Types of Registries

A *patient registry* is “...an organized system that uses observational study methods to collect uniform data (clinical and other) to evaluate specified outcomes for a population defined by a particular disease, condition, or exposure, and that serves one or more predetermined scientific, clinical, or policy purposes” [1]. The follow-up of the relevant population is implied in this definition. Registries are generally patient focused, meaning that some population of patients are the foundation and data is added over time. The term *clinical registry* is often used to refer to registries that originate in clinical setting or include data from healthcare visits. One type of clinical registry is a *quality reporting registry*, designed to capture records of procedures of interest and support the analysis of outcomes, treatment effectiveness, and other quality improvement (QI) goals. QI registry programs may not follow individual patients over time and thus do not always meet the true definition of a patient registry.

There are three broad types of patient registries: disease (or condition or syndrome), exposure (e.g., medical or surgical treatment, medical devices, environmental, geographical, or regional), and participant characteristic (e.g., genetic, twin, sibling, healthy controls) (Fig. 13.1). While disease and exposure registries (particularly drugs, devices, and procedures) [2] are the most common types of registries, growth in participant characteristic registries is increasing due to a surge of new genetic registries and annotated data records associated with biological repositories [3–6].

The Role of Registries in Evolving Research Contexts

Patient registries have been a fundamental part of research for nearly two centuries [7, 8], as observing and following populations increase our understanding of the etiology and natural history of disease. Registries have been used to support clinical

Patient registries

Inclusion criteria:

Disease,
syndrome, or
condition

(Disease or pre-disease)

Exposure

(Drugs, devices,
procedures,
environment,
geography, health
coverage)

Participant
characteristics

(Genetic, twin, sibling,
biorepository, healthy
volunteers)

Fig. 13.1 Types of inclusion criteria for patient registries

trial planning and recruitment, post-market surveillance, biomarker discovery, comparative effectiveness research, and patient safety. Registries can be sponsored by pharmaceutical companies, patient or disease advocacy groups, healthcare organizations, universities, and government. For rare and neglected diseases, registries have become an essential component of research programs by providing data that describes disease burden, supports comparisons among therapeutic approaches, and identifies potential patients for clinical trials. NIH maintains an inventory of such registries at <https://www.nih.gov/health-information/nih-clinical-research-trials-you/list-registries>. The Agency for Healthcare Research and Quality (AHRQ) hosts a database (called the Registry of Patient Registries, RoPR) of information about registries that is intended to promote collaboration, reduce redundancy, and improve transparency in registries [9]. The submission of registry information to RoPR is optional, and therefore not all registries are included. As of 2018, the Centers for Medicare and Medicaid Services (CMS) is promoting the RoPR as a replacement of the CMS Centralized Repository of Registries for voluntary reporting of certified registries by public health agencies, clinical data registries, and specialized registry providers [10]. The database can be searched at <https://patientregistry.ahrq.gov/>.

Under the Food and Drug Administration Amendments Act of 2007 in the United States, the FDA can mandate post-approval (Phase 4) studies and Risk Mitigation and Evaluation Systems as a condition of approval for new products with potential safety issues [11]. This regulation has spurred a number of rare disease registries hosted by pharmaceutical companies to follow patients with rare conditions managed with specific medications. Because there are competing treatments for some rare diseases, there are cases of multiple registries for a given rare condition. Unfortunately, this reduces opportunities to evaluate the outcomes of combinations and permutations of treatment in those conditions [12].

The Clinical Trials Transformation Initiative (CTTI), funded in part by the US Food and Drug Administration, envisions registries as a critical reusable component of the clinical trial infrastructure within which prospective randomized studies can be performed, promising increased speed and efficiency due largely to the reduction of duplicate data collection and barriers for identifying and enrolling subjects [13]. CTTI is working toward this vision with demonstrations of registry-based interventional studies and subsequent evaluations of the adequacy of this framework for regulatory decisions.

The Role of Registries in Quality Improvement and Learning Health Systems

Increasingly, registries provide data for various types of observational (health services and QI) research. Surgical registries, for example, have been used to develop risk calculators, assess measures of performance and outcomes, and share data with providers to drive improvements in care [14]. Intuitively, sharing data on provider performance or patient outcomes can increase compliance to clinical protocols as a mechanism to improve clinician performance. A recent driver of registry development is the CMS promotion of “qualified registries” for merit-based incentive

payments to providers under the Medicare Access and CHIP Reauthorization Act of 2015 (MACRA) [15]. A *qualified registry* is a CMS-approved entity that collects clinical data from eligible clinicians or practice groups and submits the data to CMS on their behalf. The Agency for Healthcare Research and Quality (AHRQ), whose mission is to produce and promote evidence that supports safe and high-quality healthcare, has sponsored registries to improve quality and support the increased uptake of evidence in practice (Box 13.1).

Box 13.1 An example of a registry for generating evidence and improving patient outcomes: the Function and Outcomes Research for Comparative Effectiveness in Total Joint Replacement (FORCE-TJR)

The FORCE-TJR, funded by AHRQ, is a comprehensive US database of total hip and knee joint replacement patients and their surgical and patient-reported outcomes. The registry has enrolled more than 50,000 patients from diverse backgrounds and settings (mostly from community practices) and provides a nationally representative cohort of patients to establish national norms, clinical benchmarks, and risk-adjustment models for joint replacement procedures.

The FORCE-TJR registry includes real-time scores of patient-reported outcomes on symptoms and function, which are used to guide treatments and discussions between patients and clinicians to optimize patient function and quality of life after surgery.

The FORCE-TJR illustrates how a registry can enable evidence implementation and generation in the learning healthcare system. The registry enabled the evaluation of new orthopedic devices and better understanding of important patient safety issues (e.g., postsurgical infection and medication use).

The FORCE-TJR registry also addresses provider needs, including serving as a Qualified Clinical Data Registry to meet CMS quality reporting requirements under the Merit-Based Incentive Payment System and by helping providers understand the risks in their patient pool, which is important to negotiate value-based and bundled payment arrangements.

The registry was highlighted as an exemplary implementation of learning health system principles in a publication of the National Academy of Medicine [16].

Source: *Bringing the Patient Voice to Evidence Generation: Patient Engagement in Disease Registries.* Content last reviewed March 2018. Agency for Healthcare Research and Quality, Rockville, MD. <http://www.ahrq.gov/news/blog/ahrqviews/disease-registries.html> [17].

The emerging interest in learning health systems (LHS) and the mounting number of LHS demonstrations will likely increase the demand and use of registries by healthcare organizations. The concept of the LHS includes infrastructure, tools (e.g., registries), processes, and incentives to support the translation of research (i.e., evidence-based medicine) into practice and the return of real-world evidence that influences research (i.e., evidence-generating medicine) [18].

Figure 13.2 illustrates how registries support the application and generation of evidence in the LHS. The reuse of clinical data for QI and research purposes is fundamental to the LHS, but there are a number of steps that need to be undertaken to ensure that the data collected as part of the clinical workflow is indeed sufficient to support the information needs of the LHS. These steps include (1) the collection and ingestion of data into the registry, (2) the capture of data into a database including linkage of data across sources, (3) the curation (“cleaning”) and (4) enrichment of the data, (5) transformation to create data sets that meet different analytic purposes, and (6) the distribution and delivery of these data sets to support eventual analysis and presentation of the data to address research or business questions. This analysis can be used to inform the design of new interventions or practice changes that can be implemented and evaluated in the context of actual patients.

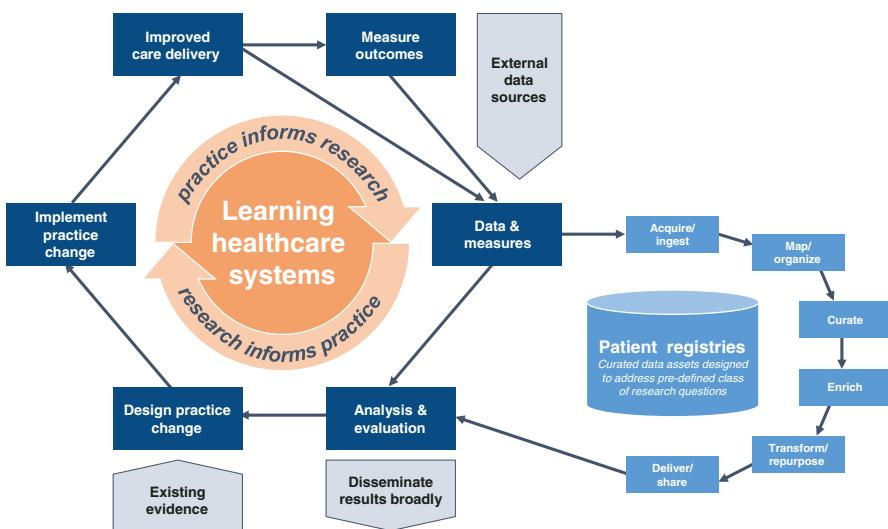


Fig. 13.2 The role of registries in learning healthcare systems

Using Clinical Data for Patient Registries

Data for a registry must either be specifically collected for the registry or abstracted from documentation. It seems intuitive that clinical information collected as a by-product of healthcare delivery (via EHR systems) might be used by registries. Although the use of clinical information from EHRs has the potential to provide an efficient source of data for patient registries, there are challenges. Some of the data necessary for the registry may not be captured in the EHR, or it may only exist in an unstructured form, requiring potentially costly manual or natural language process to convert to structured data. In practice, only a small proportion of data needed by a typical registry is actually captured as structured data in EHR systems.

Most EHR systems collect demographics, patient encounters, medications, diagnoses, problem lists, procedures, and laboratory results as structured data, along with text-based (unstructured) clinical notes. However, while the ingestion and use of these data promise to eliminate the costs of clinical data collection and abstraction, the use of clinical data for registries often requires significant time and informatics resources to ensure these clinical data are sufficiently fit for the intended purposes of the registry. In the sections to follow, we describe challenges for using clinical data for research. There are multiple dimensions that must be addressed to achieve interoperability and support registries at scale. Then we describe the limitations of patient registries and clinical data, including various types of bias.

Interoperability and Data Standards

The lack of interoperability (i.e., the ability of computer systems to communicate, exchange data, and use the information that has been exchanged) is the greatest challenge for the reuse of clinical data in registries, as well as for secondary use of EHR data in general. Multiple sustained efforts to address this challenge have been underway for decades. The Health Level Seven (HL7) organization has interoperability as its core mission. In the past few years, the Office of the National Coordinator (ONC) and CMS and other federal entities have been developing policies and incentives for the use of clinical registries that include data from EHRs. Different types of standards (described more fully in Chap. 19) are needed for interoperability. These are exchange standards (including frameworks to support model-driven software applications) and content standards (including controlled vocabularies and coding systems, entity identifiers, general and condition-specific common data elements, computable phenotype definitions, and outcome measures). We briefly describe each below.

Exchange Standards

Data exchange standards provide an agreed-upon format to move data from system to system without loss of meaning. Standards for exchanging data between clinical systems have evolved rapidly over the last 5 years. Until recently, the large majority of clinical data exchange happened using the HL7v2 messaging standard. This standard is dominant in exchanging data within an organization (e.g., between different IT systems within a single hospital) and is likely to remain so for many years to come. Despite a long history of success and broad adoption, HL7v2 is considered difficult to use and insufficient to meet the challenges of emerging data exchange use cases, especially those that require exchange of data between different organizations. The key reasons are the permissive nature of the HL7v2 messages and the point-to-point approach that is foundational to HL7v2. By providing generic models for data capture and exchange, HL7v3 was intended to address those limitations through greater structure and use case modeling. However, HL7v3 has never gained critical mass, largely because of its complexity. The rapid changes over the last decade involved moving away from the messaging metaphor to document-based standards, like the Consolidated Clinical Document Architecture (C-CDA) and Representational State Transfer (RESTful) APIs, in particular Fast Healthcare Interoperability Resources (FHIR).

It is a reasonable prediction that the two data exchange standards likely to gain momentum over the next decade are C-CDA (a document standard) and HL7 FHIR, a modern RESTful API. FHIR is especially exciting because it is sufficiently similar to commercial web programming constructs that it can be readily adopted by the programmer community (who often resist complex healthcare standards). Thus, the clinical data exchange standards that a future registry developer is most likely to encounter in the next 5 years are HL7v2 messages, C-CDA documents, and HL7 FHIR APIs.

Content Standards

In addition to transmission and exchange standards, different systems (i.e., EHR and registry) must maintain commonality of the semantic content of the data in each system. While exchange standards provide rules for formulating messages that communicate specific facts about a shared reality, content standards are required to ensure that different systems can represent and process that shared reality. Content standards can be organized into several broad categories: (i) coding systems or controlled terminologies (like ICD-10-CM and SNOMED CT), (ii) entity identifiers (such as patient identifiers or unique device identifiers (UDI)), and (iii) clinical models and data elements. Of note, distinctions among these categories are imprecise; complex standards like SNOMED CT encompass features of both terminologies and clinical models (see Chap. 19).

Coding Systems and Controlled Terminologies

Approaches to clinical terminologies are varied and complex, but fortunately there is increasing consensus about this topic. The most common and important clinical terminologies a registry team is likely to encounter include those standards for EHRs that are recognized by CMS and ONC, specifically ICD-10-CM (for clinical diagnoses and problem lists), RxNorm (for medications), LOINC (for laboratory test and results, observations), CPT (for billing codes sent to payers for reimbursement for performed procedures), and SNOMED CT (for clinical concepts on problem lists and as a reference terminology for concepts extracted from free text) [19]. Because these coding systems are mandated, they are widely included in different EHR products, and they can simplify the work of aggregating data across multiple organizations. Unfortunately, these terminology standards are complex and imprecise, enough to be used differently by different users. For example, there are hundreds of valid codes for glucose tests in LOINC, and integrating codes that are used differently in different institutions is part of the data curation challenge when building a registry.

Content Standards: Common Clinical Models and Data Elements

Terminologies alone are insufficient to precisely communicate clinical or scientific meaning; they must be bound to clinical data models to fully represent the semantic context. The easiest way to think about these models is as collection of data elements that can take on a range of predefined values with agreed-upon meanings. For example, “family history of cancer” could be represented as an (data) element in a clinical data model, with values of yes/no, present/absent, or perhaps different types of cancers. The same concept “family history of cancer” could also be represented entirely in the terminology (assuming a sufficiently robust clinical terminology such as SNOMED CT), or the concept could be modeled in different ways, e.g., the data element could be “family history of [conditions]” and “cancer” (including type and location) could be one value (or code) of many codes for various conditions. In reality, there are multiple approaches for system designers to semantically model clinical information using terminologies and clinical models [20]. Creation of clinical models and terminology bindings for a domain is a difficult, tedious, and time-consuming exercise that involves negotiation between multiple stakeholders. Moving data from an EHR system to a registry requires an explicit representation and eventual harmonization of the clinical data model and terminology binding approach in both the source and receiving systems.

There are two types of clinical models relevant to registries: general “common data elements” that apply to most registries, like sex, age, and diagnosis and condition-specific elements that are likely to apply only within a particular clinical area, such those that might be necessary to represent hypertension, or melanoma, or depression. To avoid reinventing the wheel, existing sources should be explored when considering how to standardize common data elements (CDEs). These include

Common Clinical Data Set definitions [21], the NLM Common Data Element resource portal [22], and most recently the CMS Data Element Library [23]. In addition, the HL7 Common Clinical Registry Framework (CCRF) project is developing a set of common clinical data elements that can be generalizable across most clinical registries and plans to transform these registry CDEs into implementable logical clinical information models suitable for instantiation as elements in an information exchange standard such as FHIR, CDA, or HL7v2.

Entity Identifiers Including the Unique Device Identifier (UDI)

Patient registries and device registries also require standards that can unambiguously reference specific *entities* in the real world in order to add and analyze data for unique patients and devices. In the United States, the problem of easily identifying unique patients remains challenging because the implementation and use of national unique patient identifiers has proven politically intractable, despite considerable support from the health information technology industry and many providers [24]. Because Congress previously prohibited research or development on a national patient identifier system, patient identifiers are typically proprietary and unique to specific health systems, instead of traveling with patients from one care context to the next. Consequently, matching patient records across organizations requires deterministic or probabilistic linkage methods [25]. Fortunately, Congress has recently reversed their stance, authorizing evaluation of the need for a unique patient identifier in the *21st Century Cures Act* [26].

Unique identifiers for health plans and providers can facilitate analyses to understand the impact of different types of care on patient outcomes. HIPAA established a standard, unique identifier for health plans (the Health Plan Identifier, HPID), employers (the Employer Identification Number (EIN) issued by the Internal Revenue Service used to identify employers in electronic transactions), and providers (the National Provider Identifier (NPI) used for qualified providers, but typically not RNs in supervised roles.) Unique identifiers for NPIs and EINs are required for all HIPAA transactions.

A major requirement for registries is to capture specifics about the intervention and other treatments and exposures. Medical treatments can be captured via controlled terminologies such as RxNorm or SNOMED CT. However, device registries require *unique identifiers* for specific devices (coded by serial number, not by type of device). Unlike the failures to achieve a national unique patient identifier system, unique identification of medical devices has seen significant progress. The FDA has supported the development of the Unique Device Identification (UDI) standard for close to a decade and has demonstrated the feasibility of integrating this in EHR systems and the utility of the UDI in evaluating the safety of devices [27]. At present, the uptake and demand for UDI is growing, and it is an integral part of any device registry. Supporters including the Medical Device Epidemiology Network (MDEpiNet), an FDA public-private partnership, have helped promote the UDI and are named in the ONC 2018 Interoperability Standards Advisory [19]. Extensions

will be needed to add enriching data elements that the original UDI does not cover. The update and maintenance of the UDI standard will not only advance the capacity of device registries to evaluate effectiveness but will also facilitate the use of registries for patient safety, safety monitoring, and recalls.

Clinical Phenotype Definitions

The use of controlled terminologies and code systems for the data in EHRs is not sufficient to address registry needs. Standard approaches to assemble multiple codes from one or more terminology systems to define the conditions in the registry can support efficiencies in building registries and interoperability between registries. As clinical vocabularies have become more granular (e.g., ICD-10-CM Diagnosis Code E11.51 – “Type 2 diabetes mellitus with diabetic peripheral angiopathy without gangrene” is one of hundreds of codes for diabetes), *groups of relevant codes* (i.e., computable phenotypes) are required to define broad conditions that are usually the subject of registries. Theoretically, the use of robust terminologies such as SNOMED CT will enable one to identify broad classes of diseases (e.g., diabetes or autoimmune disorders), and all subtypes or related types of disease can easily be included using subsumption or other logical expressions [28]. In practice, however, multiple codes from multiple code systems (e.g., laboratory, medication) are often required to fully define a condition from the perspective of an EHR query. Explicit documentation of these codes and logic is called a *computable phenotype*. Computable phenotypes are standardized (EHR-based query) definitions for defining patient populations or cohorts. They can be used in registries to define eligibility criteria, study endpoints (for trials), or patient outcomes.

Currently, these computable phenotypes are developed locally. It would be much faster to adopt existing definitions that have sufficient documentation and evidence of validity or performance in clinical settings. There are a few locations where computable phenotypes can be found ([PheKB.org](#) [29], the AHRQ Chronic Conditions Warehouse, and the NLM Value Set Authority Center [30]) but not one single authoritative source for all registry purposes. Further, there are many customizations that must be made to apply a phenotype definition locally, and these details are generally not reported or easy to find. The PheMA project is developing tools to make it easier for groups to develop phenotypes and share the executable formats plus the underlying logic/implementation details, plus information about validation in previous settings [31].

Researchers from the NIH Collaboratory have recognized the importance of explicit and reproducible computable phenotypes in pragmatic research [32]. They advocate the reuse of existing definitions whenever possible, but also recommend local validation. Others have followed up with details on how the ecosystem and incentives need to change to fully support phenotype sharing [33]. In the future, incentives or regulations could be used to increase the sharing and reuse of explicit phenotype definitions. For example, journals, research sponsors, or registry inventories (like the RoPR) could require registration and explicit definition of phenotypes used in the research. The adoption of EHR standards will ultimately support efficiencies and reuse of phenotype definitions used in patient registries.

Outcome Measures

Much of the data required for registry reporting – particularly those elements used in program or treatment evaluation – are actually computed or summary data. Examples of this include the highest value of a test result, the time between procedure A and B, the order and timing of treatments and intervention, the total number of readmissions or procedures in a specified time period, and the total number of hours in ED. These outcome measures can build upon other data elements or value sets but require computation or processing to generate. In the current state, different registry providers compute these outcome measures using different definitions in different places (or sometimes in two different data collection systems in the same place), which often leads to an inability to analyze and compare data across settings. To address this problem, the AHRQ has developed an Outcome Measures Framework to model these aggregate elements and harmonize data definitions [34]. Further, authors of the AHRQ-sponsored report propose that a library of outcome measures be maintained so that they may be reused across registries. Certainly, such a resource would create efficiencies in the development of new registries in different organizations and would facilitate comparisons between registries and organizations. Patient-reported outcomes are often and increasingly being used in patient registries, and these are discussed in depth in Chap. 12.

The Common Clinical Registry Framework Model and Other HL7 Standards

The very important role of registries in clinical research and LHS has spawned the development of a standards development group to identify relevant standards for registries. This relatively new Common Clinical Registry Framework (CCRF), led by the HL7 Clinical Interoperability Council (CIC), is gaining momentum and will be very relevant to the design of new clinical registries that are developed from clinical data sources. The work of the CIC includes specification of the transmission and content standards described earlier (specific to registries), along with functional standards that specify the functionality EHR systems should have in order to support the automated transmission of clinical data to patient registries. The CCRF project has created a registry domain analysis model, a set of common data elements, and a registry CDEs' logical data models. The CCRF domain analysis model (DAM) describes the function, organization, structure, and major workflows of a general clinical registry.

The Clinical Information Interoperability Council (CIIC), cosponsored by HSPC and HL7, provides governance for all clinical data modeling projects, including the Registries on FHIR project, designed to promote interoperability standards to increase efficiency and consistency between registries. The Registries on FHIR project includes the development of the Registry FHIR Specification Standard, led by the HL7 Patient Care workgroup, and a number of Registry on

FHIR Demonstration Projects, led by ROF Early Adopters from industry including registry operators, registry IT vendors, EMR vendors, registry participants (i.e., data sources), and registry users (i.e., data consumers). The Registries on FHIR project will utilize the USCDI data elements to develop the common core clinical data element for Registries. The Registries on FHIR project is led and sponsored by the Physician Consortium for Performance Improvement (PCPI) in collaboration with the Medication Device Epidemiology Network, the Duke Clinical Research Institute, and the Health Level Seven. These projects will enable the experience required to advance the status of the CCRF FHIR specification from a “draft” to a “normative” standard in HL7. This multidisciplinary, cross-organizational collaboration for standardizing data and functions for registries is unprecedented and likely predicts a converged future state.

Limitations of Registries

As the development and uptake of clinical data and registry standards gain momentum and the technical barriers for automated transmission of data from EHR to registries are reduced, it is critical to also consider the inherent limitations of registries in observational research. The standards and interoperability issues sometimes overshadow the fundamental issue that any clinical data is inherently biased and might not be generalizable to all populations. Registries – especially those that reuse data collected from clinical settings – are vulnerable to all the biases of observational research. As such, patient registries have limitations in the questions that they can answer, and consumers of registry data should be thoughtful in the interpretation of data and analytic results. Researchers must be particularly careful about using a registry to count or characterize health or disease characteristics, for comparative effectiveness research, and to extrapolate those results back to a larger or different population. Because the registry only represents a *sampled* population, researchers must be able to estimate the completeness of case ascertainment (i.e., the inclusion of all cases in the sample area, time, or place) [35, 36]. Developers and users of registry data must also be aware of the bias issues related to changes or improvement in case detection (see Box 13.2 – Errors and Bias). The identification and elucidation of new biomarkers (e.g., genetic, gene expression, metabolomics, microbiome) enable earlier determination of the presence of disease, and improvements in testing quality and sensitivity make it difficult to compare registry cases over time (consequently, the collection of information specific to the method of diagnosis, including detailed testing information, should be considered to support future analyses of the data). Registry follow-up data must provide the proportion of follow-up obtained and the nature of cases lost to follow-up.

Box 13.2 Types of error and bias in patient registries

- *Random error* is unpredictable and is associated with precision, usually due to chance alone.
- *Systematic error* is bias in a measurement that distorts the measured values from the actual values (i.e., instrument calibration, environmental changes, and procedure changes).
- *Selection bias* is a distortion that results from procedures used to select subjects for the registry or from factors that influence participation or inclusion, i.e., *self-selection bias* (also called healthy-worker/volunteer effect) [37].
- *Information bias* results from systematic errors in the measurement of either the exposure or the disease, such as poor questionnaire/survey design, data collection procedures (“interviewer bias”), selective recollection of exposures (“recall bias”), and imprecise diagnostic procedures.
- *Non-differential bias* affects the entire monitoring process rather than just a specific piece of the process and underestimates the result.
- *Differential bias* affects the accuracy of the data in different subsets of the population.
- *Misclassification* is a type of bias generally associated with categorical or discrete variables. Bias introduced into registries by inaccuracies or variation in methods of data acquisition and case or exposure definitions leads to skewing registry summary data.
- *Lead-time bias* leads to an earlier identification of disease (i.e., advances in testing, high-risk monitoring captures before clinical symptoms).
- *Variability* is a random bias that may attenuate true associations in epidemiologic measures, but is not intrinsically fatal to certain registry objectives.
- *Sensitivity* is the probability that a subject who is truly diseased/exposed will be classified as such by the method used for ascertainment and estimates how successful a registry is at identifying all of the events, cases, or exposures in the target population.

Despite these limitations, registries will likely play an important role in research and LHS in the future. Less clear is the future role of less organized data collections. These collections might even be described as “registries,” but often lack the distinguishing features of patient registries and, as a consequence, have additional limitations (see Box 13.3).

Box 13.3 “Quasi-registries” and associated limitations

A number of different data query models and repositories that resemble registries have been developed by organizations and vendors. These models lack critical registry functions and are often unable to deliver a curated data set that answers a class of analytic questions. These include:

Data lakes. A data lake is a centralized repository that stores (essentially) all of an organization’s structured and unstructured data, including nontraditional data sources such as web server logs, sensor data, social network activity, text, and images. The data is in “one place” in the sense that it is accessible to an authenticated user via some set of tools. However, the data has not been organized. Generating an analytic data set is a time-consuming process where data organization, curation, and transformation must be repeated for each analysis.

“Registry” functions offered in commercial EHR products. Many EHR vendors offer view of the data that they call a “Registry.” In these cases, data is available as a dynamic view of data within an EHR system (such as in the case of the Epic Registry function). The data is represented as a list of patients indexed by particular attributes. This is convenient for generating a cohort of patients to follow. However, this approach doesn’t support curation, enrichment, or transformation and is unlikely to meet the needs of sophisticated registry programs.

Informatics Approaches for Building Registries

The AHRQ commissioned a comprehensive report on the role of patient registries for scientific, clinical, and policy purposes [38]. This report, recently updated, provides the most comprehensive and relevant set of best practices for registry design and framework for assessing quality of registry data for evaluating patient outcomes. The general steps include the following: evaluate alternatives to a registry given the cost and commitment; develop and document explicit goals for the registry; develop leadership structure and policies for data storage, protection, and access; develop adequate infrastructure; identify data sources; identify inclusion/exclusion criteria, including case definitions; develop sampling and surveillance methods; design data collection instruments; plan follow-up data collection procedures; and continually reevaluate the registry purpose and fit of the registry to that purpose. In addition, the CITI project has provided a set of recommendations for registry-based clinical trials (<https://www.ctti-clinicaltrials.org/files/recommendations/registrytrials-recs.pdf>), which rely on the premise that a registry-embedded interventional study should only be considered after the registry has demonstrated the reliability, robustness, and relevancy of the data needed for the trial, along with the appropriate processes to assure patient protections.

If a registry aims to include clinical data, one must consider carefully whether automated ingestion of data from clinical systems (e.g., EHRs) will be efficient. The main source of difficulty is that the data elements necessary for the registry functions may not be directly represented in the clinical system. The registry steward must then ask whether the data is represented indirectly in a way that can be transformed to meet registry goals. This is a potentially difficult endeavor that requires bringing together subject matter expertise in the clinical domain and in the relevant clinical systems. Further, even if the relevant data elements, or reasonable proxies, are available, they may be in formats that are difficult to use or transform. Additional decisions need to be made about appropriate data transformations that translate from the native clinical representation of the data to the one appropriate for the registry.

The problem tends to get more difficult as registries move away from common data elements to more domain-specific data elements. The former are more likely to have direct mappings from clinical representations than the latter. The considerable effort involved in mapping clinical data into a form suitable for research is one reason relatively few registries are using fully automated data feeds from clinical systems. Where the number of cases in the registry is small and the information source is diverse and specialized, it is often more efficient to create a manual chart abstraction process, where clinical staff enter data into an electronic data capture (EDC) form that contains the data elements necessary for a registry. While the duplicative data entry of common data elements (like date of birth) may be irritating, the cost of the duplication may be much lower than the cost of the informatics labor necessary to derive automated mappings.

The efficiency calculus is reversed for registries where the number of cases is large and the data sought is reasonably standard. For example, minimal mapping work would be required for a registry that is only seeking to ingest the Common Clinical Data Set. Thus, a registry can start with automated ingestion of common data elements, supplemented by manual entry for the domain-specific elements. The registry can grow over time to include more domain-specific elements in the automated data feeds, as they are clearly defined and mapped, assuming further automation is worth the difficulty. This path can lead to a desirable state where no manual data entry is necessary (beyond what is required for care delivery). However, we caution the reader that reaching this state may be expensive and that setting a hard goal of “no manual data entry” can make some registry initiatives unaffordable, especially if that goal must be reached at the very start of data collection.

There are three powerful models for simplifying data ingestion. The first is for the registry to specify an appropriate data submission standard, such as a FHIR API. Participating sites must then submit the data using that standard. The second is to specify a Common Data Model (CDM), such as PCORnet or OMOP, that all sites must maintain and create a process that pulls data from the local models. This approach relies on creating shared semantic models inclusive of the tedious mapping work that implies. However, because the CDMs are in wide use, much of that work has already been done by others. Of course, the CDMs can only include data elements that are already collected in structured form across different EHR systems.

As long as the registry can be limited to the data available in one of the CDMs, this approach is worth considering and may be very attractive in cases where participating sites already have CDM data stores, as is becoming increasingly common in major research hospitals.

The CDM approach can also take a “federated” form, where the data remains in local CDM data stores instead of being centralized, queries sent to the central query service are distributed (or “federated”) to the local stores, and the results from each local store are combined and returned to the central service. This federated approach adds some technical complexity, but may be justified as a way to solve data governance issues where sites are unwilling to allow the registry steward to centralize the data.

The third approach, structured reporting, captures registry specific data at the point of care distributed across the individuals responsible for the care of the patient, using the same mechanism to generate clinical documentation [39]. Rather than relying on an intermediary conversion step such as FHIR, data is directly captured per data dictionary specifications of the relevant registry and subsequently compiled and packaged for upload into the registry database. This approach has the advantage of being all-inclusive (i.e., all data needed by the registry is prespecified and thus collected) compared with FHIR while not requiring the pre-translation of data necessary to utilize the CDM approach (see Chap. 18 for further clarification). On the other hand, this approach requires that clinical staff enter additional structured data at the point of care to support registry purposes, beyond what they would normally enter for care delivery, and that mechanisms for doing the data entry are made available and convenient for clinical staff.

Registry Functions

Beyond data ingestion and mapping, registries must be architected to support a number of functions that allow them to deliver well-organized, curated data in a form that is useful for answering research questions (the HL7 CCRF domain analysis model represents the common registry-related tasks as functions). All essential registry functions must be in service of that final goal, often taking the practical form of an analytic data set ready for statistical processing.

The final output of a registry should be close to what the statistician Hadley Wickham described as “tidy data” [40]. A tidy data set is organized so that each row is an observation and each column is a scientific variable while eliminating redundancy in data representation. Tidy formats are very convenient for analysis and can be easily consumed by most statistical and data visualization tools. Data sets that are not “tidy,” in contrast, may take an unpredictable amount of time to clean before they can be used to answer research questions. This data cleaning process can consume more than 80% of a typical data analysis effort. Unless data is tidied up at the source, as in a registry, the data cleaning must be repeated for each subsequent analysis, wasting enormous amounts of valuable human labor. Thus a key goal of registries is necessarily the creation of curated data assets that can generate tidy

analytic data sets, ultimately reducing the effort required to conduct data analyses. When successful, a registry can reduce that effort from months to days, and do so not only for one analysis but for all broadly anticipated types of analytic questions.

Blumenthal uses the NQRN Clinical Registry Maturational Framework model to assess registry capability in the following domains: a function domain, which outlines the functionality designed into the registry in support of its purpose(s), as well as other domains that describe the registry capabilities that support this functionality [41]. The other domains include data collection scope, data capture and transmission, standardization and quality control, performance measurement, reporting, and participant support.

The most critical functions of a registry relate to the activities shown in Fig. 13.2. These include:

- Acquire and ingest: A registry must support various types of data and methods of data acquisition.
- Map/organize: The ability to map data to one or more core representations, such as a reference standard or common data model, that will allow further curation, enrichment, and transformation.
- Curate: The data in a registry must be of adequate quality for answering the expected class of research questions; this can mean detecting and eliminating data anomalies and excluding problematic observations and cases. Data curation is about improving the quality of the data already present. It usually involves redaction of cases and measures that don't meet a quality standard, e.g., removing cases or observations for missing values, missing inclusion criteria, and negative annotations about quality.
- Enrich: Data enrichment is about adding data to make the data asset more valuable. It comes in two flavors: endogenous and exogenous. “Endogenous data enrichment” transforms existing data into derived variables that are more informative and meaningful than the original data relative to the questions being asked. Often, these are the results of bioinformatics pipelines that create derived results. A typical, if trivial, example is the calculation of a summary scale score from a vector of sub-scores. More complex examples include calculation of biomarkers from biological or phenotypic observations. “Exogenous data enrichment” uses data ingested from additional sources to increase the value of the overall data asset. Often, it is a result of a “data integration” process. A typical example is adding patient-reported outcomes or social determinants data to clinical data. More complex examples may involve combining data from multiple studies, multiple registries, or new time points or adding descriptive information about biological entities from public databases.
- Transform: Often the data must be converted from one representation to another using standard methods. The curated and enriched data is often stored in an intermediate core model that is convenient for storage but not optimized for different kinds of analyses. For example, [42] describe a “pluripotent” storage model for clinical data repositories that records data elements as documents (specifically FHIR-encoded JASON objects) [42]. A pluripotent representation is

agile, but not immediately usable by analytic tools. The solution is to add transformation functions that generate analytic data marts from the agile representations. Ability to generate multiple data marts is an overlooked advantage. Each data mart can be optimized to fit particular kinds of analyses avoiding the painful compromises in data model completeness and simplicity necessary when only one data delivery model is available.

- Deliver: Make the data set available for analysis. Usually this involves creating secure query interfaces to the data for both visual and programmatic query use cases. It may also involve creating data request portals that manage the data access approval process.

The Future: Enabling the Creation and Use of Patient Registries for Biomedical and Health Services Research

Moving forward, it would be ideal to see the registry as part of the healthcare data ecosystem and a routine tool for LHS, as shown in Fig. 13.1. The most pressing informatics issues for registries are related to redesigning information flows and moving toward standards that will support the vision of native, interoperable data transfer from point of care to registries, defining the roles and responsibilities of all affected groups (clinicians, EHR documentation systems, registry owners). The HL7 and CIIC interoperability standards mentioned previously are designed to overcome interoperability challenges and streamline flow of data from clinical information systems into registries.

Of course, this would be greatly facilitated if the data elements used in patient registries were standardized. The realization of interoperable data transfer from point of care to registries depends on standardization of data elements for common concepts that span registries as well as the development of domain-specific CDEs for use across all EHR systems (not just for registry reporting). The best possible outcome is to encourage the adoption of standard data elements for common concepts that span across registries. The ONC Common Clinical Data Set [43] items are a starting point, but there are other elements that are generalizable enough to be of interest to many registries. This requires a consistent process for developing domain-specific CDEs as data standards for use across all EHR systems and secondary uses (not just for registry reporting).

This is an exciting time in terms of the number of standardization efforts that are making progress. However, there are many outstanding challenges that require collaboration, cooperation, and coordination across many different stakeholders. The ONC has largely focused on general data elements and UDI. The CIIC and HL7 CIC and CIMI are addressing general registry data elements as well as disease-specific data elements. The AHRQ is driving outcomes data elements, but of course will need a standardized set of clinical data elements as a foundation. The challenge for the efficient development and use of registries in the future will be how to align all of these efforts.

The most immediate challenge is how to encourage the adoption of standardized data elements for common concepts that span registries. We see a special role for a

collaboration between specialty societies and the ONC in defining domain-specific common clinical data elements. Specialty-specific content will make it much easier to get data into registries in the long term, and the regulatory pressure is necessary to make the EHRs comply. The HL7 standards development organization engages clinical experts and professional societies and is well positioned to enlist these groups for developing and promoting content standards.

Informatics methods and professionals are increasingly critical for the design of data transfer systems and registries. They can provide capabilities to trace the flow of data, understand its sources and provenance, and develop linking methods and approaches for assessing the “quality” (see Chap. 11) of different data sources and the certainty of patient linkage. The transparency of systems and processes enabled by information technology can enable patients to consent for their information being part of a registry and allow them to specify preferences regarding how their data is used over time. Implied in that consent, and enabled by information technology, is the monitoring and control of data uses. Patients can remove consent any time, leaving registry holders continuously accountable. New technologies, if designed to support thoughtful and proactive patient-oriented policies, can enable patient-controlled sharing of EHR data directly from healthcare providers or from patient-managed personal health records or patient-reported outcomes, which will contribute a rich source of patient-reported information to registries that would include various disease-specific outcomes and measures of functioning and quality of life. Data streams from physiologic or device measures could also be incorporated. One way forward is shown by the NIH investment in *All of Us* and the supporting Sync for Science movement: a collaboration of scientists and technologists that will undoubtedly show innovation and demonstration of tools for public to share their EHR data for research [44, 45]. Other federal regulations impacting registries (especially the 21st Century Cures Act) underscore and foreshadow the important role that registries will play in both research and the delivery of quality healthcare.

References

1. AHRQ. In: Gliklich RE, Dreyer NA, editors. *Registries for evaluating patient outcomes: a user's guide*. Rockville: Agency for Healthcare Research and Quality; 2010.
2. Travers K, et al. Characteristics and temporal trends in patient registries: focus on the life sciences industry, 1981–2012. *Pharmacoepidemiol Drug Saf*. 2015;24(4):389–98.
3. Muilu J, Peltonen L, Litton JE. The federated database – a basis for biobank-based post-genome studies, integrating phenotype and genome data from 600,000 twin pairs in Europe. *Eur J Hum Genet*. 2007;15(7):718–23.
4. Nakamura Y. The BioBank Japan project. *Clin Adv Hematol Oncol*. 2007;5(9):696–7.
5. Ollier W, Sprosen T, Peakman T. UK Biobank: from concept to reality. *Pharmacogenomics*. 2005;6(6):639–46.
6. Sandusky G, Dumaul C, Cheng L. Review paper: human tissues for discovery biomarker pharmaceutical research: the experience of the Indiana University Simon Cancer Center-Lilly Research Labs Tissue/Fluid BioBank. *Vet Pathol*. 2009;46(1):2–9.
7. Horsley K. Florence Nightingale. *J Mil Veterans' Health*. 2018;18(4):2–5.

8. Military Records. Civil war records: basic research sources. 2018 [cited 2018 July 1, 2018]. Available from: <https://www.archives.gov/research/military/civil-war/resources>.
9. Patient registries. In: DN, Gliklich RE, Leavy MB, editors. Registries for evaluating patient outcomes: a user's guide [Internet]. 3rd ed. Rockville: Agency for Healthcare Research and Quality (US); 2014.
10. CMS. Centralized repository/RoPR. 2018a. [cited 2018 June 23]. Available from: <https://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/CentralizedRepository-.html>.
11. FDA. Guidance for industry and FDA staff. Procedures for handling post-approval studies imposed by PMA order. Rockville: U.S. Food and Drug Administration; 2007.
12. Hollak CE, et al. Limitations of drug registries to evaluate orphan medicinal products for the treatment of lysosomal storage disorders. *Orphanet J Rare Dis.* 2011;6:16.
13. Clinical Trials Transformation Initiative (CTTI). CTTI recommendations: registry trials. 2017. [cited 2018 June 23]. Available from: <https://www.ctti-clinicaltrials.org/files/recommendations/registrytrials-recs.pdf>.
14. Stey AM, et al. Clinical registries and quality measurement in surgery: a systematic review. *Surgery.* 2015;157(2):381–95.
15. CMS. Quality measures requirements. 2018b [cited 2018 June 23]. Available from: <https://qpp.cms.gov/mips/quality-measures>.
16. Platt R, et al. Clinician engagement for continuous learning discussion paper. Washington, DC: National Academy of Medicine; 2017.
17. AHRQ. Bringing the patient voice to evidence generation: patient engagement in disease registries. (AHRQ Views. Blog posts from AHRQ leaders). 2018. [cited 2018 June 23]. Available from: <http://www.ahrq.gov/news/blog/ahrqviews/disease-registries.html>.
18. IOM. The learning healthcare system: workshop summary. Washington, DC: The National Academies Press; 2007.
19. ONC. Introduction to the interoperability standards advisory. 2018a. [cited 2018 June 23]. Available from: <https://www.healthit.gov/isa>.
20. Chute CG. Medical concept representation. In: Chen H, et al., editors. Medical informatics. Knowledge management and data mining in biomedicine. New York: Springer; 2005. p. 163–82.
21. ONC. 2015 edition certification companion guide. 2015 edition common clinical data set – 45 CFR 170.102. 2018b. [cited 2018 June 23]. Available from: https://www.healthit.gov/sites/default/files/2015Ed_CCG_CCDS.pdf.
22. NLM. The NIH common data element (CDE) resource portal. 2013. [cited 2013 March 6]. Available from: <http://www.nlm.nih.gov/cde/>.
23. CMS. Data element library. 2018. [cited 2018 June 23]. Available from: <https://del.cms.gov/DELWeb/pubHome>.
24. Sood HS, et al. Has the time come for a unique patient identifier for the U.S.? *NEJM Catalyst.* 2018.
25. Dusetzina SB, Tyree S, Meyer AM, et al. Linking data for health services research: a framework and instructional guide [Internet]. In: An overview of record linkage methods. Rockville: Agency for Healthcare Research and Quality (US); 2014.
26. 21st Century Cures Act. 2018. [cited 2018 July 1]. Available from: <https://www.fda.gov/RegulatoryInformation/LawsEnforcedbyFDA/SignificantAmendmentstotheFDCAAct/21stCenturyCuresAct/default.htm>.
27. Drozda JP Jr, et al. Constructing the informatics and information technology foundations of a medical device evaluation system: a report from the FDA unique device identifier demonstration. *J Am Med Inform Assoc: JAMIA.* 2018;25(2):111–20.
28. Campbell WS, et al. An alternative database approach for management of SNOMED CT and improved patient data queries. *J Biomed Inform.* 2015;57:350–7.
29. PheKB. 2012. [cited 2013 May 24]. Vanderbilt University. Available from: <http://www.phekb.org/>.

30. NLM. NLM Value Set Authority Center (VSAC). 2015. Feb 11, 2015 [cited 2015 March 11]. Available from: <https://vsac.nlm.nih.gov/>.
31. PheMA. PheMA wiki: phenotype execution modeling architecture project. 2015. [cited 2015 September 28]. Available from: http://informatics.mayo.edu/phema/index.php/Main_Page.
32. Richesson RL, et al. Electronic health records based phenotyping in next-generation clinical trials: a perspective from the NIH health care systems collaboratory. *J Am Med Inform Assoc.* 2013;20(e2):e226–31.
33. Richesson RL, Smerek MM, Blake Cameron C. A framework to support the sharing and reuse of computable phenotype definitions across health care delivery and clinical research applications. *EGEMS (Washington, DC).* 2016;4(3):1232.
34. Gliklich RE, et al. Registry of patient registries outcome measures framework: information model report. Methods research report, Prepared by L&M Policy Research, LLC, under Contract No. 290-2014-00004-C. Rockville: Agency for Healthcare Research and Quality (US); 2018.
35. Cochi SL, et al. Congenital rubella syndrome in the United States, 1970–1985. On the verge of elimination. *Am J Epidemiol.* 1989;129(2):349–61.
36. Tilling K. Capture-recapture methods – useful or misleading? *Int J Epidemiol.* 2001;30(1):12–4.
37. Rothman K, Greenland S. Modern epidemiology. 2nd ed. Hagerstown: Lippincott Williams and Wilkins; 1998.
38. AHRQ. In: Gliklich RE, Dreyer NA, editors. *Registries for evaluating patient outcomes: a user's guide.* Rockville: Agency for Healthcare Research and Quality; 2007.
39. Sanborn TA, et al. ACC/AHA/SCAI 2014 health policy statement on structured reporting for the cardiac catheterization laboratory: a report of the American College of Cardiology Clinical Quality Committee. *J Am Coll Cardiol.* 2014;63(23):2591–623.
40. Wickham H. Tidy data. 2014., 2014;59(10):23.
41. Blumenthal S. The use of clinical registries in the United States: a landscape survey. *eGEMS (Generating evidence & methods to improve patient outcomes).* 2017;5(1):26.
42. Chute CG, Huff SM. The pluripotent rendering of clinical data for precision medicine. *Stud Health Technol Inform.* 2017;245:337–40. Available from: <https://www.ncbi.nlm.nih.gov/pubmed/29295111>.
43. ONC. Common clinical data set. 2015. [cited 2018 June 25]. Available from: https://www.healthit.gov/sites/default/files/commonclinicaldataset_ml_11-4-15.pdf.
44. S4S. Sync for science (S4S). Helping patients share EHR data with researchers. 2018. [cited 2018 June 25]. Available from: <http://syncfor.science/>.
45. Sankar PL, Parker LS. The precision medicine initiative's all of us research program: an agenda for research on its ethical, legal, and social issues. *Genet Med: Off J Am Coll Med Genet.* 2017;19(7):743–50.



Research Data Governance, Roles, and Infrastructure

14

Anthony Solomonides

Abstract

This chapter explores the concepts, requirements, structures, and processes of data or information governance. Data governance comprises the principles, policies, and strategies that are commonly adopted, the functions and roles that are needed to implement these policies and strategies, and the consequent architectural designs that provide both a home for the data and, less obviously, an operational expression of policies in the form of controls and audits. This speaks to the “What?” and “How?” of data governance, but the “Why?” is what justifies the extraordinary efforts and lengths organizations must go to in the pursuit of effective data governance. This receives a fuller answer in this chapter; in brief, information is a valuable asset whose value is threatened both by loss of integrity, the principal internal threat, and by its potential for theft or leakage, compromising privacy, business advantage, and failure to meet regulatory requirements—the external threats. Internal and external threats are not quite so neatly distinguished in real life, as we shall see later in the chapter.

Keywords

Data governance · Research data governance · Information governance · Data integrity · Internal and external threats · Security · Privacy · Confidentiality · Regulatory frameworks · HIPAA · Common rule

The American Medical Informatics Association (AMIA) Clinical Research Informatics Working Group (CRI-WG). Acknowledgements: Judy Logan, WG Chair 2014–2016; Abu Mosa, Monika Ahuja, Kris Benson, Shira Fischer, Lyn Hardy, Kate Fultz Hollis, Bernie LaSalle, Nelson Sanchez Pinto, Lincoln Sheets, Ana Szarfman, Chunhua Weng, Chair Elect 2018–2020.

A. Solomonides, PhD, MSc (Math), MSc (AI), FAMIA (✉)

Department of Family Medicine, NorthShore University HealthSystem, Research Institute, 1001 University Place, Evanston, IL, USA

Introduction: A Conceptual Model

This chapter was originally conceived around a framework discussed by the members of American Medical Informatics Association's (AMIA) Clinical Research Informatics Working Group (CRI-WG). It finally crystallized in this form as a contribution to the present book. The framework is depicted in Fig. 14.1.

The schema in Fig. 14.1 places data and information at the center: the nature and context of data and information impacts the way it is governed, the functions that implement governance, and the underlying technology that houses, communicates, and defends it. The idea is that not only does each of these domains of activity demand attention in its own right, but the relationships and interactions between them also must be addressed. All relations are bidirectional: data governance adds to the data even as it "governs" it.

In the course of this chapter, we shall examine the qualities that give data its value, the life cycle of data, the vulnerabilities of data, and the implications of all these for the organization of "data governance."

What is data governance? As suggested in the model, it comprises the principles, policies, and strategies adopted, the functions and roles that—in the favored phrase of the domain—are "stood up" to implement these policies and strategies, and the

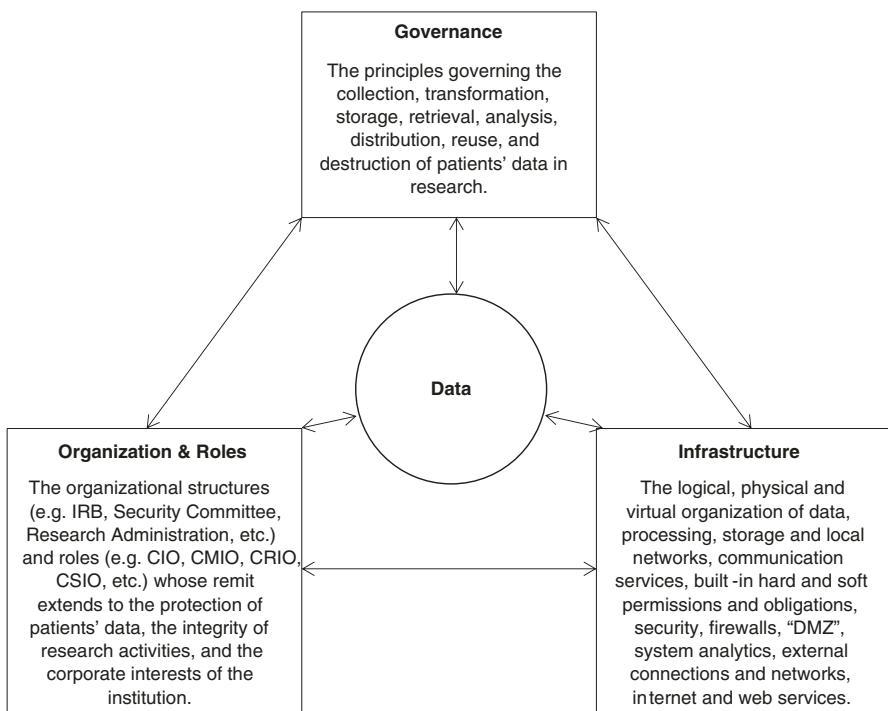


Fig. 14.1 The conceptual model. The three domains of data governance and their interactions

consequent architectural designs that provide both a home for the data and, less obviously, an operational expression of policies in the form of controls and audits.

This speaks to the “What?” and “How?” of data governance, but the “Why?” is what justifies the extraordinary efforts and lengths organizations must go to in the pursuit of effective data governance. This receives a fuller answer below, but in brief, information is a valuable asset whose value is threatened both by loss of integrity, the principal internal threat, and by its potential for theft or leakage, compromising privacy, business advantage, and failure to meet regulatory requirements—the external threats. Internal and external threats are not quite so neatly distinguished in real life, but we reserve this distinction for later in the chapter.

Research Data Governance

The principles governing the collection, transformation, storage, retrieval, analysis, distribution, reuse, and destruction of patients’ data in research.

In any enterprise, and in a healthcare organization more than most, data is literally an asset and, metaphorically, also a significant liability. The value of data can be realized in better business and care delivery decisions, in fulfilling a public health mission alongside provision of best care, in discovery of new knowledge through research, in improving quality and safety of patients, and in informing the healthy on how to maintain and enhance their health. The trouble with data is its vulnerability. If stolen by a competitor, it can damage a business irreparably, whether by identifying weaknesses in services offered or potential clients to be enticed away. In healthcare, if patients’ data is disclosed without authorization, there are consequences beyond loss of business and patients’ loss of confidence in the system: regulatory breaches bring fines and large settlements in their wake.

As a discipline, data governance delineates the (kinds of) principles, policies, strategies, functions, and actions that can guide and support the establishment of a coherent data governance program. As a practice, data governance aims to defend the value of the data in an organization, facing both inwards and outwards. The task for the institution is to assure the integrity of the data so that it does not lose its informational value. The task external to the institution is to protect the data from deliberate theft, accidental leakage, and inappropriate disclosure.

This chapter reviews more specifically the question of data governance for electronic patient data that is to be used for research. It would be more accurate to say, of course, “the questions” in plural form. To begin, there is no universal agreement on what constitutes data for research rather than data for the effective delivery of care, data for quality assessment or improvement, or even data for administrative transformation, e.g., through analytics. Thinking particularly of patients’ medical records, it is not even clear who “owns” it, notwithstanding ownership rights asserted both by patients and by providers. There is considerable variability on what is interpreted as “human subjects” research in different places, with consequences for informed consent requirements. (Indeed, as of this writing, there is some

uncertainty as to the exact requirements for consent following changes to the Common Rule¹ by the last and current administrations.²) Thinking of data, we must qualify our scope to “mainly” electronic patient data; some of the data may not be readily recognized as “electronic” (e.g., scanned paper documents whose content is not machine-readable). Further, powerful technologies and massive semipublic data repositories, including those of the social media giants, mean that secure de-identification of protected health information (PHI) remains an elusive goal.

What Does Data Governance Govern?

A succinct description of data governance may be framed in three dimensions: structures, processes, and results. The similarity to Donabedian’s dimensions of “quality” is not accidental [1]. Structures and processes are amenable to identical definition; “results” is broader than “outcomes.” Outcomes matter, but in the governance of information so do other results, such as aberrant behaviors and work-arounds. A search through the literature has not surfaced many publications that elide the preeminent framework for quality improvement with data governance, but it is not hard to see the parallels. Data governance is often paired with (and then barely distinguished from) master data management, and it is again the case that what these two have in common is the concern with the quality of data and data processes. Some, notably American Health Information Management Association (AHIMA), address these issues under the title of “Information Governance”[2]; [AHIMA] we briefly turn our attention to the ambiguity between data and information.

We shall draw—and blur—the distinction between data and information. Whether we speak of data governance or information governance, there are times when it is necessary to draw a distinction: data streaming from a device, for example, in the absence of a framework, is meaningless and may be thought of as simply data: *make of it what you will!* The moment that data stream is accommodated in a data structure that confers meaning to it—e.g., a column headed “Hourly Temperature” or more obscurely, “°C”—it becomes information. What has complicated this naive picture is the advent of data science in all its forms, from simple analytics to data mining and machine learning: with a little information about context, the possible meanings of a naked data stream may be guessed, so even where a useful distinction may be drawn in theory, it may be blurred in practice. Just as no

¹ Code of Federal Regulations 45 CFR part 46, subpart A, is known as the **Federal Policy for the Protection of Human Subjects** or the **Common Rule**. It is shared verbatim by a number of departments, hence “common.” See <https://www.hhs.gov/ohrp/regulations-and-policy/regulations/common-rule/index.html>.

² As of this writing, the status is described in the announcement “HHS and 16 Other Federal Departments and Agencies Issue a Final Rule to Delay for an Additional 6 Months the General Compliance Date of Revisions to the Common Rule While Allowing the Use of Three Burden-Reducing Provisions during the Delay Period” (<https://www.hhs.gov/ohrp/final-rule-delaying-general-compliance-revised-common-rule.html>).

physician would have difficulty guessing what the sequence 39.4, 39.8, 39.4, 38.9, 38.6, 38.2, ... likely means, a sophisticated machine learning algorithm would probably get there too.

Information is, in our definition, data organized in a way that imparts or reflects meaning. This gives information an abstract spatial quality. In this light, information means not only the (raw) data, but the meaning that renders it into information. This forces us to consider metadata on a more or less equal footing as data itself. This is reflected in the data manifold (see Fig. 14.2 above). A note of 144/102 in a patient's chart may give the appearance of a vulgar fraction, but to the knowing eye it has as very specific, indeed, highly significant meaning. How that meaning will be translated into machine-readable form—a form in which a software application can take it as its input and generate some valid output—is the result of a cascade of design decisions which also ultimately impact the governance process. Likewise, social scientists, especially social constructivists, may assert with some justification that all data is theory-laden. Grounded theory [3] notwithstanding, most data is collected with a theory of some sort in mind. We shall evade this dilemma by our convention that data becomes information in the light of a theory, however lightly that theory may be asserted—perhaps only implicitly through the headings at the top of columns of data.

In the temporal dimension, information governance spans the life cycle of the artifacts called *information*, including their creation (or capture), organization, maintenance, transformation, presentation, dissemination, curation, and destruction. The information governance process therefore treats data not only in its spatial aspect but also through its temporal dimension.

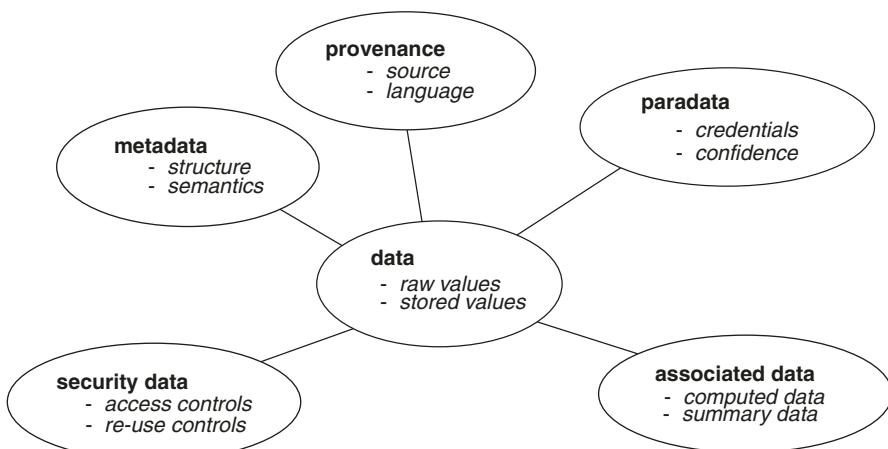


Fig. 14.2 The data manifold. Data is characterized not only by its values but also by what is loosely termed its “metadata,” which can be analyzed into metadata proper, provenance data, paradata (e.g., concerning the credibility of the data), security data, and various computed summaries, and so on

“Knowledge” is beyond our consideration, but it is often confused with information or placed in the putative hierarchy of “data-information-knowledge” (to which “-wisdom” is also added to make matters even more obscure). Knowledge is a human attribute: to quote Laurence Prusak, one of the founders of the knowledge management school, “there is no knowledge but that which a knower knows” [4]. What concerns us in this review is information in the sense of data whose meaning is derivable from its form, whether that form was deliberately constructed or imputed through some algorithmic process. Information is capable of being processed by machine. Except where the distinction makes a difference, we shall use data and information as synonymous, and, when necessary, the distinction will be made explicit.

Why Data Governance?: The Value of Data

Information is a resource. Decisions can be made at least in part on the basis of information in our possession. Information can be mined for patterns that lead to hypotheses about how something works, or fails to work, or how a pattern of behavior may contribute to the development of a condition. The value of this resource, therefore, depends on certain characteristics it may possess to an absolute degree (e.g., correctness) or in some measure (e.g., relevance). These characteristics have been painted slightly different by different authors, but in broad brush they agree that data must be accurate, valid, reliable, timely, relevant, and complete.³ Data must also be available, if it is to be useful, but along with many others, we treat that as a system rather than a data characteristic.

Accuracy Inaccurate data can scarcely be described as “information,” so accuracy or correctness is essential, but this criterion encompasses certain subsidiary characteristics. Any units that are used must be explicitly defined. The data must be sufficiently granular and precise to the degree necessary for its use; ideally, it must also be unique in the sense that a system must afford a single source of truth.

Validity Data must conform with any restrictions on the values it may take, and any relationships that are prescribed between such values: in database parlance, the data must conform with certain integrity constraints. **Legitimacy** is sometimes added to this category; it is all the more important here in the context of governance. To that end, it is often desirable to be able to reconstruct a trail back to the source of the data, a form of metadata known as provenance.

Reliability Data must be both self-consistent—integrity constraints and single source of truth contribute to this—and consistent with its environment, such as the

³This simple list was promoted to public bodies in the United Kingdom by the now dissolved Audit Commission. The elaboration in this chapter is the author’s, based on contributions from numerous authors.

applications that must use it. Where a transformation is necessary to address the requirements of an application, the validity of that translation must be assured and the transformation itself be logged in provenance.

Timeliness Data is often time-stamped, meaning that the time of its collection or entry into the system is itself recorded. Any significant time lags or delays, or any gaps, affect the usefulness of the data, especially if any data-driven decision is to be made. There should thus be minimal delay between any event and its record and minimal latency in providing the record for use.

Relevance Data is normally collected for a purpose. It is both good practice and a common regulatory requirement that a **principle of parsimony** be adopted in data collection: all the data that is required—all salient data—and only those. The **accessibility** of data, including the navigability of the architecture holding the data, is considered by some to be an aspect of relevance.

Completeness Complementing the principle of parsimony is the principle of completeness which asserts the need for the data—more precisely, for the data model—to be comprehensive, i.e., to provide as complete a picture of the entity it relates to as is necessary for its purpose. As in all modeling activity, salient features must be selected for inclusion; this is a matter of purpose and subject to scientific disagreement.

In our conceptual model of data, the data manifold (Fig. 14.2), we have distinguished between what may be termed raw values and a collection of what are often loosely called metadata—data about data—but classified into categories reflecting a purpose: *provenance*, to show how the data came from or was created; *metadata* proper, which portrays the semantic relationship between content and structure, for example, the relationship between attribute names and values; *paradata* which may be associated with confidence in the data; *security and privacy* data, reflecting access and use privileges; and *associated* data, mainly summaries of raw data.

The Life Cycle of Data

We have asserted that data governance principles, policies, structures, and functions address all phases of the data life cycle. Typically, we consider these to be collection (or creation), transformation, storage, retrieval, analysis, dissemination (or distribution), transmission, reuse, and destruction of data. In our case, we may think of these specifically as patients' or subjects' data in research.

Each of these phases in the life of data entails some threat to the integrity of the data. Poor collection practices threaten both the legitimacy and the accuracy of data; data from an inappropriately credentialed laboratory may be worthless; poorly maintained instruments may compromise precision; a copy of data collected on a portable device may remain insecurely in that device even after it has apparently

been uploaded to a secure system—the very word “uploaded” gives a false sense of security. Data is not like boxes on a dock being uploaded onto a van.

Considering creation and transformation, we know that software does not always function as intended or as designed. Even at the creation stage, habitual users of software are aware of invisible transformations that may occur when entering data (think, e.g., of presentation vs. storage formats for dates in Excel; consider the metadata needed to ensure that a date entered in US format reads correctly in a European-installed copy of the program). Data transformations undertaken in the service of analysis or dissemination likewise can cause problems. Notoriously, mix-ups between unit systems can cause catastrophic failures.

Storage in the relatively short term is highly reliable, but long-term storage is technology dependent and may provide another source of error or effective loss. If an organization considers that data still have value, then appropriate curation is necessary to ensure its retention. When the data no longer has value or there is no legitimate reason to keep it, the data must be securely destroyed: description of the method of destruction and oversight that the necessary steps are taken often falls to a data governance function. Encryption of stored data is often required as a minimal defense against theft or leakage outside a secure perimeter.

Data analysis is often carried out using specialized software packages, including statistical tools, data analytics, de-identifiers, natural language processors (from simple concordances to highly sophisticated NLP tools), visualization, and more. The integrity of these processes is, of course, a concern and a matter for the researcher, but they also pose a challenge to a data governance function to ensure that there is no inadvertent leakage or disclosure through the use of these tools. Since these are often proprietary and function as a “black box,” it is necessary to trial such software under controlled conditions in a suitable “test harness” that captures all traffic in and out of the application.

One of the principles of grid computing, and subsequently cloud computing, is the notion that when the data cannot be sent to the algorithm for whatever reason—in the case of healthcare, because it may be protected health information—there is provision for the algorithm to be sent to the data. There are some issues with this, both in terms of licensing—do all the sites need a license for any proprietary software involved?—and in technical terms, can the distributed results be legitimately aggregated? Some remarkable work has been emerging in this area [5].

Data sharing and publication are a particular challenge to a data governance function. Poor programming practices can lead to information leakage and to vulnerabilities in, for example, publication through a website, including the possibility of intrusion, malware injection, and other forms of attack. Other means of sharing, such as direct transmission of data, pose well-known security problems, including interception and corruption. Just as secure storage is typically encrypted, encryption of data for transmission provides a degree of security. However, technological advances threaten even this defense. Data may also be compressed prior to transmission to reduce its volume; depending on the nature of the data, a decision has to be made about the degree of “loss” of definition that can be tolerated in compression.

Why Data Governance?: From Data Protection to Research Ethics

While data in all its life cycle stages must be protected from error and unintended loss of integrity, it must also be defended against deliberate attack and against careless mishandling resulting in disclosure. Data needed to support business functions is not only valuable to the owner organization but is also of considerable interest to its competitors. This includes very basic data, such as details of patients and the conditions they suffer from or the specialist physicians they see. The pervasiveness of security requirements is a consequence of the digital transformation of business and of healthcare in particular. When records took the form of paper files, inappropriate disclosure meant misplacing a file and information theft meant stealing it. When we spoke of security, we meant physical security—locks and keys. The digital economy has brought with it a need for a security function of a very different kind, but the jargon of physical security has been extended to the digital variety.

By far one of the largest concerns in a healthcare organization is the protection of personal health information. The complexities of research (such as the need to “blind” studies) makes biomedical and healthcare research data management all the more fraught. This is the case in virtually all developed healthcare systems, although the jargon may differ from place to place. We shall adopt US usage, where such information is described as *protected health information* (commonly, *PHI*). In the American context, two regulatory frameworks weigh heavily on the policies and practices of healthcare organizations that engage in research: the **HIPAA** rules and the **Common Rule**. Although at the time of writing there is some uncertainty concerning the final shape of the Common Rule, the general principles, which would apply, suitably translated, in most jurisdictions with a research culture, can be outlined with some certainty.

The Health Insurance Portability and Accountability Act [6] formalized privacy requirements for any “*covered entity*” that handles patient information in electronic form. Covered entities include all providers who transmit patient data in electronic form, health plans, and healthcare information clearinghouses. When a third party is employed by a covered entity to process any PHI on its behalf, it must enter into a binding business associate agreement (BAA) with that third party, so that its handling of PHI is also ruled by HIPAA. For example, some academic medical centers that are not an integral part of their associated university have a BAA to enable academics to work with—and in particular to do research using—PHI. Pharmacy benefit managers and health information exchanges also normally operate subject to a BAA with their associated covered entities.

The **HIPAA Privacy Rule** is designed to protect individuals from harm that may be sustained through the inappropriate disclosure or illegitimate use of personal information. The scope of this protection is considerable: the individual may suffer harm from causes ranging from identity theft, through medical insurance fraud, to denial of health insurance coverage because of “known” (i.e., disclosed) existing conditions—including now genetic information which has complicated matters further still. The Privacy Rule allows for the possibility of de-identification of patient information: this may be accomplished by one of two methods—one is the so-called

Safe Harbor method which requires the removal of 18 specified types of identifiers as well as any other data that may lead to reidentification. The second method is through Expert Determination: a statistical expert must testify that by application of scientific principles, it has been determined that there is negligibly small risk that the anticipated recipient of the data would be able to identify an individual.

Supporting the goals and implementation of the Privacy Rule, HIPAA adds a Security Rule. This requires the operational, logical, and physical structure of the information function to be secured against known and foreseeable challenges. We term the function that defends against deliberate attack, inappropriate disclosure, and leakage of information the security function. By the very nature of the asset we are seeking to protect—information—security has to take many forms and be implemented at many levels, from low-level protection systems in the sense of close to the physical infrastructure, through authentication protocols for authorized users, to authorization processes and allocation of access rights, finally to an individual or, more likely, a committee charged specifically with high-level decision-making on the release of data. Since, as implied here, security also encompasses infrastructure systems and networks, the entire information architecture, physical, logical, and operational, is subject to the requirements and dictates of security. We shall see that the various demands of privacy and security (and confidentiality, as we shall add) have led to the creation of a number of distinct roles in healthcare organizations, all of whom bear the words “information officer” in their title, sometimes leading to confusion as to their exact purpose and responsibilities. We shall argue below that provided role descriptors are clear and any overlap in duties is managed, none of these roles is superfluous.

We now turn to the second framework with direct relevance for research, that of the Common Rule, as codified in Federal Regulation 45 CFR part 46. The Common Rule is so-called because it is adopted “in common” by 18 agencies, although its development is normally led by the Department of Health and Human Services (HHS).⁴ The primary purpose of the Common Rule is to protect human research subjects in studies funded by any of these 18 agencies, but in practice most institutions apply the Common Rule to all research, irrespective of funding source. The Common Rule offers protection against physical and informational harms: in particular, it encompasses all the stages in the life cycle of data—collection, use, maintenance, and retention—and how these may impact a research subject’s physical, emotional, or financial well-being or reputation.

An institution may obtain a *Federal-Wide Assurance* (FWA) asserting that any research funded by the 18 agencies (or all research, for that matter) will be conducted in full compliance with the provisions of the Common Rule. The Office of Health Research Protections (OHRP), an office of the DHHS, describes the FWA as “the only type of assurance currently accepted and approved by OHRP,” through

⁴At the time of writing, the Common Rule is subject to revision. A revised rule had been approved on the very last day of the Obama administration, but this was suspended for review by the incoming Trump administration. Recent (April 2018) indications are that the Obama rule may be amended before it is implemented.

which “an institution commits to HHS that it will comply with the requirements in the HHS Protection of Human Subjects regulations at 45 CFR part 46.” A critical step in obtaining a FWA is the registration of an *Institutional Review Board (IRB)* who must approve all research involving human subjects, whether it involves a clinical trial or processing of subjects’ identified personal health information. As an alternative, it is also possible for an institution to nominate an established IRB as the one on which the institution will rely for approval of its research. Either way, the IRB must approve all research using identifiable data of living individuals with the aim to establish new knowledge. Approval by an IRB ensures that subjects will be informed of the nature, process, and risks of the research and that on the basis of this information, subjects freely consent to participate and know that they have a right to withdraw at any time. Consent may include an indication of future work that may be undertaken using the same data. However, “broad consent,” in the sense that it allows researchers freedom to use the data for other studies without returning to the subjects for a fresh consent, has not hitherto been allowed.⁵ Some studies undertaken with a view to quality assessment or improvement and not primarily intended to generate new knowledge may be exempt from IRB approval. Likewise, studies regarded by the IRB as posing minimal risk, or using fully de-identified data and so deemed not to be human subjects research, may be exempt from, or subject to a lighter “expedited,” IRB review. The IRB is charged with continuing to monitor research studies both for noncompliance and for any unanticipated risks that arise in the course of a study. Through the mechanism of FWA and IRB review, the OHRP retains considerable powers to discipline any noncompliant entity. IRBs are subject to periodic review and are accountable for their record.

As well as PHI, privacy frameworks recognize a further category of data, *personal identifying information (PII)*. The distinction from PHI is implied in the descriptor: many of the data elements that Safe Harbor requires to be removed are PII. Personal demographics, dates of birth, telephone numbers, and so on do not impart health information but can readily identify an individual. De-identification in some cases has to be done in a way that can be reversed under very strict conditions. For example, a patient whose record appears suitably redacted with a randomly generated identifier may need to be contacted, either because something very serious has been observed (a so-called incidental finding) or because he or she meets certain criteria and is therefore a candidate to be consented for a deeper study. The linking information is sometimes entrusted to a neutral role in the institution, often approved through the IRB: the *honest broker*. The honest broker is entrusted with the link between the institutional identifier of a patient (e.g., the medical record number) and that patient’s randomly generated pseudo-identifier. It is possible to arrange for the honest broker to know nothing more than that link, i.e., no PHI at all. This also provides a means to protect confidentiality.

⁵The Obama rule and the revision still under current consideration do allow for broad consent in some cases. As embodied in this rule, broad consent is thought to place a considerable burden on the institution to maintain awareness and monitor its application.

Confidentiality of personal health information extends the concept of privacy to a principle of parsimony concerning the sharing or dissemination of data. Simply stated, confidentiality requires data to be disclosed on a strict need-to-know basis. Initially this arose from considerations concerning certain stigmatizing conditions: does a medical assistant rooming a patient for a visit need not know that he has suffered from severe depression in the past? Indeed, certain kinds of data are often treated as privileged—HIV status, mental health—but this is not uniform. However, when subjects are involved in a clinical trial involving an intervention and the trial is itself “double blind,” the emergency room physician faces a real problem when that subject reports to the ER with an acute neurological complaint of no known cause. Within integrated systems, the patient’s electronic record may include a flag indicating that a patient is indeed involved in a trial, so that in a worst case scenario the record may be unblended to provide the treating physician with knowledge of what treatments, especially medications, the patient had received prior to his being taken ill.

Theories of Information Governance

In its most abstract sense, governance is a theoretical concept referring to the actions and processes by which stable practices and organizations arise and persist.

Wikipedia—entry on Governance

To govern is to manage, to control, to direct, and to steer. We tend to think of “government” as made up of the authoritative structures of regulation and control, while “governance” reflects the *process* of regulation and management. In this chapter, we have taken this broad view of the term as our scope, so as to provide a wide perspective that captures all the activities that may fall under the term, at least in as far as it relates to research. In this section, we venture a little further into the realm of legal and socioeconomic analyses of privacy so as to locate information governance in its broader context.

We can argue that information governance is driven by two forces: what may be loosely called data management or data stewardship—maintaining the integrity and safety of the data—and *privacy and business protection*, defending sensitive data from disclosure, leakage, and theft. Security, as an active program to defend the business from attack, touches on both. While the operational structures to maintain the integrity of the data are readily seen as necessary, the concept of privacy as a driver for information governance is often misconstrued. Is the entire governance “enterprise” really necessary? How does the need for privacy arise? Concerns about identity theft and medical fraud on one hand and a patient’s “ownership” of her medical record each contributes, but they have their roots in alternative conceptions of privacy.

James Whitman [7] describes a dichotomy between two privacy cultures which he codifies as dignity vs. liberty and locates, respectively, in Europe and the United States. This thesis begins with the observation that many authors have difficulty

defining privacy in exact terms, often relying on allusion to make the case for privacy: “It is the rare privacy advocate who resists citing Orwell when describing these dangers”—threats to “fundamental rights [7].” The slipperiness of the concept can also be made “by citing a large historical literature, which shows how remarkably ideas of privacy have shifted and mutated over time [7].” And the contrast between European and American sensibilities is pressed home: “Why is it that French people won’t talk about their salaries, but will take off their bikini tops? Why is it that Americans comply with court discovery orders that open essentially all of their documents for inspection, but refuse to carry identity cards?” Whitman traces these differences to “intuitions that reflect our knowledge of, and commitment to, the basic legal values of our culture.”[7].

But what is it that must be kept private? The foundational paper on privacy by Warren and Brandeis [8] was conceived on the advent of photography and the danger that one’s image may be captured unawares. From here, it is a fairly straightforward leap to the loss of privacy through the inappropriate disclosure of personal health information. Curiously, there is a quasi-symmetrical concern with the person being forced to witness something inappropriate about others, as in the occasional system message that images have been removed from an email to protect privacy. Loss of privacy in these senses appears to mean, primarily, a loss of dignity, from an image of the subject with company he may wish not to acknowledge, to a revelation of an embarrassing condition in the medical record.

lives, and the personal health record is not so different from one’s home.

The instinctive response to this is to claim ownership of the personal health record, a tenet apparently bolstered by the law, although the complexity of who owns and who is the custodian of the record muddies things considerably. Positions on this are easy to polarize. How can the culture of the “learning health system” be promoted if citizens claim ownership of their health data and wish to hoard them? How can an individual claim that her data has been “stolen” if she does not own her medical record? But if the patient owns her medical record, what was the physician’s intellectual contribution to that record? After all, the patient did not diagnose herself—it was a physician with 7 years’ solid training and more years’ experience who did that.

This observation gives us a handle on the second contrast we must reckon with. This is presented here in terms of Viktor Mayer-Schönberger’s opposition of a systems-based theory of information governance to the prevailing rights-based view [9]. Mayer-Schönberger turns his attention to the protection of intellectual property (IP) as a means to break the deadlock over privacy rights. Like Whitman, he begins by observing differences between continental European conceptions of privacy rights and American ones, and in the interests of an international information economy, he seeks commonalities between them. In Europe he recognizes complementary moral and economic dimensions to information rights, while in the United States, he notes a trend toward “propertization.” European modes of control over information relating to an individual, such as the legal “right to be forgotten,” are expressions of a moral commitment. American legislation is a diffuse mix of federal, state, and case law which makes control over personal information all but

impossible in practice. Notwithstanding these differences, he finds little empirical evidence that these rights are much acted on in the courts. He comments wryly, “Perhaps hoping for individuals to enforce their rights through costly court action is too ambitious a vision, and thus the problem lies in the governance mechanism used to afford information privacy” [9].

Looking at the United States, he finds a more interesting contrast between the ways in which information rights and privacy rights are codified. Intellectual property rights serve twin purposes: economic and moral protection of the author, on one hand, and a stimulus to trade, on the other. While privacy rights are essentially inalienable, IP rights can be transferred—sold or licensed—even as the author retains the *moral* right to be identified as such. What would be the conditions under which personal information could be treated as property, as something title to which can be meaningfully transferred? As far back as 1998, writing in a computing journal, Kenneth Laudon proposed a market for private information: “Who owns and controls personal information in national data networks? Why not let individuals own the information about themselves and decide how the information is used? A regulated National Information Market could allow personal information to be bought and sold, conferring on the seller the right to determine how much information is divulged [10].” Chronologically, this coincides with proposals for personal health records, in the sense of records that may be “banked” by individuals, much as they bank their money and protect their financial interests, which were put forward both in the United States and the United Kingdom. Mayer-Schönberger’s argument is complex and nuanced, but in essence he advocates for information rights that are governed by a “systems-based” approach. This envisions a “thick network” of professionals and formal and quasi-formal bodies that would mediate informational transactions, much as various bodies handle copyrights and patents. Exemplars of such individuals and bodies are drawn from European practice, where there are “information commissioners,” “data guardians” (cf. Caldicott⁶ Guardians in the UK NHS), and others that play an active role in the maintenance of privacy through audits and public reporting. This approach has the potential, both to secure privacy rights by enforcing protections and to allow the economic value of the data to be realized in a fair marketplace. Indeed, Mayer-Schönberger notes that intellectual property rights are not typically defended by individuals, but by organizations established for that purpose.

Data Governance Organization and Roles

The organizational structures (e.g., IRB, Security Committee, Research Administration, etc.) and roles (e.g., CIO, CMIO, CRIO, CSIO, etc.) whose remit extends to the protection of patients’ data, the integrity of research activities, and the corporate interests of the institution.

⁶Instituted following The Caldicott Committee. Report on the Review of Patient-Identifiable Information. December 1997. UK Department of Health.

Management and regulatory oversight duties and functions in an institution are likely to be distributed among senior post-holders and committees, the former where direction is perceived to be a senior management responsibility, the latter where expert consensus as well as executive fiat may be necessary.

The commonest roles in most institutions are those of the Chief Information Officer (CIO) and the Chief Medical Information Officer (CMIO). The CIO is typically a career technical administrator who has risen to a “C-suite” executive role, while the CMIO is typically a medically qualified and still active physician who acts as a bridge between the technological functions of the organization and the body of physicians (often a medical group) in whose service the technology has been introduced but who are often highly critical of it. The CMIO likely reports to the President or Chair of the medical group as well as to the Chief Executive or Chief Operating Officer. In recent times, the role—even the very concept of the CMIO—has been challenged as too narrow. Some have advocated for the broader concept of Chief Clinical Information Officer (CCIO) which would encompass also the Chief Nursing Information Officer and rarer roles, such as the Chief Pharmacy Information Officer. Elsewhere, the role of CMIO has been redefined as that of the Chief Health Information Officer (CHIO), implying a higher level role, not so much in terms of the organizational hierarch as in terms of the types and breadth of information that the post-holder should be concerned with. Few institutions have all these roles, but almost all have at least one—which one reflecting history and organizational preferences. In the commonest setting, where there are both a CIO and a CMIO, they are likely to divide their attention, respectively, between systems, networks, and technical employees for the CIO and the conceptual design of the information, its integrity, and the way it is entered by and presented to physicians.

Among committees, the Internal Review Board (IRB) is necessary wherever research is done; since quality assurance and quality improvement work often includes elements of research, the IRB is essentially universal. In many organizations, a data governance committee may be established, whose role is significantly narrower than the “data governance” discussed in this chapter: it is restricted to determinations of whether data has been sufficiently de-identified, or the intended recipient of the data is appropriately credentialed, and other granular decisions of this nature. A Security Committee may oversee data requests and releases from the viewpoint of technical security or, more likely, from the viewpoint of business sensitivity.

The organization of the infrastructure may also entail the creation of certain particular roles. Some of the functions of the narrow “data governance committee” just described may be delegated to an individual role, often designated the “honest broker.” Sometimes honest brokers are appointed through a formal IRB process, but in many places the designation is ad hoc. Honest brokers are most frequently associated with project- or program-specific repositories, as in the case of PCORnet data marts or where a so-called “pluripotent” database, i.e., one capable of supporting many projects, is established.

Most institutions with a developed health IT infrastructure are able to differentiate a number of different components: a transactional system, used by providers and

administrators to record patient-related data, often based on an encounter or on a report from a lab, from pathology, or radiology. This will generally tend to be reorganized directly or overnight into a well-structured database. This may be good enough to serve as the “single source of truth” but is often so exquisitely normalized that its navigation is extremely laborious, and so queries run very inefficiently. Thus there is a need for a data warehouse, i.e., a collection of denormalized, flat, materialized views on the data, the so-called data marts. These can be searched very efficiently either through a standard query language or through a specialized tool, thus making it accessible to non-expert programmers. Finally, research needs may be addressed in a variety of ways. At some institutions, the same enterprise data warehouse also serves for research but is therefore highly restrictively controlled. Elsewhere, a de-identified copy of the data warehouse is created especially for research. Here access is less restricted, but access to patients for consent requires additional steps. Finally, there are several national projects which have promulgated specific data models which must be adopted in order to participate. These include the PCORnet Common Data Model, OHDSI/OMOP mandated by the All_of_Us “precision medicine” project, and i2b2 adopted by several CTSAs and other collaborations. In addition to these are the numerous ad hoc collaborations which succeed in data sharing though data sharing and data use agreements. In all these cases, some governance mechanism is deployed to ensure ethical, data release, and security approvals are obtained.

Implementation: An Effective Data Governance Structure

The design, deployment, and maintenance of an effective data governance program is a major undertaking. Addressing all the relevant issues in an enterprise-wide set of structures and processes requires an in-depth understanding of concepts and requirements from so many domains that it is almost always best left to a team with diverse backgrounds, each expert in his or her domain. There are rare professionals who have specialized in this area and whose expertise is highly valued. An informationist charged with implementing a data governance program, especially one in healthcare, would be well served by a comprehensive guide book and a team of knowledgeable fellow professionals who can cover the specialist topics: knowledge of legal aspects of data protection, knowledge of security frameworks, and knowledge of the enterprise and its culture, not least the often competing constituencies within a single enterprise. There is a choice of guidebooks on data governance.⁷

⁷John Ladley. *Data Governance: How to Design, Deploy, and Sustain an Effective Data Governance Program*. Morgan Kaufmann, 2012. A readable, comprehensive guide to the broad spectrum of data governance—recommended.

David Plotkin. *Data Stewardship: An Actionable Guide to Effective Data Management and Data Governance*. Morgan Kaufmann, 2013. Puts the onus for data governance on data stewards; this may be somewhat narrow for healthcare institutions.

Helmut Schindlwick. *IT Governance: How to Reduce Costs and Improve Data Quality through the*

The building blocks of an effective strategy: Case Study

In a report to the AAMC conference on Information Technology in Academic Medicine in 2016 and again in an AMIA CRI-WG Webinar [11], a university medical center team reported that when they began work on the creation of a data warehouse without a parallel data governance effort, they were hampered by a number of problems. These were in essence the common problems that have led to the establishment of data governance structures and processes in many organizations, reflecting both the need to protect the data from error, redundancy, and inconsistency, as well as to defend the data from accidental or malicious disclosure. Crucial headline issues they identified included ill-defined responsibility and ownership of data, along with a lack of standards and consequent mistrust of the data; they also found data replicated across multiple silos, with inconsistent integration, and noted that there was no enterprise-wide data quality audit, so that errors were not systematically traced back to their origin; and there was no information life cycle management. They were also troubled to find little understanding of the data across business lines—the clinical enterprise, the research enterprise, the academic/student enterprise, and even the finance enterprise. Thus they were persuaded of the need for a data governance process. More accurately, they understood that their need for consistent, reliable data across all business units led inexorably to the initiation of a broad data and information governance program that would address consistency in the management and use of institutional data, as well as transparency on its provenance and semantics. Moreover, it would also lead to better performance in responding to users and improved business analytics. This section of the chapter relies heavily on the experience of this team and its very well-laid out history of the development of its data governance program.

The team adopted Gartner's definition [12] of information governance as “the specification of *decision rights* and an *accountability framework* to ensure appropriate behavior in the valuation, creation, storage, use, archiving and deletion of information. It includes the *processes*, roles and policies, *standards* and metrics that ensure the effective and efficient use of information in enabling an organization to achieve its goals” (emphasis added) [@REF6]. Basing their program on this definition, they addressed the first component and determined questions of accountability and specified ownership, roles, and responsibilities. They then engaged key stakeholders across the institution to ensure that decisions would be adhered to. Their second focus was on standards: they specified expected data quality standards and a pragmatic margin of tolerance. Their standards addressed information consistency, the models of the data and their contexts, protection, and the life cycle of the data to ensure liveness while retaining a manageable volume. Finally, in what they saw as the most important component, they turned to processes: decision-making guidelines and protocols, agreement on escalation process for decision resolution,

Implementation of IT Governance. CreateSpace, 2017. Highly recommended by some business leaders, it seems to restrict its purview to IT-related matters.

Robert S. Seiner. *Non-Invasive Data Governance*. Technics Publications, 2014. Appears rather more authoritarian than its title may suggest.

communication and workflows, and change management. The team asked three questions: *What* decisions need to be made? *Who* makes them? *How* are they made? These questions focused the team's attention on data as an enterprise asset and that it is worthwhile investing in its stewardship.

Focusing these questions on particular domains, four basic domains were identified: data, metrics, tools, and funding. In the case of data, a number of decisions had to be made: which is the system of record for source data? What is the tolerance threshold for different types of data—patient counts may need to be accurate plus or minus N, perhaps, but financial data must be as accurate as possible. What data transformations are allowed, and what relationships must be preserved? What access approvals are required, and who is authorized to grant such approvals? If, as is the case in many academic medical centers, there are multiple coexisting enterprises—clinical, educational, research, business—how is consistency maintained between them? In this particular case, the local decision grants the data steward at the source continuing stewardship of those particular data as it migrates, e.g., to the data warehouse.

Turning to values and metrics, it is necessary to pay attention to different ways of defining units in different business areas: a faculty “FTE” (full-time equivalent) in academics may not be the same as a faculty FTE in clinical; dates and times of events is another well-known area of divergent definitions. There are data benchmarks, both internal and external; again, a choice has to be made on who will be responsible for maintaining these. In the present case study, the relevant source data steward retains this responsibility and so ensures continuity. This responsibility stays with the steward for that element of data right up to when it contributes to a dashboard report to management. For the last two domains, tools and finance, in this case study, the recommendation is, first, to make sure that technical professionals are involved in all tool choice decisions and that business management is on board when there is likely to be a need for funding.

Drilling down into greater detail, the team created a “decision matrix” with a horizontal axis of the four domains (data, metrics, infrastructure and tools, infrastructure funding), each broken down further by the enterprise area (system-wide, education, research, clinical, faculty) so that there are 20 columns in all. The vertical axis represents the data stewards and possible decision-makers in the organization: some c-suite executives with informatics or operational responsibilities, deans, associate vice-presidents with relevant portfolios, etc. In each box in the matrix, an entry identifies members, decision-makers, veto-holders, and information providers, and those must be informed of any relevant decision. This tool provides the medium of negotiation of roles and determination of who should be the data steward for each element. In reality, each data element requires attention in this way, so the process has to break down responsibilities at least one more time to get to a clear determination of who has ownership of what. Indeed, in conclusion, the team has observed that there are three rings of data, the inner ring of master data which is shared across all business areas and has to be governed collectively; the middle ring of shared application data which may belong to one functional area and governed locally; and finally, the outer ring of single application data, managed by the small number of concerned individuals. A sophisticated approach quantifies

responsibilities for data elements and so assigns the role appropriately. Master data is determined by exclusion as well as by inclusion: certain data elements may be useful or important, but they may not be “master data” because they change frequently or relate to specific attributes.

Acknowledgments In addition to the members of the AMIA CRI-WG, I must acknowledge a number of sources. The section on “Defense of Data” has benefited greatly from the American Statistical Association’s Committee on Privacy and Confidentiality and its comparison of the HIPAA Privacy Rule and the Common Rule [13]. The section on roles owes a great deal to the paper by Sanchez Pinto et al. [14] and in particular to the three CROs who spoke at the workshop from which the paper was developed, Bill Barnett, Peter Embi, and Umberto Tachinardi. Also fellow panelists at AMIA Summit 2018, Harold Lehmann, Kate Fultz Hollis, Bill Hersh, Jihad Obeid, Megan Singleton, and Umberto Tachinardi. The work of John Holmes [15–17] was also influential. The implementation section benefited from Adam Tobias and colleagues’ work at USF [11]. Of course, none of these authors bears any responsibility for errors or misunderstandings that may have crept into this chapter.

References

1. Donabedian A. Evaluating the quality of medical care. *Milbank Q.* 2005;83(4):691–729. Reprinted from The Milbank Memorial Fund Quarterly 44:3.2;166-203 (1966)
2. AHIMA. Information Governance Principles for Healthcare (IGPHC). Available at: www.ahima.org/~media/AHIMA/Files/HIM-Trends/IG_Principles.ashx.
3. Martin PY, Turner BA. Grounded theory and organizational research. *J Appl Behav Sci.* 1986;22(2):141.
4. Fahey L, Prusak L. The eleven deadliest sins of knowledge management. *Calif Manag Rev.* 1998;40(3):265–76. (“Error 3”). This precise formulation was given—repeated twice for emphasis—at a HICSS2000 keynote.
5. Her QL, Malenfant JM, Malek S, Vilk Y, Young J, Li L, Brown J, Toh S. A query workflow design to perform automatable distributed regression analysis in large distributed data networks. *eGEMS.* 2018;6(1):1–11.
6. Health Insurance Portability and Accountability Act of 1996. Public Law 104–191. US Government Publishing Office. 1996. Available at: <https://www.gpo.gov/fdsys/pkg/PLAW-104publ191/pdf/PLAW-104publ191.pdf>
7. Whitman JQ. The two western cultures of privacy: dignity versus liberty. *Yale Law J.* 2004;113:1151–221. Available as Faculty Scholarship Series, Paper 649 at http://digitalcommons.law.yale.edu/fss_papers/649
8. Warren SD, Brandeis LD. The right to privacy. *Harv Law Rev.* 1890;4(5):193–220.
9. Viktor Mayer-Schönberger. Beyond privacy beyond rights – toward a systems theory of information governance. *Calif Law Rev.* 98:1853–1885 (2010). Available at <http://scholarship.law.berkeley.edu/californialawreview/vol98/iss6/4>.
10. Laudon KC. Markets and privacy. *Commun ACM.* 39, 9:92–104
11. Tobias A, Chackravarthy S, Fernandes S, Strobbe J AAMC Conference on Information Technology in Academic Medicine, Toronto, June 2016; also presented as an AMIA CRI-WG Webinar, October 2016.
12. <https://www.gartner.com/it-glossary/information-governance>.
13. American Statistical Association. Committee on privacy and confidentiality. Comparison of HIPAA Privacy Rule and The Common Rule for the Protection of Human Subjects in Research. 2011.

14. Sanchez-Pinto LN, Mosa ASM, Fultz-Hollis K, Tachinardi U, Barnett WK, Embi PJ. The emerging role of the chief research informatics officer in academic health centers. *Appl Clin Informat.* 2017;8(3):845–53.
15. Brown JS, Holmes JH, Shah K, et al. Distributed health data networks: a practical and preferred approach to multi-institutional evaluations of comparative effectiveness, safety, and quality of care. *Med Care.* 2010;48(6, Supplement 1: Comparative Effectiveness Research: Emerging Methods and Policy Applications):S45–51.
16. Holmes JH, Elliott TE, Brown JS, et al. Clinical research data warehouse governance for distributed research networks in the USA: a systematic review of the literature. *JAMIA.* 2014;21:730–6.
17. Maro JC, Platt R, Holmes JH, et al. Design of a national distributed health data network. *Ann Intern Med.* 2009;151:341–4.

Part III

Knowledge Representation and Discovery: New Challenges and Emerging Models



Knowledge Representation and Ontologies

15

Kin Wah Fung and Olivier Bodenreider

Abstract

The representation of medical data and knowledge is fundamental in the field of medical informatics. Ontologies and related artifacts are important tools in knowledge representation, yet they are often given little attention and taken for granted. In this chapter, we give an overview of the development of medical ontologies, including available ontology repositories and tools. We highlight some ontologies that are particularly relevant to clinical research and describe with examples the benefits of using ontologies to facilitate research workflow management, data integration, and electronic phenotyping.

Keywords

Knowledge representation · Biomedical ontologies · Research metadata ontology
Data content ontology · Ontology-driven knowledge bases · Data integration
Electronic phenotyping

Ontologies have become important tools in biomedicine, supporting critical aspects of both healthcare and biomedical research, including clinical research [1]. Some even see ontologies as integral to science [2]. Unlike terminologies (focusing on naming) and classification systems (developed for partitioning a domain), ontologies define the types of entities that exist, as well as their interrelations. And while knowledge bases generally integrate both definitional and assertional knowledge,

K. W. Fung, MD, MS, MA (✉) · O. Bodenreider, MD, PhD
Lister Hill National Center for Biomedical Communications, National Library of Medicine,
National Institutes of Health, Bethesda, MD, USA
e-mail: kfung@mail.nih.gov; obodenreider@mail.nih.gov

ontologies focus on what is always true of entities, i.e., definitional knowledge [3]. In practice, however, there is no sharp distinction between these kinds of artifacts, and “ontology” has become a generic name for a variety of knowledge sources with important differences in their degree of formality, coverage, richness, and computability [4].

Ontology Development

Ontology development has not yet been formalized to the same extent as, say, database development has, and there is still no equivalent for ontologies to the entity-relationship model. However, ontology development is guided by fundamental ontological distinctions and supported by the formalisms and tools for knowledge representation that have emerged over the past decades. Several top-level ontologies provide useful constraints for the development of domain ontologies, and one of the most recent trends is increased collaboration among the creators of ontologies for coordinated development.

Important Ontological Distinctions

A small number of ontological distinctions inherited from philosophical ontology provide a useful framework for creating ontologies. The first distinction is between types and instances. Instances correspond to individual entities (e.g., my left kidney, the patient identified by 1234), while types represent the common characteristics of sets of instances (e.g., a *kidney* is a bean-shaped, intra-abdominal organ – properties common to all kidneys) [5]. Instances are related to the corresponding types by the relation *instance of*. For example, my left kidney is an *instance of kidney*. (It must be noted that most biomedical ontologies only represent types in reference to which the corresponding instances are recorded in patient records and in laboratory notebooks.) Another fundamental distinction is between continuants and occurrents [6]. While continuants exist (endure) through time, occurrents go through time in phases. Roughly speaking, objects (e.g., a liver, an endoscope) are continuants and processes (e.g., the flow of blood through the mitral valve) are continuants. One final distinction is made between independent and dependent continuants. While the kidney and its shape are both continuants, the shape of the kidney “owes” its existence to the kidney (i.e., there cannot be a kidney shape unless there is a kidney in the first place). Therefore, the kidney is an independent continuant (as most objects are), whereas its shape is a dependent continuant (as are qualities, functions, and dispositions, all dependent on their bearers). These distinctions are important for ontology developers, because they help organize entities in the ontology and contribute to consistent ontology development, both within and, more importantly for interoperability, across ontologies.

Building Blocks: Top-Level Ontologies and Relation Ontology

These ontological distinctions are so fundamental that they are embodied by top-level ontologies such as BFO [7] (Basic Formal Ontology) and DOLCE [8] (Descriptive Ontology for Linguistic and Cognitive Engineering). Such upper-level ontologies are often used as building blocks for the development of domain ontologies. Instead of organizing the main categories of entities of a given domain under some artificial root, these categories can be implemented as specializations of types from the upper-level ontology. For example, a protein is an independent continuant, the catalytic function of enzymes is a dependent continuant, and the activation of an enzyme through phosphorylation is an occurrent. Of note, even when they do not leverage an upper-level ontology, most ontologies implement these fundamental distinctions in some way. For example, the first distinction made among the semantic types in the UMLS Semantic Network [9] is between *entity* and *event*, roughly equivalent to the distinction between continuants and occurrents in BFO. While BFO and DOLCE are generic upper-level ontologies, Bio-Top – itself informed by BFO and DOLCE – is specific to the biomedical domain and provides types directly relevant to this domain, such as *chain of nucleotide monomers* and *organ system*. BFO forms the backbone of several ontologies from the Open Biomedical Ontologies (OBO) family, and Bio-Top has also been reused by several ontologies. Some also consider the UMLS Semantic Network, created for categorizing concepts from the UMLS Metathesaurus, an upper-level ontology for the biomedical domain [9].

In addition to the ontological template provided for types by upper-level ontologies, standard relations constitute an important building block for ontology development and help ensure consistency across ontologies. The small set of relations defined collaboratively in the relation ontology [5], including *instance of*, *part of*, and *located in*, has been widely reused.

Formalisms and Tools for Knowledge Representation

Many ontologies use description logics for their representation. Description logics (DLs) are a family of knowledge representation languages, with different levels of expressiveness [10]. The main advantage of using DL for ontology development is that DL allows developers to test the logical consistency of their ontology. This is particularly important for large biomedical ontologies. Ontologies including OCRe, OBI, SNOMED CT, and the NCI Thesaurus, discussed later in this chapter, all rely on some sort of DL for their development.

Ontologies are key enabling resources for the Semantic Web, the “web of data,” where resources annotated in reference to ontologies can be processed and linked automatically [11]. It is therefore not surprising that the main language for representing ontologies, OWL – the Web Ontology Language, has its origins in the Semantic Web. OWL is developed under the auspices of the World Wide Web Consortium (W3C). The current version of the OWL specification is OWL 2, which

offers several profiles (sublanguages) corresponding to different levels of expressivity and support of DL languages [12]. Other Semantic Web technologies, such as RDF/S (Resource Description Framework Schema) [13] and SKOS (Simple Knowledge Organization System) [14], have also been used for representing taxonomies and thesauri, respectively.

The OWL syntax can be overwhelming to biologists and clinicians, who simply want to create an explicit specification of the knowledge in their domain. The developers of the Gene Ontology created a simple syntax later adopted for the development of many ontologies from the Open Biomedical Ontologies (OBO) family. The so-called OBO syntax [15, 16] provides an alternative to OWL, to which it can be converted [17].

The most popular ontology editor is Protégé, developed at the Stanford Center for Biomedical Informatics Research for two decades [18, 19]. Originally created for editing frame-based ontologies, Protégé now supports OWL and other Semantic Web languages. Dozens of user-contributed plugins extend the stand-alone version (e.g., for visualization, reasoning services, support for specific data formats), and the recently developed web version of Protégé supports the collaborative development of ontologies. Originally created to support the development of the Gene Ontology, OBO-Edit now serves as a general ontology editor [20, 21]. Simpler than Protégé, OBO-Edit has been used to develop many of the ontologies from the Open Biomedical Ontologies (OBO) family. Rather than OWL, OBO-Edit uses a specific format, the OBO syntax, for representing ontologies. Both Protégé and OBO-Edit are open-source, platform-independent software tools.

OBO Foundry and Other Harmonization Efforts

Two major issues with biomedical ontologies are proliferation and lack of interoperability. There are several hundreds of ontologies available in the domain of life sciences, some of which overlap partially but do not systematically cross-reference equivalent entities in other ontologies. The existence of multiple representations for the same entity makes it difficult for ontology users to select the right ontology for a given purpose and requires the development of mappings between ontologies to ensure interoperability. Two recent initiatives have offered different solutions to address the issue of uncoordinated development of ontologies.

The OBO Foundry is an initiative of the Open Biomedical Ontologies (OBO) consortium, which provides guidelines and serves as coordinating authority for the prospective development of ontologies [22]. Starting with the Gene Ontology, the OBO Foundry has identified kinds of entities for which ontologies are needed and have selected candidate ontologies to cover a given subdomain, based on a number of criteria. Granularity and fundamental ontological distinctions form the basis for identifying subdomains. For example, independent continuants (entities) at the molecular level include proteins (covered by the Protein Ontology), while macroscopic anatomical structures are covered by the Foundational Model of Anatomy. In addition to syntax, versioning, and documentation requirements, the OBO Foundry

guidelines prescribe that OBO Foundry ontologies be limited in scope to a given subdomain and orthogonal. This means, for example, that an ontology of diseases referring to anatomical structures as the location of diseases (e.g., *mitral valve regurgitation has location mitral valve*) should cross-reference entities from the reference ontology for this domain (e.g., the Foundational Model of Anatomy for *mitral valve*) rather than redefine these entities. While well adapted to coordinating the prospective development of ontologies, this approach is extremely prescriptive and virtually excludes the many legacy ontologies used in the clinical domain, including SNOMED CT and the NCI Thesaurus.

The need for harmonization, i.e., making existing ontologies interoperable and avoiding duplication of development effort, has not escaped the developers of large clinical ontologies. SNOMED International, in charge of the development of SNOMED CT, is leading a similar harmonization effort in order to increase interoperability and coordinate the evolution of legacy ontologies and terminologies, including Logical Observation Identifiers Names and Codes (LOINC, for laboratory and clinical observations), the International Classification of Diseases (ICD), Orphanet (for rare diseases), the Global Medical Device Nomenclature Agency (for medical devices), and the International Classification for Nursing Practice (ICNP, for nursing diagnoses) [23].

Ontologies of Particular Relevance to Clinical Research

Broadly speaking, clinical research ontologies can be classified into those that model the characteristics (or metadata) of the clinical research and those that model the data contents generated as a result of the research [24]. Research metadata ontologies center around characteristics like study design, operational protocol, and methods of data analysis. They define the terminology and semantics necessary for formal representation of the research activity and aim to facilitate activities such as automated management of clinical trials and cross-study queries based on study design, intervention, or outcome characteristics. Ontologies of data content focus on explicitly representing the information model of and data elements (e.g., clinical observations, laboratory test results) collected by the research, with the aim to achieve data standardization and semantic data interoperability. Important examples of the two types of ontology will be described in more detail.

Research Metadata Ontology

We did a survey of the public repositories of ontologies in the Open Biomedical Ontologies (OBO) library hosted by the National Center of Biomedical Ontology and the [FAIRsharing.org](#) website hosted by the University of Oxford [25, 26]. We used the keywords “clinical trial,” “research,” and “investigation” for searching. We learned about the identified research metadata ontologies through their online information and literature search. Ontologies with little or no available information and

evidence of ongoing use are not included here. We found three ontologies that are actively maintained and used: the Ontology of Clinical Research (OCRe), Ontology for Biomedical Investigations (OBI), and Biomedical Research Integrated Domain Group (BRIDG) model ontology.

Ontology of Clinical Research

The primary aim of OCRe is to support the annotation and indexing of human studies to enable cross-study comparison and synthesis [27, 28]. Developed as part of the Trial Bank Project, OCRe provides terms and relationships for characterizing the essential design and analysis elements of clinical studies. Domain-specific concepts are covered by reference to external vocabularies. Workflow-related characteristics (e.g., schedule of activities) and data structure specification (e.g., schema of data elements) are not within the scope of OCRe.

The three core modules of OCRe are:

1. Clinical module – the upper-level entities (e.g., clinician, study subject)
2. Study design module – models study design characteristics (e.g., investigator assigned intervention, external control group)
3. Research module – terms and relationships to characterize a study (e.g., outcome phenomenon, assessment method)

OCRe entities are mapped to the Basic Formal Ontology (BFO).

Ontology for Biomedical Investigations

Unlike OCRe which is rooted in clinical research, the origin of OBI is in the molecular biology research domain [29, 30]. The forerunner of OBI is the MGED Ontology developed by the Microarray Gene Expression Data Society for annotating microarray data. Through collaboration with other groups in the “OMICS” arena such as the Proteomics Standards Initiative (PSI) and Metabolomics Standards Initiative (MSI), MGED Ontology was expanded to cover proteomics and metabolomics and was subsequently renamed Functional Genomics Investigation Ontology (FuGO) [31]. The scope of FuGO was later extended to cover clinical and epidemiological research and biomedical imaging, resulting in the creation of OBI, which aims to cover all biomedical investigations [32].

As OBI is an international, cross-domain initiative, the OBI Consortium draws upon a pool of experts from many fields, including even fields outside biology such as environmental science and robotics. The goal of OBI is to build an integrated ontology to support the description and annotation of biological and clinical investigations, regardless of the particular field of study. OBI also uses the BFO as its upper-level ontology and all OBI classes are a subclass of some BFO class. OBI covers all phases of the experimental process and the entities or concepts involved,

such as study designs, protocols, instrumentation, biological material, collected data, and their analyses. OBI also represents roles and functions which can be used to characterize and relate these entities or concepts. Specifically, OBI covers the following areas:

1. Biological material – e.g., blood plasma
2. Instrument – e.g., microarray and centrifuge
3. Information content – e.g., electronic medical record and biomedical image
4. Design and execution of an investigation – e.g., study design and electrophoresis
5. Data transformation – e.g., principal components analysis and mean calculation

For domain-specific entities, OBI makes reference to other ontologies such as the Gene Ontology (GO) and Chemical Entities of Biological Interest (ChEBI). The ability of OBI to adequately represent and integrate different biological experimental processes and their components has been demonstrated in examples from several domains, including neuroscience and vaccination.

Biomedical Research Integrated Domain Group (BRIDG) Model Ontology

The Biomedical Research Integrated Domain Group (BRIDG) model is a collaborative effort engaging stakeholders from the Clinical Data Interchange Standards Consortium (CDISC, described in Chap. 20(?) Richesson/Standards), HL7 Regulated Clinical Research Information Management Technical Committee (RCRIM TC), National Cancer Institute (NCI), and US Food and Drug Administration (FDA) [33–35]. The goal of the BRIDG model is to produce a shared view of the dynamic and static semantics for the domain of protocol-driven research and its associated regulatory artifacts, defined as “the data, organization, resources, rules, and processes involved in the formal assessment of the utility, impact, or other pharmacological, physiological, or psychological effects of a drug, procedure, process, or device on a human, animal, or other subject or substance plus all associated regulatory artifacts required for or derived from this effort, including data specifically associated with post-marketing adverse event reporting.”

One important function of the BRIDG model is to facilitate integration and meaningful data exchange from biological, translational, and clinical studies with data from health systems by providing a common understanding of biomedical research concepts and their relationships with healthcare semantics.

The BRIDG model (version 5.0) is divided into nine subdomains:

1. Common – concepts and semantics shared across different types of protocol-driven research, e.g., people, organizations, places, and materials
2. Protocol representation – planning and design of a clinical research protocol, e.g., study objective, outcome measure, and inclusion criteria

3. Study conduct – concepts related to execution of a research protocol, e.g., study site investigator, funding source, and specimen collection
4. Adverse events – safety-related activities such as detection, evaluation, and follow-up reporting of adverse events
5. Statistical analysis – planning and performance of the statistical analysis of data collected during execution of the protocol
6. Experiment – design, planning, resourcing, and execution of biomedical experiments, e.g., devices and parameters, variables that can be manipulated
7. Biospecimen – collection and management of biospecimens
8. Molecular biology – including genomics, transcriptomics, proteomics, pathways, biomarkers, and other concepts
9. Imaging – covers imaging semantics such as image acquisition, processing, and reconstruction

The experiment, biospecimen, and molecular biology subdomains are introduced since version 4.0 in response to calls for BRIDG to support molecular-based medicine, in which treatment of disease is informed by the patient's genome and other molecular characteristics. The imaging subdomain is new for version 5.0 to facilitate interfacing between a clinical trial management system and imaging systems. To enhance interoperability with other ongoing data modeling efforts, the BRIDG model has been mapped to the Common Data Model (CDM) of the Observational Health Data Sciences and Informatics (OHDSI) network. Ability to map to other clinical trial ontologies has also been demonstrated [36]. One of the future priorities of BRIDG is vocabulary binding. Historically, BRIDG is ontology and terminology agnostic, and no formal binding is provided between vocabularies and classes and attributes within BRIDG. Recognizing the value of improving semantic interoperability, future work will bind BRIDG class attributes to one or more common terminologies from medicine and research.

The BRIDG model is available as an OWL ontology. It is also available as a UML representation (intended for domain experts and architects) and as an HL7 reference information model (RIM) representation in Visio files.

Data Content Ontology

While there are relatively few research metadata ontologies, there is a myriad of ontologies that cover research data contents. Unlike metadata ontologies, in this group the distinction between ontologies, vocabularies, classifications, and code sets often gets blurred, and we shall refer to all of them as “terminologies.” As clinical research is increasingly conducted based on EHR data (e.g., pragmatic trials), the separation between terminologies for clinical research and healthcare is also becoming less important. We have chosen several terminologies for more detailed discussion here because of their role in clinical research and in electronic health records. These terminologies are the National Cancer Institute Thesaurus (NCIT), Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT), Logical

Observation Identifiers Names and Codes (LOINC), RxNorm, International Classifications of Diseases (ICD), and Current Procedural Terminology (CPT). All these terminologies are available through the Unified Medical Language System (UMLS) and the BioPortal ontology repositories (see below).

National Cancer Institute Thesaurus (NCIT)

NCIT is developed by the US National Cancer Institute (NCI). It arose initially from the need for an institution-wide common terminology to facilitate interoperability and data sharing by the various components of NCI [37–39]. NCIT covers clinical and basic sciences as well as administrative areas. Even though the content is primarily cancer-centric, since cancer research spans a broad area of biology and medicine, NCIT can potentially serve the needs of other research communities. Due to its coverage of both basic and clinical research, NCIT is well positioned to support translational research. NCIT was the reference terminology for the NCI's cancer Biomedical Informatics Grid (caBIG) and other related projects. It was one of the US Federal standard terminologies designated by the Consolidated Health Informatics (CHI) initiative, and it hosts many CDISC concepts and value sets.

NCIT contains about 120,000 concepts organized into 19 disjoint domains. A concept is allowed to have multiple parents within a domain. NCIT covers the following areas:

1. Neoplastic and other diseases
2. Findings and abnormalities
3. Anatomy, tissues, and subcellular structures
4. Agents, drugs, and chemicals
5. Genes, gene products, and biological processes
6. Animal models of disease
7. Research techniques, equipment, and administration

NCIT is updated monthly. It is in the public domain under an open content license and is distributed by the NCI in OWL format.

SNOMED Clinical Terms (SNOMED CT)

SNOMED CT was originally developed by the College of American Pathologists. Its ownership was transferred to SNOMED International (originally called International Health Terminology Standards Development Organisation, IHTSDO) in 2007 to enhance international governance and adoption [40]. SNOMED CT has been steadily gaining momentum as the emerging international standard clinical terminology. The number of member countries of SNOMED International has more than tripled since its inception. There are currently 33 member countries including the USA, the UK, Canada, Australia, India, Malaysia, the Netherlands, Sweden, and

Spain. SNOMED CT is used in over 50 countries in the world [41]. SNOMED CT is the most comprehensive clinical terminology available today, with over 340,000 active concepts. The concepts are organized into 19 disjoint hierarchies. Within each hierarchy, a concept is allowed to have multiple parents. Additionally, SNOMED CT provides a rich set of associated relations (across hierarchies), which form the basis for the logical definitions of its concepts. The principal use of SNOMED CT is to encode clinical information (e.g., diseases, findings, procedures). It also has comprehensive coverage of drugs, organisms, and anatomy. SNOMED CT is a designated terminology for the problem list, procedures, and other data fields according to the Meaningful Use of EHR incentive program of the US Centers for Medicare & Medicaid Services (CMS) [42, 43]. After the Meaningful Use program ended in 2017, the requirements for SNOMED CT use persist in the subsequent Merit-based Incentive Payment System (MIPS) and Promoting Interoperability programs.

SNOMED CT is updated twice every year. The use of SNOMED CT is free in all SNOMED International member countries, in low-income countries as defined by the World Bank, and for qualified research projects in any country. SNOMED CT is distributed by the National Release Center of the SNOMED International member countries.

Logical Observation Identifiers, Names, and Codes (LOINC)

LOINC is developed by the Regenstrief Institute, a nonprofit biomedical informatics and healthcare research organization associated with Indiana University [44]. LOINC's primary role is to provide identifiers and names for laboratory and clinical observations that will facilitate the unambiguous exchange and aggregation of clinical results for many purposes, including care delivery, quality assessment, public health, and research purposes [45]. The laboratory section of LOINC covers the usual categories in clinical laboratory testing such as chemistry, urinalysis, hematology, microbiology, molecular genetics, and others. This section accounts for about two-thirds of LOINC codes. The clinical section covers a very broad scope, from clinical documents, anthropomorphic measures to cardiac and obstetrical ultrasound. Each LOINC code corresponds to a single kind of observation, measurement, or test result. A LOINC term includes six parts: component, kind of property, time aspect, system, type of scale, and type of method (optional). LOINC has over 80,000 terms. In 2013, the Regenstrief Institute and SNOMED International formed a long-term collaborative relationship with the objective of developing coded content to support order entry and result reporting by linking SNOMED CT and LOINC. This landmark agreement aims to reduce duplication of effort and provide a common framework within which to use the two terminologies.

In the USA, LOINC has been adopted by large reference laboratories, health information exchanges, healthcare organizations, and insurance companies. LOINC is also a designated terminology for the EHR under the Meaningful Use program.

Internationally, LOINC has over 60,000 registered users from 172 countries. At least 15 countries have chosen LOINC as a national standard. LOINC is updated twice a year. Use of LOINC is free upon agreeing to the terms of use in the license.

RxNorm

RxNorm is a standard nomenclature for medications developed by NLM [46]. RxNorm provides normalized names for clinical drugs and links its names to many of the drug vocabularies commonly used in pharmacy management and drug interaction software, including those of First Databank, Micromedex, Gold Standard Drug Database, and Multum. RxNorm also integrates drugs from sources like DrugBank and the Anatomical Therapeutic Chemical (ATC) drug classification system, often used in research projects. By providing links between these vocabularies, RxNorm can mediate messages between systems not using the same software and vocabulary. The focus of RxNorm is at the clinical drug level, represented as a combination of ingredients, strength, and dose form. The clinical drug is linked by semantic relationships to other drug entities such as ingredients and drug packs. Non-therapeutic radiopharmaceuticals, bulk powders, contrast media, food, dietary supplements, and medical devices (e.g., bandages and crutches) are all out of scope for RxNorm. RxNorm has about 37,000 generic clinical drugs, 22,000 branded clinical drugs, and 11,000 ingredients. The Current Prescribable Content Subset is a subset of currently prescribable drugs in RxNorm. The subset is intended to be an approximation of the prescription drugs currently marketed in the USA, and it also includes some frequently prescribed over-the-counter drugs.

RxNorm is the designated terminology for medications and medication allergies according to the Meaningful Use incentive program. The Centers for Medicare & Medicaid Services (CMS) uses RxNorm in its Formulary Reference File to define the value sets for clinical quality measures. The National Council for Prescription Drug Programs (NCPDP) uses RxNorm in its SCRIPT e-prescribing and Formulary and Benefit Standards. The Department of Veterans Affairs (VA) and the Department of Defense (DoD) use RxNorm to enable bi-directional real-time data exchange for medication and drug allergy information [47].

RxNorm is released as a full data set every month. There are weekly updates for newly approved drugs. To download the RxNorm files, a UMLS user license is required because some RxNorm content comes from commercial drug knowledge sources and is proprietary.

International Classification of Disease (ICD)

The root of ICD can be traced back to the International List of Causes of Death created 150 years ago [48]. ICD is endorsed by the World Health Organization (WHO) to be the international standard diagnostic classification for epidemiology, health management, and clinical purposes. The current version of ICD is

ICD-10 which was first published in 1992. ICD-11 is still under development. Apart from reporting national mortality and morbidity statistics to WHO, many countries use ICD-10 for reimbursement and healthcare resource allocation. To better suit their national needs, several countries have created national extensions to ICD-10, including ICD-10-AM (Australia), ICD-10-CA (Canada), and ICD-10-CM (USA). In the USA, ICD-9-CM was used until 2015 and was replaced by ICD-10-CM. Because of the requirement of ICD codes for reimbursement, they are ubiquitous in the EHR and insurance claims data. There is a fourfold increase in the number of codes from ICD-9-CM to ICD-10-CM, due to the more granular disease codes and capture of additional healthcare dimensions (e.g., episode of encounter, stage of pregnancy) [49]. CMS provides forward and backward maps between ICD-9-CM and ICD-10-CM, which are called General Equivalence Maps (GEMs). These maps are useful for conversion of coded data between the two versions of ICD [50].

While ICD-9-CM covers both diagnosis and procedures, ICD-10-CM does not cover procedures. A brand-new procedure coding system called ICD-10-PCS was developed by CMS to replace the ICD-9-CM procedure codes for reporting of inpatient procedures [51]. ICD-10-PCS is a radical departure from ICD-9-CM and uses a multiaxial structure. Each ICD-10-PCS code has seven digits, each covering one aspect of a procedure such as body part, root operation, approach, and device. As a result of the transition, there is a big jump in the number of procedure codes from about 4000 to over 70,000.

Both ICD-10-CM and ICD-10-PCS are updated annually and are free for use without charge.

Current Procedural Terminology (CPT)

CPT is developed by the American Medical Association (AMA) to encode medical services and procedures. In the USA, CPT is used to report physician services, many non-physician services, and surgical procedures performed in hospital outpatient departments and ambulatory surgery centers. The scope of CPT includes physician consultation and procedures, physical and occupational services, radiological and clinical laboratory investigations, transportation services, and others. There are three categories of CPT codes. Category I codes are five-digit numeric codes. For a procedure to receive a category 1 code, it must be an established and approved procedure with proven clinical efficacy performed by many healthcare professionals. Category II codes are five-character alphanumeric codes ending with an “F.” These are supplementary tracking codes for quality and performance measurement. Category III codes are temporary five-character alphanumeric codes ending with “T.” These codes are for emerging technologies that do not yet qualify for regular category I codes. There are about 9000 category I codes. CPT is now in the fourth edition and is updated annually. Use of CPT requires a license from AMA.

Clinical Data Warehouses for Translational Research

Several clinical data warehouses have been developed for translational research purposes. On the one hand, there are traditional data warehouses created through the Clinical and Translational Science Awards (CTSA) program and other translational research efforts. Such warehouses include BTRIS [52], based on its own ontology, the Research Entity Dictionary, and STRIDE [53], based on standard ontologies, such as SNOMED CT and RxNorm. On the other hand, several proof-of-concept projects have leveraged Semantic Web technologies for translational research purposes. In the footsteps of a demonstration project illustrating the benefits of integrating data in the domain of Alzheimer's disease [54], other researchers have developed knowledge bases for cancer data (leveraging the NCI Thesaurus) [55] and in the domain of nicotine dependence (using an ontology developed specifically for the purpose of integrating publicly available datasets) [56]. The Translational Medicine Knowledge Base, based on the Translational Ontology, is a more recent initiative developed for answering questions relating to clinical practice and pharmaceutical drug discovery [57].

Ontology Repositories

Because most biomedical terminologies and ontologies are developed by different groups and institutions independently of each other and made available to users in heterogeneous formats, interoperability among them is generally limited. In order to create some level of semantic interoperability among ontologies and facilitate their use, several repositories have been created. Such repositories provide access to integrated ontologies through powerful graphical and programming interfaces. This section presents the two largest repositories: the Unified Medical Language System (UMLS) and the BioPortal.

Unified Medical Language System (UMLS)

The US National Library of Medicine (NLM) started the UMLS project in 1986. One of the main goals of UMLS is to aid the development of systems that help health professionals and researchers retrieve and integrate electronic biomedical information from a multitude of disparate sources [58–61]. One major obstacle to cross-source information retrieval is that the same information is often expressed differently in different vocabularies used by the various systems and there is no universal biomedical vocabulary. Knowing that to dictate the use of a single vocabulary is not realistic, the UMLS circumvents this problem by creating links between the terms in different vocabularies. The UMLS is available free of charge. Users need to acquire a license because some of the UMLS contents are protected by additional license requirements [62]. Currently, there are over 20,000 UMLS licensees in more than 120 countries. The UMLS is released twice a year.

UMLS Knowledge Sources

The Metathesaurus of the UMLS is a conglomeration of a large number of terms that exist in biomedical vocabularies. All terms that refer to the same meaning (i.e., synonymous terms) are grouped together in the same UMLS concept. Each UMLS concept is assigned a permanent unique identifier (the Concept Unique Identifier, CUI), which is the unchanging pointer to that particular concept. This concept-based organization enables cross-database information retrieval based on *meaning*, independent of the lexical variability of the terms themselves. In the 2018AA release, the UMLS Metathesaurus incorporates 154 source vocabularies and includes terms in 25 languages. There are two million biomedical concepts and eight million unique terms. The Metathesaurus also contains relationships between concepts. Most of these relationships are derived from relationships asserted by the source vocabularies. To edit the Metathesaurus, the UMLS editors use a sophisticated set of lexical- and rule-based matching algorithms to help them focus on areas that require manual review.

The Semantic Network is another resource in the UMLS. The Semantic Network contains 127 semantic types and 54 kinds of relationship between the semantic types. The Semantic Network is primarily used for the categorization of UMLS concepts [9]. All UMLS concepts are assigned at least one semantic type. The semantic relationships represent the possible relationships between semantic types, which may or may not hold true at the concept level. A third resource in the UMLS is the SPECIALIST Lexicon and the lexical tools. The SPECIALIST Lexicon is a general English lexicon that includes over 500,000 lexical items. Each lexicon entry records the syntactic, morphological, and orthographic information that can be used to support activities such as natural language processing of biomedical text. The lexical tools are designed to address the high degree of variability in natural language words and terms. Normalization is one of the functions of the lexical tools that help users to abstract away from variations involving word inflection, case, and word order [63].

UMLS Tooling

The UMLS is distributed as a set of relational tables that can be loaded in a database management system. Alternatively, a web-based interface and an application programming interface (API) are provided. The UMLS Terminology Services (UTS) is a web-based portal that can be used for downloading UMLS data; for browsing the UMLS Metathesaurus, Semantic Network, and SPECIALIST Lexicon; and for accessing the UMLS documentation. Users of the UTS can enter a biomedical term or the identifier of a biomedical concept in a given ontology, and the corresponding UMLS concept will be retrieved and displayed, showing the names for this concept in various ontologies, as well as the relations of this concept to other concepts. For example, a search on “Addison’s disease” retrieves all names for the corresponding concept (C0001403) in over 25 ontologies (version 2018AA, as of June 2018),

including SNOMED CT, MedDRA, and the International Classification of Primary Care. Each ontology can also be navigated as a tree. In addition to the graphical interface, the UTS also offers an application programming interface (API) based on RESTful web services. This API provides access to the properties and relations of Metathesaurus concepts, as well as semantic types and lexical entries. Most functions of the UTS API require UMLS user credentials to be checked in order to gain access to UMLS data. Support for user authentication is provided through the UTS API itself.

UMLS Applications

The UMLS provides convenient one-stop access to diverse biomedical vocabularies, which are updated as frequently as resources allow. One important contribution of the UMLS is that all source vocabularies are converted to a common schema of representation, with the same file structure and object model. This makes it much easier to build common tools that deal with multiple vocabularies, without the need to grapple with the native format of each. Moreover, this also enhances the understanding of the vocabularies as the common schema abstracts away from variations in naming conventions. For example, a term may be called “preferred name,” “display name,” or “common name” in different vocabularies, but if they are determined to mean the same type of term functionally, they are all referred to as “preferred term” in the UMLS.

One common use of the UMLS is inter-terminology mapping. The UMLS concept structure enables easy identification of equivalent terms between any two source terminologies. In addition to mapping by synonymy, methods have been reported that create inter-terminology mapping by utilizing relationships and lexical resources available in the UMLS [64]. Natural language processing is another important use of the UMLS making use of its large collection of terms, the SPECIALIST Lexicon and the lexical tools. MetaMap is a publicly available tool developed by NLM which aims to identify biomedical concepts in free text [65, 66]. This is often the first step in data mining and knowledge discovery. Other uses of the UMLS include terminology research, information indexing and retrieval, and terminology creation [67].

BioPortal

BioPortal is developed by the National Center for Biomedical Ontology (NCBO), one of the National Centers for Biomedical Computing, created in 2004. The goal of NCBO is “to support biomedical researchers in their knowledge-intensive work, by providing online tools and a Web portal enabling them to access, review, and integrate disparate ontological resources in all aspects of biomedical investigation and clinical practice.” BioPortal not only provides access to biomedical ontologies, but it also helps link ontologies to biomedical data [68].

BioPortal Ontologies

The current version of BioPortal integrates over 700 ontologies for biomedicine, biology, and life sciences and includes roughly 9 million concepts. A number of ontologies integrated in the UMLS are also present in BioPortal (e.g., Gene Ontology, LOINC, NCIT, and SNOMED CT). However, BioPortal also provides access to the ontologies from the Open Biomedical Ontologies (OBO) family, an effort to create ontologies across the biomedical domain. In addition to the Gene Ontology, OBO includes ontologies for chemical entities (e.g., ChEBI), biomedical investigations (OBI), phenotypic qualities (PATO), and anatomical ontologies for several model organisms, among many others. Some of these ontologies have received the “seal of approval” of the OBO Foundry (e.g., Gene Ontology, ChEBI, OBI, and Protein Ontology). Finally, the developers of biomedical ontologies can submit their resources directly to BioPortal, which makes BioPortal an open repository, as opposed to the UMLS. Examples of such resources include the Research Network and Patient Registry Inventory Ontology and the Ontology of Clinical Research. BioPortal supports several popular formats for ontologies, including OWL, OBO format, and the Rich Release Format (RRF) of the UMLS.

BioPortal Tooling

BioPortal is a web-based application allowing users to search, browse, navigate, visualize, and comment on the biomedical ontologies integrated in its repository. For example, a search on “Addison’s disease” retrieves the corresponding entries in 51 ontologies (as of June 2018), including SNOMED CT, the Human Phenotype Ontology, and DermLex. Visualization as tree or graph is offered for each ontology. The most original feature of BioPortal is to support the addition of marginal notes to various elements of an ontology, e.g., to propose new terms or suggest changes in relations. Such comments can be used as feedback by the developers of the ontologies and can contribute to the collaborative editing on ontologies. Users can also publish reviews of the ontologies. In addition to the graphical interface, BioPortal also offers an application programming interface (API) based on RESTful web services and is generally well integrated with Semantic Web technologies, as it provides URIs for each concept, which can be used as a reference in linked data applications.

BioPortal Applications

Similar to the UMLS, BioPortal identifies equivalent concepts across ontologies in its repositories (e.g., between the term *listeriosis* in DermLex and in Medline Plus Health Topics). The BioPortal Annotator is a high-throughput named entity recognition system available both as an application and a web service. The Annotator identifies the names of biomedical concepts in text using fast string-matching algorithms.

While users can annotate arbitrary text, BioPortal also contains 40 million records from 50 textual resources, which have been preprocessed with the Annotator, including several gene expression data repositories, ClinicalTrials.gov, and the Adverse Event Reporting System from the Food and Drug Administration (FDA). In practice, BioPortal provides an index to these resources, making it possible to use terms from its ontologies to search these resources. Finally, BioPortal also provides the Ontology Recommender, a tool that suggests the most relevant ontologies based on an excerpt from a biomedical text or a list of keywords.

Approaches to Ontology Alignment in Ontology Repositories

Apart from providing access to existing terminologies and ontologies, the UMLS and BioPortal also identify bridges between these artifacts, which will facilitate inter-ontology integration or alignment. For the UMLS, as each terminology is added or updated, every new term is comprehensively reviewed (by lexical matching followed by manual review) to see if they are synonymous with existing UMLS terms. If so, the incoming term is grouped under the same UMLS concept. In the BioPortal, equivalence between different ontologies is discovered by a different approach. For selected ontologies, possible synonymy is identified through algorithmic matching alone (without human review). It has been shown that simple lexical matching works reasonably well in mapping between some biomedical ontologies in BioPortal, compared to more advanced algorithms [69]. Users can also contribute equivalence maps between ontologies.

Ontology in Action: Uses of Ontologies in Clinical Research

Ontologies can be used to facilitate clinical research in multiple ways. In the following section, we shall highlight three areas for discussion: research workflow management, data integration, and electronic phenotyping. However, these are not meant to be watertight categories (e.g., the ontological modeling of the research design can facilitate workflow management, as well as data sharing and integration).

Research Workflow Management

In most clinical trials, knowledge about protocols, assays, and specimen flow is stored and shared in textual documents and spreadsheets. The descriptors used are neither encoded nor standardized. Stand-alone computer applications are often used to automate specific portions of the research activity (e.g., trial authoring tools, operational plan builders, study site management software). These applications are largely independent and rarely communicate with each other. Integration of these systems will result in more efficient workflow management, improve the quality of

the data collected, and simplify subsequent data analysis. However, the lack of common terminology and semantics to describe the characteristics of a clinical trial impedes efforts of integration. Ontology-based integration of clinical trial management applications is an attractive approach. One early example is the Immune Tolerance Network, a large distributed research consortium engaged in the discovery of new therapy for immune-related disorders. The Network created the Epoch Clinical Trial Ontologies and built an ontology-based architecture to allow sharing of information between disparate clinical trial software applications [70]. Based on the ontologies, a clinical trial authoring tool had also been developed [71].

Another notable effort in the use of ontology in the design and implementation of clinical trials is the Advancing Clinical Genomic Trials on Cancer (ACGT) Project in Europe. ACGT is a European Union co-funded project that aims at developing open-source, semantic, and grid-based technologies in support of post-genomic clinical trials in cancer research. One component of this project is the development of a tool called Ontology-based Trial Management Application (ObTiMA), which has two main components: the Trial Builder and the Patient Data Management System, which are based on their master ontology called ACGT Master Ontology (ACGT-MO) [72–75]. Trial Builder is used to create ontology-based case report forms (CRF), and the Patient Data Management System facilitates data collection by frontline clinicians.

The advantage of an ontology-based approach in data capture is that the alignment of research semantics and data definition is achieved early in the research process, which facilitates greatly the downstream integration of data collected from different data sources. The early use of a common master ontology obviates the need of a post hoc mapping between different data and information models, which is time-consuming and error-prone. Similar examples can be found in the use of OBI and BRIDG. OBI is used to define a standard submission form for the Eukaryotic Pathogen Database project, which integrates genomic and functional genomics data for over 30 protozoan parasites [76]. While the specific terms used for a specimen are mainly drawn from other ontologies (e.g., Gazetteer, PATO), OBI is used to provide categories for the terms used (e.g., sequence data) to facilitate the loading of the data onto a database and subsequent data mining. In the USA, FDA has used the BRIDG as the conceptual model for the Janus Clinical Trials Repository (CTR) warehouse. To support drug marketing application, clinical trial sponsors need to submit subject-level data from trials in the CDISC format to the FDA for storage in the Janus CTR, which is used to support regulatory review and cross-study analysis [77].

Data Integration

In the post-genomic era of research, the power and potential value of linking data from disparate sources is increasingly recognized. A rapidly developing branch of translational research exploits the automated discovery of association between clinical and genomics data [78]. Ontologies can play important roles at different strategic steps of data integration [79].

For many existing data sources, data sharing and integration only occur as an afterthought. To align multiple data sources to support activities such as cross-study querying or data mining is no trivial task. The classical approach, warehousing, is to align the sources at the *data* level (i.e., to annotate or index all available data by a common ontology). When the source data are encoded in different vocabularies or coding systems, which is sadly a common scenario, data integration requires alignment or mapping between the vocabularies. Resources like the UMLS and BioPortal are very useful in such mapping activity.

Another approach to data integration is to align data sources at the *metadata* level, which allows effective cross-database queries without actually pooling data in a common database or warehouse. The prerequisite to the effective query of a network of federated research data sources is a standard way to describe the characteristics of the individual sources. This is the role of a common research metadata ontology. OCRe (described above) is specifically created to annotate and align clinical trials according to their design and data analysis methodology. In a pilot study, OCRe is used to develop an end-to-end informatics infrastructure that enables data acquisition, logical curation, and federated querying of human studies to answer questions such as “find all placebo-controlled trials in which a macrolide is used as an intervention” [27]. Using similar approaches for data discovery and sharing, a brand-new platform called Vivli is created to promote the reuse of clinical research data [80]. Vivli is intended to act as a neutral broker between data contributor, data user, and the wider data sharing community. It will provide an independent data repository, in-depth search engine, and a cloud-based, secure analytics platform.

Another notable effort is BIRNLex which is created to annotate the Biomedical Informatics Research Network (BIRN) data sources [56]. The BIRN sources include image databases ranging from magnetic resonance imaging of human subjects, mouse models of human neurologic disease to electron microscopic imaging. BIRNLex not only covers terms in neuroanatomy, molecular species, and cognitive processes, but it also covers concepts such as experimental design, data types, and data provenance. BIRN employs a mediator architecture to link multiple databases. The mediator integrates the various source databases by the use of a common ontology. The user query is parsed by the mediator, which issues database-specific queries to the relevant data sources each with their specific local schema [81].

The use of OBI in the Investigation/StudY/Assay (ISA) Project is another example of ontology-based facilitation of data integration and sharing. The ISA Project supports managing and tracking biological experiment metadata to ensure its preservation, discoverability, and reuse [82]. Concepts from OBI are used to annotate the experimental design and other characteristics, so that queries such as “retrieve all studies with balanced design” or “retrieve all studies where study groups have at least 3 samples” are possible. In a similar vein, the BRIDG model ontology is used in various projects to facilitate data exchange. One example is the SALUS (Security and interoperability in next generation Public Protection and Disaster Relief (PPDR) communication infrastructures) Project of the European Union [83]. BRIDG is used to provide semantics for the project’s metadata repository to allow meaningful exchange of data between European electronic health records.

Other innovative approaches of using ontologies to achieve data integration have also been described. One study explored the possibility of tagging research data to support real-time meta-analysis [84]. Another described a prototype system for ontology-driven indexing of public data sets for translational research [85].

One particular form of data integration supported by ontologies is represented by what has become known as “linked data” in the Semantic Web community [86]. The foundational idea behind linked data and the Semantic Web is that resources semantically annotated to ontologies can be interrelated when they refer to the same entities. In practice, datasets are represented as graphs in RDF, the Resource Description Framework, in which nodes (representing entities) can be shared across graphs, enabling connections among graphs. Interestingly, a significant portion of the datasets currently interrelated as linked data consists of biomedical resources, including PubMed, KEGG, and DrugBank. For privacy reasons, very few clinical datasets have been made publicly available, and no such datasets are available as linked data yet. However, researchers have illustrated the benefits of Semantic Web technologies for translational research [54–57]. Moreover, the development of personal health records will enable individuals to share their clinical data, and effective de-identification techniques might also contribute to the availability of clinical data, which could enable knowledge discovery through the mining of large volume of data. Ontologies support linked data in three important ways. Ontologies provide a controlled vocabulary for entities in the Semantic Web; integrated ontology repositories, such as the UMLS and BioPortal, support the reconciliation of entities annotated to different ontologies; finally, relations in ontologies can be used for subsumption and other kinds of reasoning. An active community of researchers is exploring various aspects of biomedical linked data as part of the Semantic Web Health Care and Life Sciences Interest Group [87], with particular interest in the domain of drug discovery through the Linking Open Drug Data initiative [88].

Electronic (Computable) Phenotyping

Data in electronic health records (EHRs) are becoming increasingly available for clinical and translational research. Through projects such as the Electronic Medical Records and Genomics (eMERGE) Network [89], National Patient-Centered Clinical Research Network (PCORnet) [90], Strategic Health IT Advanced Research Projects (SHARP) [91], Observational Health Data Sciences and Informatics (OHDSI) [92], and NIH Health Care Systems Collaboratory [93], it has been demonstrated that EHR data can be used to develop research-grade disease phenotypes with sufficient accuracy to identify traits and diseases for biomedical research and clinical care.

Electronic or computable phenotyping refers to activities and applications that use data captured in the delivery of healthcare (typically from EHRs and insurance claims) to identify individuals or populations (cohorts) with clinical characteristics, events, or service patterns that are relevant to interventional, observational,

prospective, and/or retrospective studies [93]. So far, the most tried-and-true approach to electronic phenotyping utilizes explicit, standardized queries – consisting of logical operators, data fields, and list of codes often from standardized terminologies – that can be run against different data sources to identify comparable populations. Due to the heterogeneity across care settings, data models, and patient populations, designing phenotype definitions is complex and often requires customization for different data sources. However, the validity of selected phenotype definitions and the comparability of patient populations across different healthcare settings have been reported [94–98]. Newer approaches in electronic phenotyping involving techniques such as machine learning have been studied with promising results [99–102]. However, manually curated phenotype definitions are still the most commonly employed phenotyping method.

Most phenotype definitions to date use both structured and unstructured elements in the EHR. Structured elements usually include demographic information, billing codes, laboratory tests, vital signs, and medications. Unstructured elements include clinical notes, family history, radiology reports, pathology reports, and others. Utilization of unstructured data elements usually require additional processing by natural language processing. So far, the most commonly used structured data are the billing codes – especially the ICD and CPT codes because of their ubiquity in the EHR [103]. With the increasing use of clinical terminologies such as SNOMED CT, LOINC, and RxNorm as a result of the Meaningful Use and subsequent incentive programs, it is expected that the inclusion of these terminologies in phenotype definitions will increase. This should have a positive impact in the accuracy of phenotyping as clinical terminologies such as SNOMED CT have been shown to provide better coverage and more fine-grained representation of clinical information [104–106]. The use of standardized terminologies in the EHR will be a great boon toward making phenotype definitions fully computable and portable across data sources [107]. The use of robust terminologies can also make phenotype authoring more efficient. For example, the tools developed by the Informatics for Integrating Biology and the Bedside (i2b2) project leverage the intrinsic hierarchical structure of medical ontologies to allow the selection of all descendants under the same concept [108]. Before standardized terminologies become the norm, the diversity in content terminologies remains a challenge to electronic phenotyping. One approach to mitigate this problem is demonstrated by the OHDSI collaborative. The OHDSI vocabulary incorporates and maps terms from different terminologies to a core list of concepts.

Development of phenotype definitions is a time- and resource-intensive activity. Often knowledge engineers, domain experts, and researchers have to spend many hours to create and iteratively refine phenotype algorithms to achieve high sensitivity and specificity and positive and negative predictive values. It is highly likely that different research groups have the need to identify some common conditions such as type 2 diabetes mellitus. To ensure comparability of results and to avoid duplication of effort, it is important that phenotype definitions are validated and shared across institutional and organizational boundaries. One platform for the creation,

validation, and dissemination of phenotype definitions is the Phenotype Knowledgebase (PheKB) developed by the eMERGE Network [103]. PheKB has built-in tools specifically designed to enhance knowledge sharing and collaboration, so as to facilitate the transportability of phenotype definitions across different healthcare systems, clinical data repositories, and research applications.

Phenotype definitions often include enumerated lists of concepts that identify the pertinent characteristics of a patient population. These lists are conventionally called value sets, which are lists of codes from standard terminologies for diagnosis, procedures, laboratory tests, medications, etc. Value sets developed for phenotype definitions are very similar to value sets developed for quality measures. As part of the Electronic Clinical Quality Improvement (eCQI) initiative, healthcare systems have to submit data for selected clinical quality measures [109]. Quality measure value sets are used to identify subpopulations of patients sharing certain demographic and clinical characteristics, as defined by a clinical quality measure. The Value Set Authority Center (VSAC) of NLM is a purpose-built platform to support the authoring, maintenance, and dissemination of value sets which can be used for quality measurement, phenotype definition, and other purposes [110].

The Way Forward

Looking forward, it is encouraging that the value of ontologies in clinical research becomes more recognized. This is evidenced by the increase in the number of investigations making use of ontologies. At the same time, this is also accompanied by an increase in the number of ontologies, which in itself is a mixed blessing. Many researchers still tend to create their own ontologies to suit their specific use case. Reuse of existing ontologies is only a rarity. If left unchecked, this tendency has the potential of growing into the very problem that ontologies are created to solve – the multitude of ontologies will itself become the barrier to data interoperability and integration. Post hoc mapping and alignment of ontologies are often difficult (if not impossible) and an approximation at best (with inherent information loss). The solution is to coordinate the development and maximize the reuse of existing ontologies, which will significantly simplify things downstream.

To facilitate reuse of ontologies, resources like the UMLS and BioPortal are indispensable. They enable users to navigate the expanding sea of biomedical ontologies. In addition to listing and making these ontologies available, what is still lacking is a better characterization of these ontologies to help users decide whether they are suitable for the tasks at hand. In case there are multiple candidate ontologies, some indicators of quality (e.g., user base, ways in which they are used, user feedback and comments) will be very useful to help users decide on the best choice.

Acknowledgments This research was supported in part by the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine (NLM).

References

1. Bodenreider O. Biomedical ontologies in action: role in knowledge management, data integration and decision support. *Yearb Med Inform* 2008;17(01):67–79.
2. Smith B. Ontology (Science). *Nature Precedings*, 2008. Available from [Nature Precedings](http://hdl.handle.net/10101/npre.2008.2027.2).
<http://hdl.handle.net/10101/npre.2008.2027.2>.
3. Bodenreider O, Stevens R. Bio-ontologies: current trends and future directions. *Brief Bioinform.* 2006;7(3):256–74.
4. Cimino JJ, Zhu X. The practical impact of ontologies on biomedical informatics. *Yearb Med Inform* 2006;15(01):124–135.
5. Smith B, et al. Relations in biomedical ontologies. *Genome Biol.* 2005;6(5):R46.
6. Simmons P, Melia J. Continuants and occurrents. *Proc Aristot Soc Suppl Vol.* 2000;74:59–75. +77–92.
7. IFOMIS. BFO. Available from: <http://www.ifomis.org/bfo/>.
8. Laboratory for Applied Ontology. DOLCE. Available from: <http://www.loa-cnr.it/DOLCE.html>.
9. McCray AT. An upper-level ontology for the biomedical domain. *Comp Funct Genomics*. 2003;4(1):80–4.
10. Baader F, et al. The description logic handbook: theory, implementation, and applications. 2nd ed. xix, 601 p ed. 2007, Cambridge University Press: Cambridge, New York. ill. 26 cm.
11. Berners-Lee T, Hendler J, Lassila O. The semantic web: a new form of web content that is meaningful to computers will unleash a revolution of new possibilities. *Sci Am.* 2001;284(5):34–43.
12. World Wide Web Consortium. OWL 2 web ontology language document overview. 2009a. Available from: <http://www.w3.org/TR/owl2-overview/>.
13. World Wide Web Consortium. RDF vocabulary description language 1.0: RDF schema. 2004. Available from: <http://www.w3.org/TR/rdf-schema/>.
14. World Wide Web Consortium. SKOS simple knowledge organization system reference. 2009b. Available from: <http://www.w3.org/TR/2009/REC-skos-reference-20090818/>.
15. Day-Richter J. The OBO flat file format specification. 2006. Available from: http://www.geneontology.org/GO.format.obo-1_2.shtml.
16. Mungall C, et al.. OBO flat file format 1.4 syntax and semantics. Available from: <http://owlcollab.github.io/oboformat/doc/obo-syntax.html>.
17. Golbreich C, et al. OBO and OWL: leveraging semantic web technologies for the life sciences, in Proceedings of the 6th international The semantic web and 2nd Asian conference on Asian semantic web conference. Busan: Springer-Verlag; 2007. p. 169–82.
18. Noy N, et al. The ontology life cycle: integrated tools for editing, publishing, peer review, and evolution of ontologies. *AMIA Ann Symp Proc.* 2010;2010:552–6.
19. Stanford Center for Biomedical Informatics Research. Protégé. Available from: <http://protege.stanford.edu/>.
20. Day-Richter J, et al. OBO-edit-an ontology editor for biologists. *Bioinformatics.* 2007;23(16):2198–200.
21. Lawrence Berkeley National Lab. OBO-edit. Available from: <http://oboedit.org/>.
22. Smith B, et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol.* 2007;25(11):1251–5.
23. International S. Partnerships – working with other standards organizations. Available from: <https://www.snomed.org/about/partnerships>.
24. Richesson RL, Krischer J. Data standards in clinical research: gaps, overlaps, challenges and future directions. *J Am Med Inform Assoc.* 2007;14(6):687–96.
25. FAIRsharing website. <https://www.FAIRsharing.org>.
26. McQuilton P, Gonzalez-Beltran A, Rocca-Serra P, Thurston M, Lister A, Maguire E, Sansone SA. BioSharing: curated and crowd-sourced metadata standards, databases and data policies in the life sciences. *Database (Oxford)*. 2016.

27. Sim I, et al. Ontology-based federated data access to human studies information. *AMIA Ann Symp Proc.* 2012;2012:856–65.
28. Tu SW, et al. OCRe: ontology of clinical research. In 11th International Protege Conference. 2009.
29. Bandrowski A, et al. The ontology for biomedical investigations. *PLoS One.* 2016;11(4):e0154556.
30. Ontology for Biomedical Investigations: Community Standard for Scientific Data Integration. Available from: <http://obi-ontology.org/>.
31. Whetzel PL, et al. Development of FuGO: an ontology for functional genomics investigations. *OMICS.* 2006;10(2):199–204.
32. Brinkman RR, et al. Modeling biomedical experimental processes with OBI. *J Biomed Semant.* 2010;1(Suppl 1):S7.
33. Becnel LB, et al. BRIDG: a domain information model for translational and clinical protocol-driven research. *J Am Med Inform Assoc.* 2017;24(5):882–90.
34. Biomedical Research Integrated Domain Group Website. Available from: <https://bridgmodel.nci.nih.gov/faq/components-of-bridg-model>.
35. Fridsma DB, et al. The BRIDG project: a technical report. *J Am Med Inform Assoc.* 2008;15(2):130–7.
36. Tu SW, et al. Bridging epoch: mapping two clinical trial ontologies. In 10th International Protege Conference. 2007.
37. de Coronado S, et al. NCI thesaurus: using science-based terminology to integrate cancer research results. *Med Info.* 2004;11(Pt 1):33–7.
38. Fragoso G, et al. Overview and utilization of the NCI thesaurus. *Comp Funct Genomics.* 2004;5(8):648–54.
39. Sioutos N, et al. NCI Thesaurus: a semantic model integrating cancer-related clinical and molecular information. *J Biomed Inform.* 2007;40(1):30–43.
40. International S. SNOMED CT (Systematized Nomenclature of Medicine-Clinical Terms), SNOMED International. Available from: <https://www.snomed.org/>.
41. Lee D, et al. A survey of SNOMED CT implementations. *J Biomed Inform.* 2013;46(1):87–96.
42. Blumenthal D, Tavenner M. The “meaningful use” regulation for electronic health records. *N Engl J Med.* 2010;363(6):501–4.
43. Office of the National Coordinator for Health Information Technology (ONC) – Department of Health and Human Services. Standards & certification criteria Interim final rule: revisions to initial set of standards, implementation specifications, and certification criteria for electronic health record technology. *Fed Regist.* 2010;75(197):62686–90.
44. Huff SM, et al. Development of the Logical Observation Identifiers Names and Codes (LOINC) vocabulary. *J Am Med Inform Assoc.* 1998;5(3):276–92.
45. Logical Observation Identifier Names and Codes (LOINC). Available from: <https://loinc.org/>.
46. Nelson SJ, et al. Normalized names for clinical drugs: RxNorm at 6 years. *J Am Med Inform Assoc.* 2011;18(4):441–8.
47. Bouhaddou O, et al. Exchange of computable patient data between the Department of Veterans Affairs (VA) and the Department of Defense (DoD): terminology standards strategy. *J Am Med Inform Assoc.* 2008;15:174–183.
48. History of the development of the ICD, World Health Organization. Available from: <http://www.who.int/classifications/icd/en/HistoryOfICD.pdf>.
49. Steindel SJ. International classification of diseases, 10th edition, clinical modification and procedure coding system: descriptive overview of the next generation HIPAA code sets. *J Am Med Inform Assoc.* 2010;17(3):274–82.
50. Fung KW, et al. Preparing for the ICD-10-CM transition: automated methods for translating ICD codes in clinical phenotype definitions. *EGEMS (Wash DC).* 2016;4(1):1211.
51. Averill RF, et al. Development of the ICD-10 procedure coding system (ICD-10-PCS). *Top Health Inf Manag.* 2001;21(3):54–88.
52. Cimino JJ, Ayres EJ. The clinical research data repository of the US National Institutes of Health. *Stud Health Technol Inform.* 2010;160(Pt 2):1299–303.

53. Lowe HJ, et al. STRIDE – an integrated standards-based translational research informatics platform. *AMIA Ann Symp Proc.* 2009;2009:391–5.
54. Ruttenberg A, et al. Methodology – advancing translational research with the Semantic Web. *BMC Bioinforma.* 2007;8:S2.
55. McCusker JP, et al. Semantic web data warehousing for caGrid. *BMC Bioinforma.* 2009;10(Suppl 10):S2.
56. Sahoo SS, et al. An ontology-driven semantic mashup of gene and biological pathway information: application to the domain of nicotine dependence. *J Biomed Inform.* 2008;41(5):752–65.
57. Semantic Web for Health Care and Life Sciences Interest Group. Translational medicine ontology and knowledge base. Available from: <http://www.w3.org/wiki/HCLSIG/PharmaOntology>.
58. Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Res.* 2004;32(Database issue):D267–70.
59. Humphreys BL, Lindberg DA, Hole WT. Assessing and enhancing the value of the UMLS Knowledge Sources. *Proc Annu Symp Comput Appl Med Care.* 1991;78–82.
60. Humphreys BL, et al. The unified medical language system: an informatics research collaboration. *J Am Med Inform Assoc.* 1998;5(1):1–11.
61. Lindberg DA, Humphreys BL, McCray AT. The unified medical language system. *Methods Inf Med.* 1993;32(4):281–91.
62. UMLS. Unified Medical Language System (UMLS). Available from: <http://www.nlm.nih.gov/research/umls/>.
63. McCray AT, Srinivasan S, Browne AC. Lexical methods for managing variation in biomedical terminologies. *Proc Ann Symp Comput Appl Med Care.* 1994:235–9.
64. Fung KW, Bodenreider O. Utilizing the UMLS for semantic mapping between terminologies. *AMIA Annu Symp Proc.* 2005:266–70.
65. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc AMIA Symp.* 2001:17–21.
66. Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. *J Am Med Inform Assoc.* 2010;17(3):229–36.
67. Fung KW, Hole WT, Srinivasan S. Who is using the UMLS and how – insights from the UMLS user annual reports. *AMIA Annu Symp Proc.* 2006:274–8.
68. Noy NF, et al. BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Res.* 2009;37(Web Server issue):W170–3.
69. Ghazvinian A, Noy NF, Musen MA. Creating mappings for ontologies in biomedicine: simple methods work. *AMIA Annu Symp Proc.* 2009;2009:198–202.
70. Shankar RD, et al. An ontology-based architecture for integration of clinical trials management applications. *AMIA Annu Symp Proc.* 2007:661–5.
71. Shankar R, et al. TrialWiz: an ontology-driven tool for authoring clinical trial protocols. *AMIA Annu Symp Proc.* 2008:1226.
72. Brochhausen M, et al. The ACGT master ontology and its applications – towards an ontology-driven cancer research and management system. *J Biomed Inform.* 2011;44(1):8–25.
73. Martin L, Anguita A, Graf N, Tsiknakis M, Brochhausen M, Rüping S, Bucur A, Sfakianakis S, Sengstag T, Buffa F, Stenzhorn H. ACGT: advancing clinico-genomic trials on cancer - four years of experience. *Stud Health Technol Inform.* 2011;169:734–8.
74. Stenzhorn H, et al. The ObTiMA system – ontology-based managing of clinical trials. *Stud Health Technol Inform.* 2010;160(Pt 2):1090–4.
75. Weiler G, et al. Ontology based data management systems for post-genomic clinical trials within a European Grid Infrastructure for Cancer Research. *Conf Proc IEEE Eng Med Biol Soc.* 2007;2007:6435–8.
76. Eukaryotic Pathogen Database. Available from: <https://eupathdb.org/eupathdb/>.
77. FDA Janus Data Repository. Available from: <https://www.fda.gov/ForIndustry/DataStandards/StudyDataStandards/ucm155327.htm>.
78. Genome-Wide Association Studies. Available from: <http://grants.nih.gov/grants/gwas/>.

79. Bodenreider O. Ontologies and data integration in biomedicine: success stories and challenging issues. In: Bairoch A, Cohen-Boulakia S, Froidevaux C, editors. Proceedings of the Fifth International Workshop on Data Integration in the Life Sciences (DILS 2008). Berlin: Springer; 2008b. p. 1–4.
80. Vivli: Center for Global Clinical Research Data. Available from: <http://vivli.org/>.
81. Rubin DL, Shah NH, Noy NF. Biomedical ontologies: a functional perspective. *Brief Bioinform.* 2008;9(1):75–90.
82. Sansone SA, et al. Toward interoperable bioscience data. *Nat Genet.* 2012;44(2):121–6.
83. SALUS Project: Security and interoperability in next generation PPDR communication infrastructures. Available from: <https://www.sec-salus.eu/>.
84. Cook C, et al. Real-time updates of meta-analyses of HIV treatments supported by a biomedical ontology. *Account Res.* 2007;14(1):1–18.
85. Shah NH, et al. Ontology-driven indexing of public datasets for translational bioinformatics. *BMC Bioinforma.* 2009;10(Suppl 2):S1.
86. Bizer C, Heath T, Berners-Lee T. Linked data – the story so far. *Int J Semant Web Inf Syst.* 2009;5(3):1–22.
87. HCLS. Semantic Web Health Care and Life Sciences (HCLS) Interest Group.
88. Semantic Web for Health Care and Life Sciences Interest Group. Linking open drug data. Available from: <http://www.w3.org/wiki/HCLSIG/LODD>.
89. Gottesman O, et al. The electronic medical records and genomics (eMERGE) network: past, present, and future. *Genet Med.* 2013;15(10):761–71.
90. Fleurence RL, et al. Launching PCORnet, a national patient-centered clinical research network. *J Am Med Inform Assoc.* 2014;21(4):578–82.
91. Chute CG, et al. The SHARPn project on secondary use of electronic medical record data: progress, plans, and possibilities. *AMIA Ann Symp Proc.* 2011;2011:248–56.
92. Hripcsak G, et al. Observational health data sciences and informatics (OHDSI): opportunities for observational researchers. *Stud Health Technol Inform.* 2015;216:574–8.
93. Richesson RL, et al. Electronic health records based phenotyping in next-generation clinical trials: a perspective from the NIH Health Care Systems Collaboratory. *J Am Med Inform Assoc.* 2013;20(e2):e226–31.
94. Carroll RJ, et al. Portability of an algorithm to identify rheumatoid arthritis in electronic health records. *J Am Med Inform Assoc.* 2012;19(e1):e162–9.
95. Cutrona SL, et al. Validation of acute myocardial infarction in the Food and Drug Administration's mini-sentinel program. *Pharmacoepidemiol Drug Saf.* 2013;22(1):40–54.
96. Kho AN, et al. Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study. *J Am Med Inform Assoc.* 2012;19(2):212–8.
97. Newton KM, et al. Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network. *J Am Med Inform Assoc.* 2013;20(e1):e147–54.
98. Ritchie MD, et al. Robust replication of genotype-phenotype associations across multiple diseases in an electronic medical record. *Am J Hum Genet.* 2010;86(4):560–72.
99. Banda JM, et al. Electronic phenotyping with APHRODITE and the observational health sciences and informatics (OHDSI) data network. *AMIA Jt Summits Transl Sci Proc.* 2017;2017:48–57.
100. Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. *J Am Med Inform Assoc.* 2013;20(1):117–21.
101. Martin-Sanchez FJ, et al. Secondary use and analysis of big data collected for patient care. *Yearb Med Inform.* 2017;26(1):28–37.
102. Yu S, et al. Toward high-throughput phenotyping: unbiased automated feature extraction and selection from knowledge sources. *J Am Med Inform Assoc.* 2015;22(5):993–1000.
103. Kirby JC, et al. PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability. *J Am Med Inform Assoc.* 2016;23(6):1046–52.

104. Campbell JR, Payne TH. A comparison of four schemes for codification of problem lists. *Proc Ann Symp Comput Appl Med Care.* 1994;201–5.
105. Campbell JR, et al. Phase II evaluation of clinical coding schemes: completeness, taxonomy, mapping, definitions, and clarity. CPRI work group on codes and structures. *J Am Med Inform Assoc.* 1997;4(3):238–51.
106. Chute CG, et al. The content coverage of clinical classifications. For the computer-based patient record institute's work group on codes & structures. *J Am Med Inform Assoc.* 1996;3(3):224–33.
107. Mo H, et al. Desiderata for computable representations of electronic health records-driven phenotype algorithms. *J Am Med Inform Assoc.* 2015;22(6):1220–30.
108. Murphy SN, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *J Am Med Inform Assoc.* 2010;17(2):124–30.
109. Electronic Clinical Quality Improvement Resource Center, The Office of the National Coordinator for Health Information Technology. Available from: <https://ecqi.healthit.gov/content/about-ecqi>.
110. Value Set Authority Center, National Library of Medicine Available from: <https://vsac.nlm.nih.gov/>.



Nonhypothesis-Driven Research: Data Mining and Knowledge Discovery

16

Mollie R. Cummins

Abstract

Clinical information, stored over time, is a potentially rich source of data for clinical research. Knowledge discovery in databases (KDD), commonly known as data mining, is a process for pattern discovery and predictive modeling in large databases. KDD makes extensive use of data mining methods, automated processes, and algorithms that enable pattern recognition. Characteristically, data mining involves the use of machine learning methods developed in the domain of artificial intelligence. These methods have been applied to healthcare and biomedical data for a variety of purposes with good success and potential or realized clinical translation. Herein, the Fayyad model of knowledge discovery in databases is introduced. The steps of the process are described with select examples from clinical research informatics. These steps range from initial data selection to interpretation and evaluation. Commonly used data mining methods are surveyed: artificial neural networks, decision tree induction, support vector machines (kernel methods), association rule induction, and k-nearest neighbor. Methods for evaluating the models that result from the KDD process are closely linked to methods used in diagnostic medicine. These include the use of measures derived from a confusion matrix and receiver operating characteristic curve analysis. Data partitioning and model validation are critical aspects of evaluation. International efforts to develop and refine clinical data repositories are critically linked to the potential of these methods for developing new knowledge.

Keywords

Knowledge discovery in databases · Data mining · Artificial neural networks
Support vector machines · Decision trees · *k*-Nearest neighbor classification
Clinical data repositories

M. R. Cummins, PhD, RN (✉)

College of Nursing, University of Utah, Salt Lake City, UT, USA

e-mail: mollie.cummins@utah.edu

Clinical information, stored over time, is a potentially rich source of data for clinical research. Many of the concepts that would be measured in a prospective study are already collected in the course of routine healthcare. Based on comparisons of treatment effects, some believe well-designed case-control or cohort studies produce results equally rigorous to that of randomized controlled trials, with lower cost and with broader applicability [1]. While this potential has not yet been fully realized, the rich potential of clinical data repositories for building knowledge is undeniable. Minimally, analysis of routinely collected data can aid in hypothesis generation and refinement and partially replace expensive prospective data collection.

While smaller samples of data can be extracted for observational studies of clinical phenomena, there is also an opportunity to learn from the much larger, accumulated mass of data. The availability of so many instances of disease states, health behaviors, and other clinical phenomena bears an opportunity to find novel patterns and relationships. In an exploratory approach, the data itself can be used to fuel hypothesis development and subsequent research. Importantly, one can induce executable knowledge models directly from clinical data, predictive models that can be implemented in computerized decision support systems [2, 3]. However, the statistical approaches used in cohort and case-control studies of small samples are not appropriate for large-scale pattern discovery and predictive modeling, where bias can figure more prominently, data can fail to satisfy key assumptions, and *p* values can become misleading.

Knowledge discovery in databases (KDD), also commonly known as data mining, is the process for pattern discovery and predictive modeling in large databases. An iterative, exploratory process distinctly differs from traditional statistical analysis in that it involves a great deal of interaction and subjective decision-making by the analyst. KDD also makes extensive use of data mining methods, which are automated processes and algorithms that enable pattern recognition and are characteristically machine learning methods developed in the domain of artificial intelligence. These methods have been applied to healthcare and biomedical data for a variety of purposes with good success and potential or realized clinical translation.

The Knowledge Discovery in Databases Process

Casual use of the term *data mining* to describe everything from routine statistical analysis of small data sets to large-scale enterprise data mining projects is pervasive. This broad application of the term causes semantic difficulties when attempting to communicate about KDD-relevant concepts and tools. Though multiple models and definitions have been proposed, the terms and definitions used in this chapter will be those given by Fayyad and colleagues in their seminal overview of data mining and knowledge discovery. The Fayyad model encompasses other leading models. Fayyad and colleagues define data mining as the use of machine learning, statistical, and visualization techniques algorithms to enumerate patterns, usually in an automated fashion, over a set of data. They clarify that data mining is one step in a larger knowledge discovery in databases (KDD) process that includes

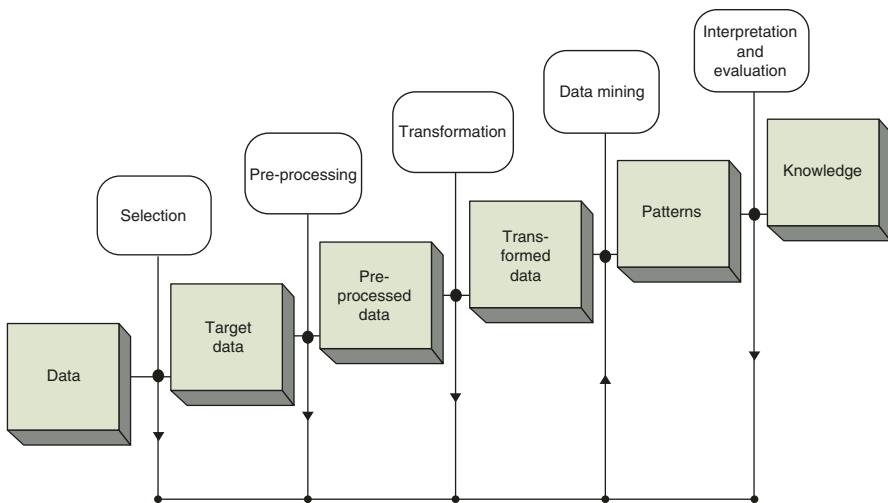


Fig. 16.1 Fayyad's knowledge discovery in databases process

data mining, along with any necessary data preparation, sampling, transformation, and evaluation/model refinement [4]. The encompassing process, the KDD process, is iterative and consists of multiple steps, depicted in Fig. 16.1. Data mining is not helpful or productive in inducing clinical knowledge models outside of this larger, essential process. Unless data mining methods are applied within a process that ensures validity, the results may prove invalid, misleading, and poorly integrated with current knowledge. As Fig. 16.1 depicts, the steps of KDD are iterative, not deterministic. While engaging in KDD, findings at any specific step may warrant a return to previous steps. The process is not sequential, as in a classic hypothetical-deductive scientific approach.

Data Selection

KDD projects are typically inceptioned when there is a clinical or operational decision requiring a clear and accurate knowledge model or in order to generate promising hypotheses for scientific study. These projects develop around a need to build knowledge or provide some guidance for clinical decision-making. Or lacking a particular clinical dilemma, a set of data particularly rich in content and size relevant to a particular clinical question may present itself. However, the relevant data is usually not readily available in a single flat file, ready for analysis. Typically, a data warehouse must be queried to return the subset of instances and attributes containing potentially relevant information. In some cases, clinical data will be partially warehoused, and some data will also need to be obtained from the source information system(s).

Just 20 years ago, data storage was sufficiently expensive, and methods for analysis of large data sets sufficiently immature, that clinical data was not routinely stored

apart from clinical information systems. However, there has been constant innovation and improvement in data storage and processing technology, approximating or exceeding that predicted by Moore's law. The current availability of inexpensive, high-capacity hard drives and inexpensive processing power is unprecedented. Data warehousing, the long-term storage of data from information systems, is now common. Transactional data, clinical data, radiological data, and laboratory data are now routinely stored in warehouses, structured to better facilitate secondary analysis and layered with analytic tools that enable queries and online analytic processing (OLAP).

Since clinical data is collected and structured to facilitate healthcare delivery and not necessarily analysis, key concepts may be unrepresented in the data or may be coarsely measured. For example, a coded field may indicate the presence or absence of pain, rather than a pain score. Proxies, other data attributes that correlate with unrepresented concepts, may be identified and included. For example, if a diagnosis of insulin-dependent diabetes is not coded, one might use insulin prescription (in combination with other attributes found in a set of data) as a proxy for Type I diabetes diagnosis. The use of proxy data and the triangulation of multiple data sources are often necessary to optimally represent concepts and identify specific populations within clinical data repositories [5]. A relevant subset of all available data is then extracted for further analysis.

Preprocessing

It is often said that preprocessing constitutes 90% of the effort in a knowledge discovery project. While the source and basis for that adage is unclear, it does seem accurate. Preprocessing is the KDD step that encompasses data cleaning and preparation. The values and distribution of values for each attribute must be closely examined, and with a large number of attributes, the process is time-consuming. It is sometimes appropriate or advantageous to recode values, adjust granularity, ignore infrequently encountered values, replace missing values, or to reduce data by representing data in different ways. For example, ordinality may be inherent in categorical values of an attribute and enable data reduction. An example exists in National Health Interview Survey data, wherein type of milk consumed is a categorical attribute. However, the different types of milk are characterized by different levels of fat content, and so the categorical values can be ordered by % fat content [6]. Each categorical attribute with n possible values constitutes n binary inputs for the knowledge discovery process. By restructuring a categorical attribute like type of milk consumed as an ordinal attribute, the values can be represented by a single attribute, and the number of inputs is reduced by $n - 1$. If attributes are duplicative or highly correlated, they are removed.

The distribution of values is also important because highly skewed distributions do not behave well mathematically with certain data mining methods. Attributes with highly skewed distributions can be adjusted to improve results, typically through normalization. The distribution of values is also important so that the investigator(s) is familiar with the representation of different concepts in the data set and can determine whether there are adequate instances for each attribute-value pair.

Transformation

Transformation is the process of altering the coded representation of data as input in order to reduce dimensionality or the number of rows and columns. Dimensionality reduction is often necessary in order to avoid combinatorial explosion or simply to improve computational efficiency during knowledge discovery. Combinatorial explosion is the vast increase in the number of possible patterns/solutions to a classification problem that occur with increases in the number of attributes. If a data set contains n input attributes, the number of possible combinations of attribute-value pairs that could be used to predict an outcome is 2^n . For a mere 16 inputs ($n = 16$), the number of possible combinations is 65,536. Every additional input results in increased computational demand. For knowledge discovery involving very large data sets, it is often necessary to create an alternate representation of the original input data, a representation that is computationally more manageable. Methods of transformation include wavelet transformation, principal components analysis, and automated binning (discretization) of interval attributes.

Data Mining

Data mining is the actual application of statistical and machine learning methods to enumerate patterns in a set of data [4]. It can be approached in several different ways, best characterized by the type of learning task specified. Artificial intelligence pioneer Marvin Minsky [7] defined learning as “making useful changes in our minds.” Data mining methods “learn” to predict values or class membership by making useful, incremental model adjustments to best accomplish a task for a set of training instances. In unsupervised learning, data mining methods are used to find patterns of any kind, without relationship to a particular target output. In supervised learning, data mining methods are used to predict the value of an interval or ordinal attribute or the class membership of a class attribute (categorical variable).

Examples of unsupervised learning tasks:

- Perform cluster analysis to identify subgroups of patients with similar demographic characteristics.
- Induce association rules that detect novel relationships among attribute-value pairs in a pediatric injury database.

Examples of supervised learning tasks:

- Predict the blood concentration of an anesthetic given the patient’s body weight, gender, and amount of anesthetic infused.
- Predict smoking cessation status based on health interview survey data.
- Predict the severity of medical outcome for a poison exposure, based on patient and exposure characteristics documented at the time of initial call to a poison control center.

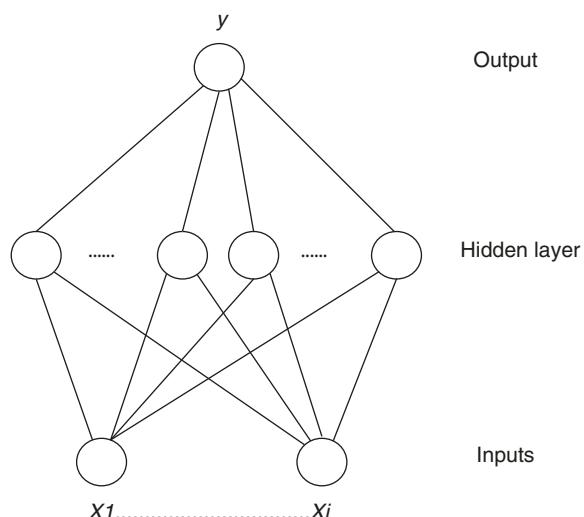
Data mining methods are numerous, and it is important to understand enough about each method to use it appropriately. Some methods are highly flexible, capable of modeling very complex decision boundaries (artificial neural networks, support vector machines), while other methods are advantageous because they can be readily understood (classification and regression trees, association rules). Bayesian methods are distinctive in modeling dependencies among data. A comprehensive description of data mining methods is beyond the scope of this chapter but can be found in any data mining textbook. This chapter includes only a brief description of several important methods.

Artificial Neural Networks

Artificial neural networks constitute one of the oldest and perpetually useful data mining methods. The most fundamental form of an artificial neural network, the threshold logic unit, was invented by McCulloch and Pitts at the University of Chicago during the 1930s and 1940s as a mathematical representation of frog neuron [8]. Contemporary artificial neural networks are multilayer networks composed of processing elements, variations of McCulloch and Pitt's original TLUs (Fig. 16.2). Weighted inputs to each processing element are summed, and if they meet or exceed a certain threshold value, they produce an output. The sum of the weighted inputs is a probability of class membership, and when deployed, the threshold of artificial neural networks can be adjusted for sensitivity or specificity.

Artificial neural networks make incremental adjustments to the weights according to feedback of training instances during a procedure for weight adjustment. Weight settings are initialized with random values, and the weighted inputs feed a network of processing elements, resulting in a probability of class membership and a

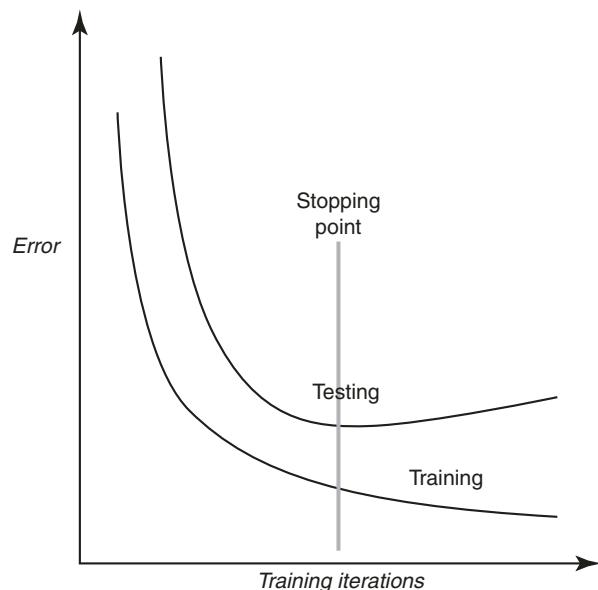
Fig. 16.2 Multilayer artificial neural network



prediction of class membership for each instance. The predicted class membership is then compared to the actual class membership for each instance. The model is incrementally adjusted, in a method specific to one of many possible training algorithms, until all instances are correctly classified or until the training algorithm is stopped. Because artificial neural networks incrementally adjust until error is minimized, they are prone to overtraining, modeling nuances, and noise in the training data set, in addition to valid patterns. In order to avoid overtraining, predictions are also incrementally made for a portion of data that has been set aside, not used for training. Each successive iteration of weights is used to predict class membership for the holdout data. Initially, successive iterations of weight configurations will result in decreased error for both the training data and the holdout data. As the artificial neural network becomes overtrained, error will increase for the holdout data and continue to decrease for the training data. This transition point is also the stopping point and is used to determine the optimal weight configuration (Fig. 16.3). Over multiple experiments, artificial neural networks can assume very different weight configurations but with varied configurations demonstrating equivalent performance.

Deep learning, [9] a powerful method for knowledge discovery used when very large amounts of data and training examples are available, is based upon artificial neural networks. In deep learning, the networks may have numerous layers and inputs, including multiple representation layers; the representation layers are refined in a “pre-training” step. This approach allows for effective, automatic identification of features, and so it effectively eliminates the need for more laborious forms of feature selection. Deep learning has led to extraordinary breakthroughs in image and language processing [10]. Its utility in modeling human behavior and health outcomes is not yet well characterized.

Fig. 16.3 Training/testing curves



Decision Trees

Decision trees, methods including classification and regression trees (CART) and an almost identical method known as C4.5, developed in parallel by Quinlan and others in the early 1980s [11]. These methods are used for supervised learning tasks and induce tree-like models that can be used to predict the output values for new cases. In this family of decision tree methods, the data is recursively partitioned based on attribute values, either nominal values or groupings of numeric values. A criterion, usually the information gain ratio of the attributes, is used to determine the order of the attributes in the resulting tree. Unless otherwise specified, these methods will induce a tree that classifies every instance in the training data set, resulting in an overtrained model. However, models can be post-pruned, eliminating leaves and nodes that handle very few instances and improving the generalizability of the model.

Decision trees are readily comprehensible and can be used to understand the basic structure of a pattern in data. They are sometimes used in the preprocessing stage of data mining to enhance data cleaning and feature subset selection. The use of decision tree induction methods early in the KDD process can help identify the persistence of rogue variables highly correlated with the output that are inappropriate for inclusion. However, ensembles of multiple decision trees, such as those utilized in random forest methods, tend to outperform single decision trees.

Support Vector Machines

Support vector machine methods were developed by Vapnik and others in the 1970s through the 1990s [12–14]. Support vector machines, like artificial neural networks, can be used to model highly complex, nonlinear solutions; however, they require the adjustment of fewer parameters and are less prone to overtraining. The method implements a kernel transformation of the feature space (attributes and their values) and then learns a linear solution to the classification problem (or by extension, regression) in the transformed feature space. The linear solution is made possible because the original feature space has been transformed to a higher-dimensional space. Overtraining is avoided through the use of maximal margins, margins that parallel the optimal linear solution and that simultaneously minimize error and maximize the margin of separation.

***k*-Nearest Neighbor**

The *k*-nearest neighbor classification method (a common classification method and so-called “hot deck” method in missing value imputation) infers binary class membership on the basis of known class membership for similar instances. The output is inferred based on the majority class value for similar instances. This is a relatively simple algorithmic approach to classification. It has been shown robust in the

presence of missing values and with large numbers of attributes [15]. It is a case-based reasoning method that learns pattern in the training data only when it is required to classify each new testing instance.

Association Rules

Association rule induction is a method used for unsupervised learning. This method is used to identify if-then relationships among attribute-value pairs of any kind. For example, a pattern this algorithm could learn from a data set would be If COLOR=red, then FRUIT=apple. Higher-order relationships can also be found using this algorithm. For example, If COLOR=red and SKIN=smooth, then FRUIT=apple. Relationships among any and all attribute-value combinations will be described, regardless of importance. Many spurious relationships will typically be described, in addition to meaningful and informative relationships. The analyst must set criteria and limits for the order of relationships described, the minimum number of instances (evidence), and percentage of instances for which the relationship is true (coverage).

Bayesian Methods

Bayesian networks (in general) are networks of variables that describe the conditional probability of class membership based on the values of other attributes in the data. For example, a Bayesian network to predict the presence or absence of a disease would model P(disease symptoms). That conditional probability is then used to infer class membership for new instances. The structure and probabilities of the network can be directly induced from data, and the structure can be specified by domain experts with probabilities derived from actual data. These models become complex as joint probability distributions become necessary to model dependencies among input data. Naïve Bayes is the most fundamental form of these methods, in which conditional independence between the input variables is assumed (thus the descriptor “naïve”).

Interpretation and Evaluation

For supervised learning tasks, an output is specified, and a predictive model is induced. The error of induced models in predicting the output, whether the output is a real number or class membership, is used to evaluate the models. These metrics can be calculated by applying the model to predict outputs for data where actual output is known and comparing the predicted outputs to the actual outputs. For real number outputs, the error is the difference between the actual and predicted outputs. Error terms, including LMS error and RMSE, are used to quantify error.

For class variable outputs, error is misclassification. Each prediction constitutes a true positive, true negative, false positive, or false negative, and a confusion matrix is constructed from which various accuracy metrics are derived. Many data mining methods produce models that calculate a probability of class membership, to which a threshold is applied. At any given threshold, the confusion matrix may change. A higher threshold will result in fewer false positives, while a lower threshold will maximize sensitivity. This is advantageous in that the threshold can be adjusted in order to optimize these parameters for clinical applications. However, the predictive performance of the model cannot be adequately represented by metrics calculated with a single threshold confusion matrix. Instead, receiver operating curve (ROC) analysis is used.

An ROC curve is derived from the confusion matrix, by plotting the true-positive fraction vs. the false-positive fraction. Hanley and McNeil [16] define the index known as the area under the ROC curve as the probability that a randomly chosen subject of a given class will be predicted to belong to that class versus a randomly chosen subject that does not belong to that class [16]. ROC analysis originated in Great Britain during World War II, as a method of quantifying the ability of submarine sonar operators to distinguish signal indicating the presence of enemy ships. It was later adopted in radiology to quantify diagnostic accuracy. A detailed discussion of ROC analysis, specific to knowledge discovery and data mining in biomedical informatics, is found in Lasko et al. [17].

In order to obtain unbiased estimates of accuracy, it is necessary to calculate accuracy of model performance on a set of data that has not been used in training, testing, or model selection. This validation data set must be set aside before data mining methods are applied. Validation data sets differ from testing data sets. While validation data sets are not used during the data mining step, testing data sets are used in an interactive fashion to select model parameters and architecture. When cross validation is used, each testing instance also serves as a training instance. Even if cross validation is not used, and testing data sets do not contribute training instances, testing data sets are certainly used to compare and make choices about model parameters during the data mining step of the KDD process, so any estimates of accuracy calculated using testing data are biased. It is necessary to calculate accuracy using an entirely separate body of data, the validation set. Data partitioning, the assignment of available instances to training, testing, and validation data sets, is critical to interpretation and evaluation in KDD.

Applications of Knowledge Discovery and Data Mining in Clinical Research

Knowledge discovery and data mining methods have been used in numerous ways to generate hypotheses for clinical research.

Knowledge discovery and data mining methods are especially important in genomics, a field rich in data but immature in knowledge. In this area of biomedical research, exploratory approaches to hypothesis generation are accepted, even

necessary, in order to accelerate knowledge development. Data mining methods are often used to identify genetic markers of disease and genotype-phenotype associations for closer examination. For example, microarray analysis employs automated machine learning and statistical methods to identify patterns and associations in gene expression relevant for genetic epidemiology, pharmacogenomics, and drug development [18].

While KDD and data mining methods have demonstrated their ability to discern patterns in large, complex data, their usefulness in identifying patterns across biomedical, behavioral, social, and clinical domains is tempered by the disparate ways in which data is represented across research databases. It is difficult to aggregate clinical and genomic data, for instance, from diverse sources because of differences in coding and a lack of syntactic and semantic interoperability. Currently, a great deal of effort is being devoted to development of systems and infrastructure to facilitate sharing and aggregation of data.

Commonly Encountered Challenges in Data Mining

Rare Instances

Rare instances pose difficulty for knowledge discovery with data mining methods. In order for automated pattern search algorithms to learn differences that distinguish rare instances, there must be adequate instances. Also, during the data mining step of the KDD process, rare instances must be balanced with no instances for pattern recognition. If only 1 out of every 100 patients in a healthcare system has a fall incident, a sample of instances would be composed of 1% fall and 99% no-fall patients. Any classification algorithm applied to this data could achieve 99% accuracy by universally predicting that patients do not fall. If the sample is altered so that it is composed of 50% fall and 50% no-fall patients or if weights are applied, true patterns that distinguish fall patients from no-fall patients will be recognized. Afterwards, the models can be adjusted to account for the actual prior probability of a fall. In cases where inadequate instances exist, rare instances can be replicated, weighted, or simulated.

Sources of Bias

Mitigation of bias is a continual challenge when using clinical data. Many diverse sources of bias are possible in secondary analysis of clinical data. Verification bias is a type of bias commonly encountered when inducing predictive models using diagnostic test results. Because patients are selected for diagnostic testing on the basis of their presentation, the available data does not reflect a random sample of patients. Instead, it reflects a sample of patients heavily biased toward presence of a disease state. Another troublesome source of bias relates to inadequate reference standards (gold standards). Machine learning algorithms are trained on sets of

instances for which the output is known, the reference standard. However, clinical data may not include a coded, sufficiently granular representation of a given disease or condition. Even then, the quality of routinely collected clinical data can vary dramatically [6]. Diagnoses may also be incorrect, and source data, such as lab and radiology results, may require review by experts in order to establish the reference standard. If this additional step is necessary to adequately establish the reference standard, the time and effort necessary to prepare an adequate sample of data may be substantial. For an extended discussion of these and other sources of bias, the reader is referred to Pepe [19].

Many concepts in medicine and healthcare are not precisely defined or consistently measured across studies or clinical sites. Changes in information systems certainly influence the measurement of concepts and the coding of the data that represents those concepts. When selecting a subset of retrospective clinical data for analysis, it is wise to consult with institutional information technology personnel who are knowledgeable about changes in systems and databases over time. They may also be aware of documents and files describing clinical data collected using legacy systems, information that could be crucially important.

Limitations

The limitations in using repositories of clinical data for research are related to data availability, data quality, representation and coding of clinical concepts, and available methods of analysis. Since clinical information systems only contain data describing patients served by a particular healthcare organization, clinic, or hospital, the data represent only the population served by that organization. Any analysis of data from a single healthcare organization is, in effect, a convenience sample and may not have been drawn from the population of interest.

Data quality can vary widely and is strongly related to the role of data entry in workflow. For example, one preliminary study of data describing smoking status revealed that the coded fields describing intensity and duration of smoking habit were completed by minimally educated medical assistants, instead of nurse practitioners or physicians. Data describing intensity and duration of smoking habit were also plagued by absurdly large values. These values may have been entered by medical assistants when the units of measurement enforced by the clinical information system did not fit descriptions provided by patients. For example, there are 20 cigarettes in a pack. When documenting the intensity of the smoking habit, a medical assistant may have incorrectly entered “10” instead of “0.5” into a field with the unit of measurement “packs per day,” not “number of cigarettes per day” [6].

Infrastructure for Knowledge Discovery

The power of the KDD process, and of data mining methods, to enable large-scale knowledge discovery lies in their singular capacity to identify previously unknown

patterns, in data sets too large and complex for human pattern recognition. However, in order to identify true and complete patterns, all the relevant concepts must be represented in the data. Representations of key concepts, whether gene expression, environmental exposure, or treatment, often exist. However, they exist in siloed data repositories, owned by different scientific groups. Development of systems and infrastructure to support sharing and aggregation of scientific data is essential for understanding complex multifactorial relationships in biomedicine. The potential of KDD for advancing biomedical knowledge will not be fully realized until these systems and infrastructure are in place.

One earlier and influential infrastructure project in the United States was caBIG®, the cancer biomedical informatics grid. This project addressed barriers posed by lack of interoperability and siloed data by promoting fundamental change in the way clinical research is conducted. caBIG® collaborators developed open-source tools and architecture that enable federated sharing of interoperable data, using an object-oriented data model and standard data definitions. In early 2009, the University of Edinburgh became the first European university to deploy a caBIG application, caTISSUE repository [20]. However, in 2012, caBIG in the United States was reassessed.¹ The activities of the cancer Biomedical Informatics Grid (caBIG) program of the National Cancer Institute (NCI) were integrated into the Institute's new National Cancer Informatics Program (NCIP). NCIP provides many biomedical informatics resources for the cancer research community.

Another major approach to facilitating biomedical knowledge discovery has been that of the semantic web [21]. The semantic web is an extension of current web-based information retrieval that enables navigation and retrieval of resources using semantics (meaning) in addition to syntax (specific words or representations). Development of the semantic web is broadly important for information retrieval and use but specifically valuable for biomedical research because of its ability to make scientific data retrievable and usable across disciplines and scientific groups. In a recent methodological review, Ruttenberg and colleagues emphasized the importance of scientific ontology, standards, and tools development for the semantic web in order for biomedical research to realize the benefits. All-purpose semantic web schema languages RDFS and OWL can be used to manage relationships among data elements in information systems used to manage clinical studies. “Middle” ontologies are being developed to specifically address data relationships in scientific work [21].

Enterprise data warehouses (EDW) are repositories of clinical and operational data, populated by source systems but completely separate from those systems. EDWs facilitate secondary analysis by integrating data from diverse systems in a single location. The data is not used to support patient care or operations. It exists in a stand-alone repository optimized for secondary analysis. Typically, a layer of analytic tools is used to support queries and OLAP (online analytic processing). In some healthcare organizations, all clinical data may be warehoused. In other organizations, data collected by certain systems may be excluded, or certain types of

¹ Kush R. Where is caBIG Going? [Internet]. CDISC Website. 2012. Available from: <http://www.cdisc.org/where-cabig-going?>

data may be excluded. In these cases, data extracted from the EDW may need to be aggregated with data stored only in source systems. It is crucially important that data warehouses be optimized to facilitate scientific analytics. The way in which the data is stored and the development of powerful tools for examining and extracting the data directly influence the feasibility and quality of knowledge discovery using the data.

Success in aggregating data from diverse sources representing the spectrum of factors that affect human health, such as genomics, geography and community characteristics, social and behavioral determinants of health, environmental exposures, and healthcare, could enable unprecedented system-level insight into human health, using methods of knowledge discovery and data mining. In fact, the National Institutes of Health has launched a large initiative, the Child Health Outcomes (ECHO) Program, to create the infrastructure to support large cohort studies that can accomplish these types of analyses [22]. Pediatric asthma is an example of a disease thought to be influenced by multiple factors, including genomics, social and behavioral determinants of health, healthcare, and environmental air quality. In recent years, the NIH National Institute for Biomedical Imaging and Bioengineering funded PRISMS (Pediatric Research Using Integrated Sensor Monitoring Systems), a large scientific project aimed at achieving system-level insight in pediatric asthma. The PRISMS project is advancing the development of air quality sensors, both personal and environmental, optimized for use in research. However, it is also devoting resources to the development of informatics centers such as University of Utah's Utah PRISMS Center. The Utah PRISMS Center along with a partner informatics center located at the University of California, Los Angeles, is developing an informatics platform capable of receiving, processing, and storing the large quantities of data generated by sensors and producing data sets for analysis. A data coordinating center, currently based at the University of Southern California, then facilitates data integration and analysis. This project will enable exposomic research related to pediatric asthma, at varied spatiotemporal scale [23, 24].

Conclusion

Knowledge discovery and data mining methods are important for informatics because they link innovations in data management and storage to knowledge development. The sheer volume and complexity of modern data stores overwhelms statistical methods applied in a more traditional fashion. In the past, the inductive approach of data mining and knowledge discovery has been criticized by the statistical community as unsound. However, these methods are increasingly recognized as necessary and powerful for hypothesis generation, given the current data deluge. Hypotheses generated through the use of these methods, and unknown without these methods, can then be tested using more traditional statistical approaches. As the statistical community increasingly recognizes the advantages of machine learning methods and engages in knowledge discovery, the line between the statistical and machine learning worlds becomes increasingly blurred [25].

Much criticism is tied to the iterative and interactive nature of the knowledge discovery process, which is not consistent with the very sequential scientific method. Indeed, it is very important that data mining studies be replicable. In order for studies to be replicable, it is important that the analyst keep detailed records, particularly as data is transformed and sampled. It is also crucial that domain experts be involved in decision-making about data selection and feature selection and transformation, as well as the iterative evaluation of models. The quality of resultant models is evidenced by performance on new data, and models should be validated on unseen data whenever possible. Models also must be calibrated for the target population with which they are being used. Uncalibrated models will certainly lead to increased error [26].

While the data deluge is very real, our technology for optimally managing and structuring that data lags behind. In clinical research, data mining and knowledge discovery awaits the further development of high-quality clinical data repositories. Many data mining application studies in the biomedical literature find that model performance is limited by the concepts represented in the available data. For optimal use of these methods, all relevant concepts in a particular area of interest must be represented. The old adage “garbage in, garbage out” applies. If a health behavior (i.e., smoking) is believed to be related to biological, social, behavioral, and environmental factors, a data set composed of only biological data will not suffice. Additionally, much of the data being accumulated in data warehouses is of varied quality and is not collected according to the more rigorous standards employed in clinical research. As more sophisticated systems for coding and sharing data are devised, we find ourselves increasingly positioned to apply data mining and knowledge discovery methods to high-quality data repositories that include most known and possibly relevant concepts in a given domain.

In the ever-intensifying data deluge, knowledge discovery methods represent one of several pivotal tools that may determine whether human welfare is advanced or diminished. It is important for scientists engaged in clinical research to develop familiarity with these methods and to understand how they can be leveraged to advance scientific knowledge. It is also critical that clinical scientists recognize the dependence of these methods upon high-quality data, well-structured clinical data repositories, and data sharing initiatives.

References

1. Benson K, Hartz AJ. A comparison of observational studies and randomized, controlled trials. *Am J Ophthalmol.* 2000;130(5):688.
2. Aronsky D, Fiszman M, Chapman WW, Haug PJ. Combining decision support methodologies to diagnose pneumonia. In: *Proceedings of the AMIA symposium;* 2001. p. 12–6.
3. Lagor C, Aronsky D, Fiszman M, Haug PJ. Automatic identification of patients eligible for a pneumonia guideline: comparing the diagnostic accuracy of two decision support models. *Stud Health Technol Inform.* 2001;84(Pt 1):493–7.
4. Fayyad U, Piatetsky-Shapiro G, Smyth P. From data mining to knowledge discovery in databases. *AI Mag.* 1996;17(3):37–54.

5. Aronsky D, Haug PJ, Lagor C, Dean NC. Accuracy of administrative data for identifying patients with pneumonia. *Am J Med Qual.* 2005;20(6):319–28. <https://doi.org/10.1177/1062860605280358>.
6. Poynton MR, Frey L, Freg H. Representation of smoking-related concepts in an electronic health record. In: *Medinfo 2007: Proceedings of the 12th world congress on health (medical) informatics; building sustainable health systems*; 2007. p. 2255.
7. Minsky M. *The society of mind*. New York: Simon & Schuster; 1986.
8. McCulloch WS, Pitts W. A logical calculus of the ideas immanent in nervous activity. *Bull Math Biophys.* 1943;5(4):115–33. <https://doi.org/10.1007/BF02478259>.
9. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature.* 2015;521(7553):436.
10. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. *Commun ACM.* 2017;60(6):84–90. <https://doi.org/10.1145/3065386>.
11. Quinlan JR. *C4. 5: programs for machine learning*. Oxford: Elsevier; 2014.
12. Cristianini N, Shawe-Taylor J. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge, UK: Cambridge University Press; 2000.
13. Vapnik VN. *The nature of statistical learning theory*. New York: Springer; 1995. p. 188.
14. Vapnik VN. *Statistical learning theory*. New York: Wiley; 1998. p. 736.
15. Jonsson P, Wohlin C. Benchmarking k-nearest neighbour imputation with homogeneous likert data. *Empir Softw Eng.* 2006;11(3):463–89. <https://doi.org/10.1007/s10664-006-9001-9>.
16. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology.* 1982;143(1):29–36. <https://doi.org/10.1148/radiology.143.1.7063747>.
17. Lasko TA, Bhagwat JG, Zou KH, Ohno-Machado L. The use of receiver operating characteristic curves in biomedical informatics. *J Biomed Inform.* 2005;38(5):404–15. <https://doi.org/10.1016/j.jbi.2005.02.008>.
18. Cordero F, Botta M, Calogero RA. Microarray data analysis and mining approaches. *Brief Funct Genomics.* 2007;6(4):265–81. <https://doi.org/10.1093/bfgp/elm034>.
19. Pepe MS. *The statistical evaluation of medical tests for classification and prediction*. Oxford: Oxford University Press; 2003. ISBN 9780198509844.
20. Breiman L. Statistical modeling: the two cultures (with comments and a rejoinder by the author). *Stat Sci.* 2001;16(3):199–231. <https://doi.org/10.1214/ss/1009213726>.
21. Genomeweb. Persistent systems helps first european deploy cabig's catissue repository. 2009.
22. Ruttenberg A, Clark T, Bug W, Samwald M, Bodenreider O, Chen H, Doherty D, Forsberg K, Gao Y, Kashyap V, Kinoshita J, Luciano J, Marshall MS, Ogbuji C, Rees J, Stephens S, Wong GT, Wu E, Zaccagnini D, Hongsermeier T, Neumann E, Herman I, Cheung K-H. Advancing translational research with the semantic web. *BMC Bioinforma.* 2007;8(3):S2. <https://doi.org/10.1186/1471-2105-8-s3-s2>.
23. Program E. Environmental influences on child health outcomes (echo) program. 1/30/2018), ECHO supports multiple longitudinal studies using existing study populations to investigate environmental exposures on child health and development.
24. Burnett N. Harmonization of sensor measurement to support health research. In: *Proceedings of the national conference of undergraduate research 2017.* 2017.
25. Kelly KE, Whitaker J, Petty A, Widmer C, Dybwad A, Sleeth D, Martin R, Butterfield A. Ambient and laboratory evaluation of a low-cost particulate matter sensor. *Environ Pollut.* 2017;221:491–500. <https://doi.org/10.1016/j.envpol.2016.12.039>.
26. Matheny ME, Ohno-Machado L, Resnic FS. Discrimination and calibration of mortality risk prediction models in interventional cardiology. *J Biomed Inform.* 2005;38(5):367–75.



Advancing Clinical Research Through Natural Language Processing on Electronic Health Records: Traditional Machine Learning Meets Deep Learning

17

Feifan Liu, Chunhua Weng, and Hong Yu

Abstract

Electronic health records (EHR) capture “real-world” disease and care processes and hence offer richer and more generalizable data for comparative effectiveness research than traditional randomized clinical trial studies. With the increasingly broadening adoption of EHR worldwide, there is a growing need to widen the use of EHR data to support clinical research. A big barrier to this goal is that much of the information in EHR is still narrative. This chapter describes the foundation of biomedical language processing and explains how traditional machine learning and the state-of-the-art deep learning techniques can be employed in the context of extracting and transforming narrative information in EHR to support clinical research.

F. Liu, PhD (✉)

Department of Quantitative Health Sciences and Department of Radiology (Joint),
University of Massachusetts Medical School, Worcester, MA, USA

e-mail: feifan.liu@umassmed.edu

C. Weng, PhD

Department of Biomedical Informatics, Columbia University, New York, NY, USA
e-mail: chunhuaweng@columbia.edu

H. Yu, PhD

Department of Computer Science, University of Massachusetts Lowell, Lowell, MA, USA
Bedford VA Medical Center, Bedford, MA, USA

Department of Medicine, University of Massachusetts Medical School (Adjunct),
Worcester, MA, USA

College of Information and Computer Sciences, University of Massachusetts Amherst
(Adjunct), Amherst, MA, USA
e-mail: hong_yu@umass.edu

Keywords

Electronic health records · Biomedical natural language processing · Rule-based approach · Machine learning · Deep learning · Clinical research

Electronic health records (EHR) capture “real-world” disease and care processes and hence offer richer and more generalizable data for comparative effectiveness research [1] than traditional randomized clinical trial studies. With the increasingly broadening adoption of EHR worldwide, there is a growing need to widen the use of EHR data to support clinical research [2]. A big barrier to this goal is that much of the information in EHR is still narrative. This chapter describes the foundation of biomedical language processing and how traditional machine learning and the state-of-the-art deep learning techniques can be employed in the context of extracting and transforming narrative information in EHR to support clinical research.

Accelerating Clinical Research Using EHR: Opportunities and Challenges

The NIH defines clinical research as *patient-oriented research, epidemiological and behavioral studies, or outcomes and health services research* [3]. Patient-oriented research involves a particular person or group of people or uses materials from humans. In recent years, national clinical research enterprises have been under increased jeopardy [4] in part due to the rising costs associated with participant screening and recruitment, as well as issues surrounding data collection. Only 13% of clinicians are involved in clinical research [5]. To integrate research with clinical care, and to speed the application of research findings to clinical practice, the NIH has created the CTSA (Clinical and Translational Science Awards) program to reengineer the clinical research enterprise [6]. A potential powerful accelerator to clinical research is electronic health records (EHR).

An EHR is a legal computerized medical record for documenting patient information captured at every patient encounter [7, 8]. Figure 17.1 shows a sample EHR [9]. As of 2008, more than 40% of physicians in the United States are using EHRs, more than double the percentage at the start of the decade [10]. The resident population of the United States as of 2009 was 307 million [11]. During that same year, it was reported that 83% adults and 90% children had contact with a healthcare professional, there were 1.1 billion ambulatory care visits (to physician offices, hospital outpatient, and emergency departments), and the number of physician office visits was 902 million. In other words, there were possibly over 800 million record entries in EHR in 2009.

EHRs offer great potential to improve the efficiency and reduce the cost for clinical research, but this potential has not yet been fully realized. EHR includes standards-based structured laboratory test results and narrative interpretations by care providers. Unstructured narrative information can be provided for admission

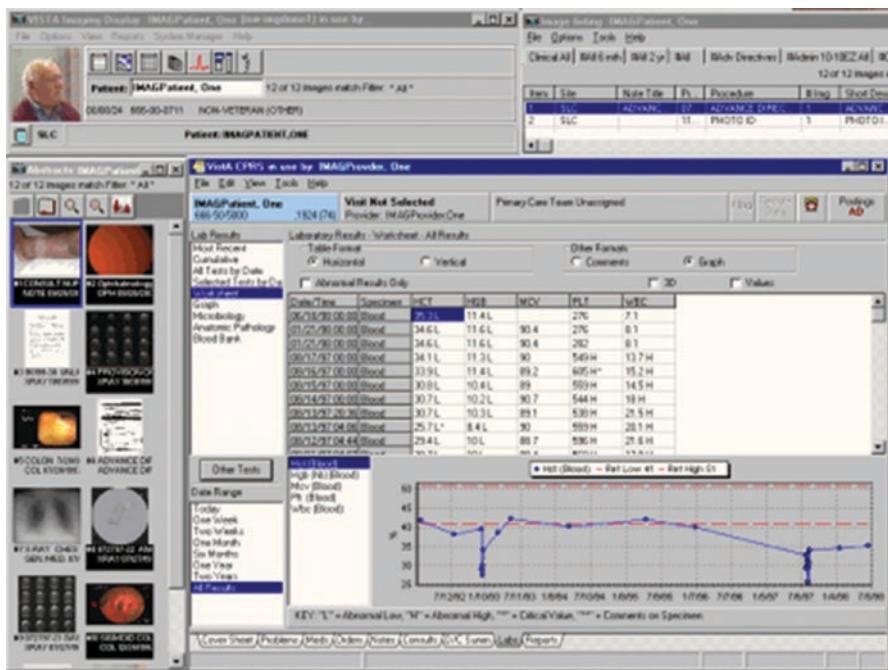


Fig. 17.1 Illustration of a sample electronic health record (EHR)

note, discharge summary, radiology images, and all sorts of ancillary notes, etc. Unlocking discrete data elements from such narrative information is a big challenge for reusing EHR data for clinical research.

Many studies and demonstration projects have explored the use of EHR data for clinical research, including detecting possible vaccination reactions in clinical notes [12], identifying heart failure [13], classifying whether a patient has rheumatoid arthritis [14], identifying associations between diabetes medications and myocardial infarction [15], and predicting disease outcomes [16]. EHR data has also been used for computerized pharmacovigilance [17] (see Chap. 20). Below, we elaborate two common use cases as examples of applying information extraction and retrieval techniques in EHR to support clinical research.

- *Use case 1: eligibility screening or phenotype retrieval*

The foremost, albeit costly, information retrieval task in clinical research is eligibility screening, which is to determine whether a person may or may not be eligible to enter a clinical research study. Chute has described this as essentially “patient phenotype retrieval” since it is meant to identify patients who manifest certain characteristics, which include diagnosis, signs, symptoms, interventions, functional status, or clinical outcomes [18]. Such characteristics are generally described in the eligibility criteria section for a research protocol. In recent years, the increasing

volume of genome-wide association studies also raised the demand for clinical phenotype retrieval in discovering the genetics underlying many medical conditions. Traditional methods of participants search through manual chart review cannot scale to meet this need. In the study of rare diseases, there are usually only a small number of patients available, so it is feasible to have research personnel carefully collect, record, and organize the phenotypic information of each study participant. Diseases like diabetes mellitus, hypertension, and obesity, however, are complex, multifactorial, and chronic, and it is likely that a large number of patients will need to be followed over an extended period to ascertain important phenotypic traits. Large-scale studies involving many participants, or even smaller studies in which participants are selected from a larger population, will require innovative means to extract reliable, useful phenotype information from EHR data.

In recent years, several academic institutions have used EHR data to electronically screen (E-Screen) eligible patients for clinical studies [19]. Manually screening charts is time-consuming for research personnel, who must search for information in patient records to determine whether a patient meets the eligibility criteria for a clinical trial. E-Screening, however, can exclude ineligible patients and establish a much smaller patient pool for manual chart review. Thus, E-Screening helps clinical research personnel transition from random and burdensome browsing of patient records to a focused and facilitated review. Consistent with concerns for patient safety and trial integrity, clinical research personnel should review all patients classified as “potentially eligible” by E-screening to confirm their eligibility. E-screening systems essentially perform “pre-screening” for clinical research staff and should not fully replace manual review.

- *Use case 2: secondary use of clinical data for research*

The national movement toward the broad adoption of EHRs obviously means that more clinical data will be captured and stored electronically. Secondary use of data for clinical research is a competitive requirement for a clinical and research enterprise [20]. In late 2009, the National Center for Research Resources called for “widening the use of electronic health records for research” to strengthen our capacity for using clinical care data for research. The nation’s transition from traditional clinical trials to comparative effectiveness research [21] led by the US Government has further emphasized the need for effective tools to extract research variables from pre-existing clinical data. As an example, i2b2 (Informatics for Integrating Biology and the Bedside) is an NIH-funded National Center for Biomedical Computing based at Partners HealthCare System. The i2b2 Center is developing a scalable informatics framework that will enable clinical researchers to use existing clinical data for discovery research. In addition to that, the US Office of the National Coordinator for Health Information Technology (ONC) recently awarded \$60 million in research grants through the Strategic Health IT Advanced Research Projects (SHARP) Program to the Mayo Clinic College of Medicine for secondary use of EHR data research.

A major challenge of using EHRs to facilitate clinical research is that much EHR data is presented as clinical narratives, which is largely unstructured and poses machine readability problems. Clinical natural language processing has been an active field since the inception of EHR in the 1960s and is an area that explores tools that can effectively extract, mine, and retrieve clinically relevant structured data from narrative EHRs. Clinical natural language processing has been influenced by the theory of sublanguage, which is characterized by distinctive specializations of syntax and the occurrence of domain-specific word subclasses in particular syntactic combinations. More recently clinical NLP has been experiencing a shift from rule-based approach to machine learning methods, as discussed later.

The rest of this chapter is organized as follows. We will first introduce the fundamental concepts of clinical NLP research in terms of sublanguage analysis and traditional machine learning models. Then we review current progress of clinical NLP research including the state-of-the-art deep learning techniques employed in recent years, followed by discussions on the challenges and future directions in this field.

Foundations of Biomedical Natural Language Processing

Natural language processing (NLP) is a research field dedicated to enable computers with the right knowledge for understanding natural language text, ultimately to facilitate the different types of natural language interaction between humans and computers. Biomedical NLP is a subfield specified for biomedical texts from biology, medicine, and chemistry. There exists great variability in the language in each of these areas, as reflected in their respective literature, guidelines, etc. In addition, the same type of biomedical text, such as narrative in an EHR, as discussed earlier, could differ greatly due to the expression variances and some organization-specific variance (i.e., among different medical centers). Sublanguage and machine learning theory and approaches lay strong foundations for developing efficient clinical NLP systems in many real-world applications. Although some approaches and models are described below in the context of biomedical NLP applications, all of them can be adapted on electronic health record (EHR) for clinical research informatics.

Sublanguage Approach

A sublanguage is defined by Grishman [22] as a specialized form of natural language used to describe a limited subject matter, generally employed by a group of specialists dealing with a particular subject. Zellig Harris [23] was one of the first linguists to apply the term sublanguage to natural language, using algebra as the underlying formalism. He defines a sublanguage as a subset of the language that is closed under some or all of the operations of the language.

Sublanguage theory laid a foundation for NLP in specific contexts such as the clinical narratives. Many NLP applications are developed by exploiting the sublanguage characteristics, i.e., restricted domain syntax and semantics. For example, an electronic health record (EHR) is limited to discussions of patient care and is unlikely to cover gene annotations or cell line issues as in the biomedical literature.

Sublanguages have many unique properties in comparison to more everyday language, resulting in a specialized vocabulary, structural patterns, as well as specialized entities and relationships among them.

Vocabulary Level

A sublanguage tends to have a specialized vocabulary which is quite different from standard language. For example, “cell line” is unlikely to be mentioned in non-biological documents. In particular, the development of scientific and technological advancements in the biomedical domain has led to the discovery of new biological objects, functions, and events, which can only be acquired by analyzing sublanguage in the corresponding corpus.

Syntax Level

A sublanguage is not merely an arbitrary subset of sentences and may differ in syntax structure as well as vocabulary. For example, in medicine, telegraphic sentences such as “patient improved” are grammatical, due to operations that permit dropping articles and auxiliaries. In addition, there are certain patterns of expression in sublanguage consisting of predicate words and ordered arguments, as in “<antibody> <appeared in> <tissue>,” “appeared in” is predicate words and “<antibody>” and “<tissue>” are two arguments which can have semantically related terms filled in.

Semantics and Discourse Level

In addition to differences on the vocabulary and syntax levels, a sublanguage may also have specialized ways of interpreting language and organizing larger units of discourse. For example, “secondary to” has a specialized meaning that indicates a causal relationship, which is different from its use in standard language. In discharge summaries, the structural format often includes history of present illness, medications on admission, social history, physical examination, etc.

These properties of sublanguages allow the use of methods of analysis and processing that would not be possible when processing the language of newspaper articles or novels. Sublanguage analysis also provides a way of integrating domain-specific knowledge with existing systems. For example, a biomedical information retrieval system can be developed by indexing medical articles on only terms from a list of terminology known to be of interest to researchers; controlled medical vocabulary can be derived using sublanguage analysis based on terms’ combining regularly with particular other words; biological information extraction system can be adapted by sublanguage analysis of specialized expression patterns; a system that analyzes clinical reports can look for predictable semantic patterns that are characteristic of the clinical domain [24–27].

Machine Learning Approach

Sublanguage patterns (rules) and manually specified models often lack the quality of generalization and also are time-consuming to keep well maintained and updated. With the ever-growing availability of electric biomedical resource data and advanced computational power, machine learning models have been arousing intense interests for many biomedical NLP tasks, which can be mainly divided into five categories:

- Classification: assign documents predefined labels.
- Ranking: order objects by preference.
- Regression: obtain real value output as prediction.
- Structured prediction: sequence labeling and segmentation to recognize entities or other semantic units.
- Clustering: discover the underlying structure of unlabeled data to form natural groups.

Many clinical research informatics applications can be formulated into the above-mentioned tasks, such as entity (medications, diseases, doses) extraction from EHRs can be realized using structured prediction models and adverse events detection from EHRs is an example of classification tasks. For these tasks, the goal of machine learning is to enable correct predictions for target variables given observation variables (attributes or features) from corresponding instances. Different learning models have been applied in recent years. In terms of their modeling approaches, they can be grouped as generative models and discriminative models. The generative approach models a joint probability distribution over both input and output variables (observation and label sequences), such as naive Bayes, Bayesian network, hidden Markov model, and Markov random field, while the discriminative approach directly models the dependence of the output variables (label to be predicted) on the input variables (observation) by conditional probability, such as decision tree, logistic regression, support vector machine, K nearest neighbor, artificial neural network, and conditional random fields. This section will cover the introductory descriptions of those algorithms, but we encourage interested readers to explore these in more detail through further readings [28–32].

Generative Model

The generative model is a full probability model on all variables, which can simulate the generation of values for any variables in the model. By using Bayes' rule, it can be formed as a conditional distribution to be used for classification. When there is little annotated data, the generative model is advantageous for making use of a large quantity of raw data for better performance. The generative model reduces the variance of parameter estimation by modeling the input, but at the expense of possibly introducing model bias.

1. Naive Bayes Classifier

The naive Bayes classifier is based on Bayes' theorem [33] and is a very simple probabilistic generative model that can be used to compute the probability of each

candidate class label given observed features, under the assumption that all the features are independent given class label. It requires only a small size of training data with faster parameter estimation, but the strong independence assumption is violated in numerous occasions for real applications, which can lead to a large bias.

2. Bayesian Network

Bayesian network [34], also belief network, is a probabilistic graphical model, whose nodes are a set of random variables connected by a directed acyclic graph (DAG) to represent the conditional dependences among those variables. This model doesn't require the independence assumption as in naive Bayes, providing stronger representational power in real-world applications and making the parameter estimation more complex, as well. It models the dependency between variables providing a good ability to handle missing values and is widely used in causal relationship reasoning applications, such as clinical decision support [35] and gene expression data analysis [36].

3. Hidden Markov Model

The hidden Markov model (HMM) [37] is a probabilistic generative model of a Markov process (Markov chain), where the model passes different state sequences, which are unobserved, producing a sequence of observations. Each hidden state has a probability distribution over the possible output observations, and there are transition probabilities among those states.

HMM is widely used in temporal pattern recognition (e.g., medical dictation system) and other sequence labeling tasks (e.g., gene/protein recognition [38] and bio-sequence alignment [39]). Although this type of statistical model has worked extremely well in many situations, it does have limitations. A major limitation is the assumption that successive observations are independent, which can't take into account the contextual dependency in the observation sequence. Another limitation is the Markov assumption itself, i.e., the current state only depends on the immediate preceding state, which is also inappropriate for some problems.

4. Markov Random Field

Markov random field (MRF), also a Markov network or undirected graphic model [40], is a graphic model on the joint probability over a set of random variables each corresponding to a node in the graph. Markov properties exist among those variables to provide conditional independence for graph factorization.

MRF is similar to Bayesian network in terms of modeling dependency relationships among variables. Bayesian network is a direct graphic model, and it represents probability distributions that can be factorized into products of conditional distributions, which is desirable to capture causal relationships among variables, while MRF is an undirected graphic model, where there is no directionality on each edge connecting a pair of nodes, and the probability distribution it represents will be

factorized into products of potential functions of conditionally independent cliques [28], which makes MRF better suited to expressing soft constraints between random variables. In addition, MRF can represent certain dependencies that a Bayesian network cannot, such as cyclic dependencies, wherein it can't represent certain dependencies that a Bayesian network can such as induced dependencies. MRF model has been successfully applied in biomedical image analysis for computer-aided diagnosis as shown in [41, 42].

Discriminative Model

Compared with the generative model, the discriminative model is designed to only involve a target variable(s) conditional on the observed variables, directly computing the input to output mappings (posterior) and eschewing the underlying distributions of the input. As there are fewer independence assumptions, the discriminative model often provides more robust generalization performance when enough annotated data is available. However, it usually lacks flexible modeling methods for prior knowledge, structure, uncertainty, etc. In addition, the relationships between variables are not as explicit or visualizable as in the generative model.

1. Decision Tree

A decision tree (DT) [43] is a logical model represented as a tree structure that shows how the value of a target variable can be predicted by using the values of a set of observation variables (attributes). Each branch node represents a split between a number of alternatives based on a specific attribute, and each leaf node represents a decision. The induction of a decision tree is a top-down process to reduce information content by mapping them to fewer outputs but seek a trade-off between accuracy and simplicity.

Decision trees provide a way to easily understand the derived decision rules and interpret the predicted results and have been used for diagnosis of aortic stenosis [44] and folding mechanism prediction of protein homodimer [45]. One of the disadvantages of DT models is that DT split the training set into smaller and smaller subsets, which makes correct generalization harder and incorrect generalization easier because smaller sets have accidental regularities that don't generalize. Pruning can address this problem to some extent, though.

2. Logistic Regression

Logistic function was first discovered by Peral and Reed [46] in 1920, and logistic regression is a generalized linear model used to calculate the probability of the occurrence of an event by fitting the data to a logit function through maximum likelihood. It is a discriminative counterpart of naive Bayes model as they represent the same set of conditional probability distributions. It has been extensively used for prediction and diagnosis in medicine [47, 48] due to its robustness, flexibility, and ability to handle nonlinear effects. But generally, it requires more data to achieve stable and meaningful results than standard regression.

3. Support Vector Machines

Support vector machines (SVMs) [49] are also linear models that are trained to separate the data points (instances) based on both empirical and structural risk minimum principles; that is, they not only classify objects into categories but construct a hyperplane or set of hyperplanes in a high-dimensional space with a maximum margin among different categories. New instances are then mapped into the same space and classified into a category based on which side of hyperplanes they fall on.

The SVM model has been used for many biomedical tasks, such as microarray data analysis [50], classification [51], information extraction [52], and image segmentation [53]. SVM model can leverage an arbitrary set of features to produce accurate and robust results on a sound theoretical basis, with powerful generalization ability due to optimizing margins. However, from a practical point of view, the most serious problem with the SVM model is the high level of computational complexity and extensive memory requirements for large-scale tasks.

4. K Nearest Neighbor

The k nearest neighbor (k -NN) rule [54] is a type of instance-based learning, or lazy learning, where generalization beyond the training data is delayed. The goal is to assign a new instance a value or category that is averaged (for regression) or voted (for classification) based on examining the k closest labeled training instances.

The K -NN method has been used in gene expression analysis [55], screening data analysis [56], protein-protein interaction [57], biomedical image interpretation [58], etc. The main advantage of this method is that the target function will be approximated locally for each new instance, so that it can deal well with changes in the problem domain. A practical problem is that it tends to be slower especially for large training sets as the entire training set would be traversed for each new instance.

5. Artificial Neural Network

Artificial neural networks (ANNs) [59] are a mathematical model of human intellectual abilities that seek to simulate the structure and functional aspects of biological neural networks. In an ANN model, the artificial neurons (processing units) are connected together via unidirectional signal channels in different layers to mimic the biological neural network. Usually, only neurons in two consecutive layers are connected.

In the biomedical domain, ANNs have been used for many diagnostic [60, 61] and prognostic [62, 63] tasks. Neural networks have the ability to implicitly detect complex nonlinear relationships between dependent and independent variables, as well as possible interactions among predictor variables. On the other hand, ANNs are computationally expensive, prone to over-fitting, and lack a sound theoretical foundation.

6. Conditional Random Fields

Conditional random fields (CRFs) [64] consist of a probabilistic framework for labeling and segmenting structured data, such as sequences, trees, and lattices. The underlying idea is that of defining a conditional probability distribution over label sequences given a particular observation sequence, rather than a joint distribution over both label and observation sequences.

Much like MRF, a CRF is an undirected graphic model, but they have different characteristics. CRFs have better predictive power due to direct modeling on posterior, have flexibility to use features from different aspects, and relax the strong assumption of conditional independence of the observed data. On the other hand, MRFs can handle incomplete data problems and augment small labeled data with larger amounts of cheap unlabeled data. Similarly, the primary advantage of CRFs over hidden Markov models (HMM) [37] is also their conditional nature, resulting in the relaxation of the independence assumptions required by HMMs in order to ensure tractable inference. Additionally, CRFs avoid the label bias problem, a weakness exhibited by maximum entropy Markov models (MEMMs) [65] and other conditional Markov models based on directed graphical models. CRF model is very popular in biomedical entity recognition [66, 67], relation extraction [68], and event detection [69].

Unsupervised Clustering

The learning models discussed above are mostly for supervised learning, which requires labeled data for model training. Clustering is a commonly used unsupervised learning method which automatically discovers the underlying structure or pattern in a collection of unlabeled data. The goal is to partition a set of objects into subsets whose members are similar in some way as well as dissimilar to members from a different subset. Determining how similarity (or dissimilarity) between objects is defined and measured is very crucial for the clustering task. Examples of distance metrics are Mahalanobis, Euclidean, Minkowski, and Jeffreys-Matusita. There are three main types of clustering approaches: partition clustering [70], hierarchical clustering [71], and a mixed model [72].

The most typical example of clustering in bioinformatics is microarray analysis [55, 73–76], where genes with expressional similarities are grouped together, assuming that they have regulatory or functional similarity.

Deep Learning

Deep learning has in recent years become an emerging trend in machine learning. Deep learning refers to “a class of machine learning techniques that exploit many layers of non-linear information processing for supervised or unsupervised feature extraction and transformation, and for pattern analysis and classification” [77]. Compared with aforementioned traditional machine learning approaches, deep

learning learns optimal representations from unlabeled data (i.e., representation learning) and therefore eliminates the needs of feature engineering efforts required for traditional machine learning approaches and has shown a strong learning ability [78].

Deep learning models that learn from unlabeled data include restricted Boltzmann machines (RBMs) [79], deep belief networks (DBNs) [80], and deep autoencoders [81]. Supervised learning models include multilayer perceptron, convolutional neural networks (CNNs) [82], and recurrent neural networks (RNNs) [83]. Deep learning models are typically trained using the backpropagation algorithm. When data in a target domain is limited, which is common in the healthcare domain, a pre-trained model in a large but close domain can be fine-tuned in the target domain [84].

An Overview of Existing Clinical NLP Systems

In electronic health records (EHRs), the central challenge of extracting detailed medical information is dealing with the heterogeneity of clinical data, which involves both structured descriptions and narratives. Over the last two decades, there have been great efforts to develop biomedical NLP systems for clinical narrative text mining. There are mainly two types of approaches that have been explored. Rule-based approaches focus on making use of sublanguage analysis and pattern matching rules, while machine learning-based approaches investigate various useful features and appropriate algorithms. For both approaches, a domain knowledge resource is generally used.

Rule-Based Approach

One of the earliest clinical NLP systems developed, which emerged from the Linguistic String Project [85, 86], used comprehensive syntactic and semantic knowledge rules to extract encoded information from clinical narratives. But systems containing syntactic knowledge are very time-consuming to build and maintain because syntax is so complex.

Later, MedLEE (Medical Language Extraction and Encoding system) system [87] was developed to process clinical information expressed in natural language. This system incorporates a semantically based (simple syntax rules are also included) parser for determining the structure of text. The parser is driven by a grammar that consists of well-defined semantic patterns, their interpretations, and the underlying target structures. By integrating the pattern matching with semantic techniques, the MedLEE system is expected to reduce the ambiguity within the language of domain because of the underlying semantics.

Gold et al. [88] developed a rule-based system called MERKI to extract medication names and the corresponding attributes from structured and narrative clinical texts. Recently, Xu et al. [89] built an automatic medication extraction system (MedEx) on discharge summaries by leveraging semantic rules and a chart parser, achieving promising results for extracting medication and related fields, e.g.,

strength, route, frequency, form, dose, duration, etc. This information was defined by a simple semantic representation model for prescription-type of medication findings, into which medication texts were mapped.

Learning-Based Approach

SymText (Symbolic Text Processor) [90] is a learning-based NLP system which integrates a syntactic parser based on augmented transition networks and transformational grammars with a semantics model based on the Bayesian network [34] statistical formalism, which has been used for various applications such as extracting pneumonia-related findings from chest radiograph reports [91].

Agarwal and Yu developed two biomedical NLP systems named NegScope [92] and HedgeScope [93], which were able to detect negation and hedge cues as well as their scopes in both the biomedical literature and clinical notes. Both systems were built on the conditional random fields (CRFs) [64] learning model trained on the publicly available BioScope [94] corpus.

Lancet [95] is a supervised machine learning system that automatically extracts medication events consisting of medication names and information pertaining to their prescribed use (dosage, mode, frequency, duration, and reason) from clinical discharge summaries. Lancet employs the CRF model [64] for tagging individual medication names and associated fields and the AdaBoost model with decision stump algorithm [96] for determining which medication names and fields belong to a single medication event. During the third i2b2 shared-task for challenges in natural language processing for clinical data, medication extraction challenge, Lancet achieved the highest precision among top ten systems.

In order to help healthcare providers quickly and efficiently answer the questions that arise during their meetings with patients, Cao et al. [97] built a clinical question answering system called AskHERMES, a computational system that automatically analyzes the input clinical questions, retrieves and mines large sets of literature documents and clinical notes pertaining to specific questions, and generates short text summary as the output answer. For question analysis [98], support vector machines (SVMs) learning algorithm [49] and CRF model [64] were used for the question classification and keyword identification, respectively. The system is designed to enable healthcare providers to efficiently seek information in clinical settings.

Liu et al. [99] developed speech recognition system for clinical setting, ClinicalASR, to provide the speech interface to clinical NLP applications for more efficient information access, such as clinical question answering system. ClinicalASR explored language model (LM)-based adaptation on the SRI Decipher system [100] using clinical questions.

The most recent review on clinical information extraction [101] shows that traditional machine learning-based information extraction approaches have been increasingly applied on the EHR data and systems have been developed in various areas of applications: entity extraction [102, 103], adverse drug reaction detection [104, 105], disease classification [106], and clinical event identification [107].

Deep Learning Approach

Word embedding is a popular technique in clinical natural language processing. The dense representation through standard word embedding technique has been successfully applied in medical named entity recognition [108, 109], medical semantics modeling [110], clinical abbreviation disambiguation [111], and expansion [112]. Related to named entity recognition, Henriksson et al. [113] leveraged word embedding for detecting adverse event signals from clinical notes. In another study, Ghassemi et al. [114] employed word embeddings to facilitate extracting clinical sentiment information (positive vs. negative), and they reported a promising correlation between clinical sentiment extraction with the outcome classes.

Instead of using standard word embedding training strategy, Choi et al. [115] presented a novel neural word embedding tool, Med2Vec, which can not only learn distributed representations for both medical codes and visits in electronic health record (EHR) but also allow interpreting the learned representations confirmed positively by clinical experts. Another study [116] exploited neural language modeling to learn low-dimensional embeddings for a wide range of medical concepts, and the similarity and relatedness properties of learned embeddings have been evaluated and compared when learning from different sources such as medical journals, medical claims, and clinical narratives. They also demonstrate how to learn medical concept embeddings in a privacy preserving manner from co-occurrence counts derived from clinical narratives.

Different deep learning architectures have also been explored to identify clinical events and relations from EHR notes and to use longitudinal EHRs for predicting patient outcomes. For example, Jagannatha and Yu [117] employed bidirectional RNN structure to extract adverse drug events from electronic health records, and the GRU RNN model trained on the document level achieved the best performance of 80.31% (exact match), outperforming the baseline conditional random field (CRF) model. Later they combined the CRF-based structured learning with RNN to extract multiple categories of medical entities [118], and the combined model with approximate CRF message passing inference performed best among other system paradigms. To extract relations relevant to adverse drug events, Munkhdalai et al. [119] exploited RNN model combined with the attention mechanism to address the relation classification problem. For predicting assertions (presence and period) of medical events in clinical notes, Li and Yu [120] proposed a hybrid architecture composed of a RNN and deep residual network, showing improved performance compared with conventional baseline systems. Choi et al. [121] applied GRU (gated recurrent units) [122]-based RNNs on longitudinal EHR data to predict future disease diagnosis and medication prescription. Miotto et al. [123] explored a three-layered stack of denoising autoencoders to learn a general-purpose patient representation from EHR data, resulting in improved disease classification accuracy of 92.9%.

Deep learning approaches have shown advantages in multimodal information since they can combine several components with different deep neural network architectures. This becomes especially useful in mining heterogeneous clinical data,

which include textual clinical notes, radiology images, and lab test results, to better support clinical and translational research. For instance, Shin et al. [124] applied an integrated text-image CNN to identify semantic interactions between radiology images and reports. Similarly, Wang et al. [125] proposed a text-image embedding network (TieNet) with multilevel attention mechanisms to learn distinctive image and text representations simultaneously, which is exploited for common thorax disease classification and reporting in chest X-rays.

Challenges and Future Directions

Although remarkable progress has been made for clinical NLP, there are many challenges and open questions to be investigated in the future.

One obstacle to clinical NLP is access to EHRs. In the United States, the Health Insurance Portability and Accountability Act of 1996, or as it is known today as HIPAA, has required that the use of protected health information (PHI) in research studies is not permitted except with the explicit consent of the patient, which prevents gathering data for NLP applications if the data is not de-identified. But HIPAA does allow for the creation of de-identified health information. De-identification tools have been developed, and commercial tools are also available. De-ID [126] information has been used by affiliated hospitals at the University of Pittsburgh, which made available a whole year of EHR data for NLP use. Currently, de-identification tools are still not widely used by hospitals, hampering the NLP applications which are highly based on available EHR data.

Although the sublanguage analysis works well in many sub-domains, it is very time-consuming to compile rules syntactically and semantically and needs a lot of efforts to keep them well maintained, especially as ever-increasing amount of EHR data becomes available. But sublanguage analysis does provide more information that could be helpful in the design of learning-based systems. Therefore, how to effectively and systematically integrate the sublanguage analysis as features into the learning framework and how to employ the learning methods for automatically extracting sublanguage-specific patterns have great potential to facilitate the advancement of EHR-based clinical research informatics.

Currently, most clinical NLP systems are still in an experimental stage rather than deployed and regularly used in clinical setting. The difficulties in translation of clinical NLP research into clinical practice and obstacles in determining the level of practical engagement of NLP systems provide more challenging research opportunities in this field. In addition, to assist clinical decision support, NLP system needs to deal with time series information extraction, reasoning, and integration, for example, linking clinical findings to patient profile, linking different records of same patient, and integrating factual information from multiple sources. However, all those tasks are not trivial in the clinical setting.

Last but not least, effectively mining EHRs for clinical research has the following two challenges.

1. Data quality issues

EHR data hold the promise for secondary use for research and quality improvement; however, such uses remain extremely challenging because EHR data can be inaccurate, incomplete, fragmented, and inconsistent in semantic representations for common concepts. For example, patient data such as glomerular filtration rate (GFR) or body mass index are often unavailable in EHR but are important research variables. In addition, for a study looking for hypertension patients, the determination of hypertension should account for the use of hypertensive drugs, the ICD-9 diagnosis codes for hypertension, or the blood pressure values out of the normal range in certain measurements contexts. Blood pressure values captured in an emergency room are found to be generally elevated compared with the blood pressure values documented during physical exams; therefore, the former value may not represent the patient's real value. Moreover, the saying "absence of evidence is not evidence of absence" is very true for using EHR data. If a clinical research investigator is looking for patients with cardiovascular diseases but cannot find corresponding diagnoses in a patient, the investigator cannot jump to the conclusion that the patient has no cardiovascular disease until further confirmation can be obtained. Typical reasons can be that the patient's medical history is not completely captured by the hospital where the EHR is used or the patient has not been diagnosed. Moreover, much data is not amenable for computer processing, especially those in free-text notes. Whenever it is free-text, there is a challenge for identifying semantic equivalence of multiple linguistic forms of the shared concepts. For example, among hypertensive patients, the medical records can store values such as "HTN," "hypertension," or "401.9" as an ICD-9 code to indicate hypertension.

2. Challenges for converting clinical data to research variables

Many people are still skeptical about reusing clinical data for clinical research because they believe clinical data are "garbage in, garbage out." Although this statement is a little exaggerating, there are dramatic differences between a clinical database and a clinical research database developed following a rigorous clinical research protocol. A research protocol will specify what data will be collected at what time and how. A clinical research database is often designed as a relational database with a tabular format, organized by patient and variables over time. There is a strict quality assurance procedure to ensure the completeness and accuracy of research data. Furthermore, clinical research databases are optimized for statistical analysis. In contrast, a clinical database is organized by clinical events, not by patients. Moreover, clinical data are collected for administrative uses or personal interpretations of medical doctors. Copy and paste as well as creative abbreviations that only doctors themselves can interpret in certain contexts are very common in clinical databases. Therefore, ad hoc extraction of research variables from a clinical database is not a trivial task.

In conclusion, natural language processing (NLP) offers an effective way to unlock disease knowledge from unstructured clinical narratives. Although standards

are emerging and EHR data is becoming better encoded with clinical terminology standards, there will likely always be a narrative aspect (at least for the foreseeable future), which makes clinical NLP technologies indispensable for clinical research informatics. Different approaches and models have been widely applied for biomedical literature, and all those NLP techniques are crucial and can be adapted for effectively mining electric health records (EHRs) to support important clinical research activities. Newly emerged deep learning techniques have brought significant improvements across various tasks and will be increasingly embraced for effectively and efficiently mining big data of EHRs, further advancing disease management, quality improvement, and all aspects of clinical research.

References

1. Sox HC, Greenfield S. Comparative effectiveness research: a report from the Institute of Medicine. Ann Intern Med. 2009;151:203–5.
2. NIH VideoCasting Event Summary. <http://videocast.nih.gov/summary.asp?live=8062>. Accessed 18 May 2011.
3. Clinical Research & Clinical Trials. <http://www.nichd.nih.gov/health/clinicalresearch/>. Accessed 17 May 2011.
4. Sung NS, Crowley WF, Genel M, Salber P, Sandy L, Sherwood LM, et al. Central challenges facing the national clinical research enterprise. JAMA. 2003;289:1278–87.
5. Most physicians do not participate in clinical trials because of lack of opportunity, time, personnel support and resources. <http://www.harrisinteractive.com/news/allnewsbydate.asp?NewsID=811>. Accessed 31 Aug 2010.
6. Clinical and Translational Science Awards. 2007. <http://www.ncrr.nih.gov/clinical%5Fresearch%5Fresources/clinical%5Fand%5Ftranslational%5Fscience%5Fawards/>. Accessed 31 Aug 2010.
7. Garets D, Davis M. Electronic medical records vs. electronic health records: yes, there is a difference. A HIMSS analytics white paper Chicago: HIMSS Analytics. 2005.
8. Garets D, Davis M. Electronic patient records, EMRs and EHRs: concepts as different as apples and oranges at least deserve separate names. Healthcare Informatics online. 2005;22:53–54.
9. File:VistA_Img.png – wikipedia, the free encyclopedia. http://en.wikipedia.org/wiki/File:VistA_Img.png. Accessed 18 Aug 2010.
10. Walker EP. More doctors are using electronic medical records. 2010. <http://www.medpagetoday.com/PracticeManagement/InformationTechnology/17862>. Accessed 18 Aug 2010.
11. Population Estimates. <http://www.census.gov/popest/states/NST-ann-est.html>. Accessed 17 May 2011.
12. Hazlehurst B, Mullooly J, Naleway A, Crane B. Detecting possible vaccination reactions in clinical notes. In: AMIA annual symposium proceedings; 2005. p. 306–10.
13. Pakhomov S, Weston SA, Jacobsen SJ, Chute CG, Meverden R, Roger VL. Electronic medical records for clinical research: application to the identification of heart failure. Am J Manag Care. 2007;13(6 Part 1):281–8.
14. Liao KP, Cai T, Gainer V, Goryachev S, Zeng-treitler Q, Raychaudhuri S, et al. Electronic medical records for discovery research in rheumatoid arthritis. Arthritis Care Res (Hoboken). 2010;62:1120–7.
15. Brownstein JS, Murphy SN, Goldfine AB, Grant RW, Sordo M, Gainer V, et al. Rapid identification of myocardial infarction risk associated with diabetes medications using electronic medical records. Diabetes Care. 2010;33:526–31.
16. Reis BY, Kohane IS, Mandl KD. Longitudinal histories as predictors of future diagnoses of domestic abuse: modelling study. BMJ. 2009;339:b3677.

17. Wang X, Hripcsak G, Markatou M, Friedman C. Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: a feasibility study. *J Am Med Inform Assoc.* 2009;16:328–37.
18. Chute CG. The horizontal and vertical nature of patient phenotype retrieval: new directions for clinical text processing. *Proc AMIA Symp.* 2002:165–9.
19. Thadani SR, Weng C, Bigger JT, Ennever JF, Wajngurt D. Electronic screening improves efficiency in clinical trial recruitment. *J Am Med Inform Assoc.* 2009;16:869–73.
20. Embi PJ, Payne PRO. Clinical research informatics: challenges, opportunities and definition for an emerging domain. *J Am Med Inform Assoc.* 2009;16:316–27.
21. Kuehn BM. Institute of Medicine outlines priorities for comparative effectiveness research. *JAMA.* 2009;302:936–7.
22. Grishman R, Hirschman L, Nhan NT. Discovery procedures for sublanguage selectional patterns: initial experiments. *Comput Linguist.* 1986;12:205–15.
23. Harris Z. Mathematical Structures of Language. New York and London: Interscience Publishers; 1968.
24. Grishman R, Kittredge R. Analyzing language in restricted domains: sublanguage description and processing. Hillsdale, N.J: Lawrence Erlbaum Associates; 1986.
25. Johnson SB, Gottfried M. Sublanguage analysis as a basis for a controlled medical vocabulary. In: Proceedings symposium on computer applications in medical care; 1989. p. 519–23.
26. Bronzino JD. The biomedical engineering handbook. Florida: Springer; 2000.
27. Friedman C, Kra P, Rzhetsky A. Two biomedical sublanguages: a description based on the theories of Zellig Harris. *J Biomed Inform.* 2002;35:222–35.
28. Hastie T, Tibshirani R, Friedman JH. The elements of statistical learning: data mining, inference, and prediction. New York: Springer; 2001.
29. Bishop C. Pattern recognition and machine learning (Information Science and Statistics). Springer; 2007. <http://www.amazon.ca/exec/obidos/redirect?tag=citeulike09-20&path=ASIN/0387310738>. Accessed 15 Jul 2010.
30. Pearl J. Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmann Publishers; 1988. <http://portal.acm.org/citation.cfm?id=534975>. Accessed 12 Jul 2010.
31. Michalski RS, Carbonell JG, Mitchell TM. Machine learning: an artificial intelligence approach. Berlin Heidelberg: Springer-Verlag; 1983.
32. Manning CD, Schütze H. Foundations of statistical natural language processing. Cambridge, MA: MIT Press; 2000.
33. Bayes M, Price M. An essay towards solving a problem in the doctrine of chances. By the Late Rev. Mr. Bayes, F. R. S. Communicated by Mr. Price, in a Letter to John Canton, A. M. F. R. S. Philos Trans (1683–1775). 1763;53:370–418.
34. Pearl J. Bayesian networks: a model of self-activated memory for evidential reasoning. In: Proceedings of the 7th Conference of the Cognitive Science Society, University of California, Irvine. 1985. p. 334, 329. <http://www.amazon.com/Bayesian-networks-self-activated-evidential-University/dp/B00071DFAE>. Accessed 15 Jul 2010.
35. Verduijn M, Peek N, Rosseel PMJ, de Jonge E, De Mol BAJM. Prognostic Bayesian networks: I: rationale, learning procedure, and clinical use. *J Biomed Inform.* 2007;40:609–18.
36. Friedman N, Linial M, Nachman I, Pe'er D. Using Bayesian networks to analyze expression data. *J Comput Biol.* 2000;7:601–20.
37. Baum LE, Petrie T. Statistical inference for probabilistic functions of finite state Markov chains. *Annals Math Stat.* 1966;37:1554–63.
38. Lukashin AV, Borodovsky M. GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res.* 1998;26:1107–15.
39. Yu L, Smith TF. Positional statistical significance in sequence alignment. *J Comput Biol.* 1999;6:253–9.
40. Kindermann R. Markov random fields and their applications (Contemporary Mathematics; V. 1). American Mathematical Society. <http://www.amazon.ca/exec/obidos/redirect?tag=citeulike09-20&path=ASIN/0821850016>. Accessed 16 Jul 2010.

41. Komodakis N, Besbes A, Glocker B, Paragios N. Biomedical image analysis using Markov random fields & efficient linear programming. *Conf Proc IEEE Eng Med Biol Soc.* 2009;2009:6628–31.
42. Lee N, Laine AF, Smith RT. Bayesian transductive Markov random fields for interactive segmentation in retinal disorders. In: World congress on medical physics and biomedical engineering, September 7–12, 2009, Munich. 2009. 227–30. https://doi.org/10.1007/978-3-642-03891-4_61. Accessed 16 Jul 2010.
43. Quinlan JR. Induction of decision trees. *Mach Learn.* 1986;1:81–106.
44. Pavlopoulos S, Stasis A, Loukis E. A decision tree – based method for the differential diagnosis of aortic stenosis from mitral regurgitation using heart sounds. *Biomed Eng Online.* 2004;3:21.
45. Suresh A, Karthikraja V, Lulu S, Kangueane U, Kangueane P. A decision tree model for the prediction of homodimer folding mechanism. *Bioinformation.* 2009;4:197–205.
46. Pearl R, Reed LJ. A further note on the mathematical theory of population growth. *Proc Natl Acad Sci USA.* 1922;8:365–8.
47. Bagley SC, White H, Golomb BA. Logistic regression in the medical literature:: standards for use and reporting, with particular attention to one medical domain. *J Clin Epidemiol.* 2001;54:979–85.
48. Gareen IF, Gatsonis C. Primer on multiple regression models for diagnostic imaging research. *Radiology.* 2003;229:305–10.
49. Vapnik VN. The nature of statistical learning theory. New York: Springer-Verlag; 1995. <http://portal.acm.org/citation.cfm?id=211359>. Accessed 19 Jul 2010
50. Brown MPS, Grundy WN, Lin D, Cristianini N, Sugnet CW, Furey TS, et al. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc Natl Acad Sci USA.* 2000;97:262–7.
51. Polavarapu N, Navathe SB, Ramnarayanan R, Ul Haque A, Sahay S, Liu Y. Investigation into biomedical literature classification using support vector machines. In: Proceedings IEEE computational systems bioinformatics conference; 2005. p. 366–74.
52. Takeuchi K, Collier N. Bio-medical entity extraction using Support Vector Machines. In: Proceedings of the ACL 2003 workshop on Natural language processing in biomedicine – Volume 13. Sapporo: Association for Computational Linguistics; 2003. p. 57–64. <http://portal.acm.org/citation.cfm?id=1118958.1118966>. Accessed 19 Jul 2010.
53. Pan C, Yan X, Zheng C. Hard Margin SVM for biomedical image segmentation. In: Advances in neural networks – ISNN 2005; 2005. p. 754–9. https://doi.org/10.1007/11427469_120. Accessed 19 Jul 2010.
54. Fix E, Hodges JL. Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties. *International Statistical Review / Revue Internationale de Statistique.* 1989;57:238–47.
55. Pan F, Wang B, Hu X, Perrizo W. Comprehensive vertical sample-based KNN/LSVM classification for gene expression analysis. *J Biomed Inform.* 2004;37:240–8.
56. Shammuganadaram V, Maggiora GM, Lajiness MS. Hit-directed nearest-neighbor searching. *J Med Chem.* 2005;48:240–8.
57. Qi Y, Klein-Seetharaman J, Bar-Joseph Z. Random forest similarity for protein-protein interaction prediction from multiple sources. In: Pacific symposium on biocomputing; 2005. p. 531–42.
58. Barbini P, Cevenini G, Massai MR. Nearest-neighbor analysis of spatial point patterns: application to biomedical image interpretation. *Comput Biomed Res.* 1996;29:482–93.
59. McCulloch W, Pitts W. A logical calculus of the ideas immanent in nervous activity. *Bull Math Biol.* 1990;52:99–115.
60. Xue Q, Reddy BRS. Late potential recognition by artificial neural networks. *Biomed Eng, IEEE Trans on.* 1997;44:132–43.
61. Khan J, Wei JS, Ringner M, Saal LH, Ladanyi M, Westermann F, et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat Med.* 2001;7:673–9.

62. Jerez-Aragonés JM, Gómez-Ruiz JA, Ramos-Jiménez G, Muñoz-Pérez J, Alba-Conejo E. A combined neural network and decision trees model for prognosis of breast cancer relapse. *Artif Intell Med.* 2003;27:45–63.
63. Burke HB, Goodman PH, Rosen DB, Henson DE, Weinstein JN, FEH J, et al. Artificial neural networks improve the accuracy of cancer survival prediction. *Cancer.* 1997;79:857–62.
64. Lafferty J, McCallum A, Pereira F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Machine learning-international workshop then conference; 2001. p. 282–9.
65. McCallum A, Freitag D, Pereira FCN. Maximum Entropy Markov models for information extraction and segmentation. In: Proceedings of the seventeenth international conference on machine learning; Morgan Kaufmann Publishers; 2000. p. 591–8. <http://portal.acm.org/citation.cfm?id=658277>. Accessed 20 Jul 2010.
66. Settles B. ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text. *Bioinformatics.* 2005;21:3191–2.
67. Leaman R, Gonzalez G. Banner: an executable survey of advances in biomedical named entity recognition. In: Pacific Symposium on Biocomputing. 2008. p. 652–663.
68. Bundschus M, Dejori M, Stetter M, Tresp V, Kriegel H-P. Extraction of semantic biomedical relations from text using conditional random fields. *BMC Bioinforma.* 2008;9:207.
69. Sarafraz F, Eales J, Mohammadi R, Dickerson J, Robertson D, Nenadic G. Biomedical event detection using rules, conditional random fields and parse tree distances. In: Proceedings of the workshop on BioNLP: shared task. Boulder: Association for Computational Linguistics; 2009. p. 115–8. <http://portal.acm.org/citation.cfm?id=1572340.1572359>. Accessed 21 Jul 2010.
70. Forgy E. Cluster analysis of multivariate data: efficiency vs. interpretability of classifications. *Biometrics.* 1965;21:768.
71. Jardine N, Sibson R. Mathematical taxonomy. Wiley; 1971.
72. McLachlan GJ, Basford KE. Mixture models: inference and applications to clustering. New York, N.Y.: Marcel Dekker; 1988.
73. Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, et al. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc Natl Acad Sci USA.* 1999;96:2907.
74. De Smet F, Mathys J, Marchal K, Thijs G, De Moor B, Moreau Y. Adaptive quality-based clustering of gene expression profiles. *Bioinformatics.* 2002;18:735.
75. Sheng Q, Moreau Y, De Moor B. Biclustering microarray data by Gibbs sampling. *Bioinformatics.* 2003;19(2):196–205.
76. Schafer J, Strimmer K. An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics.* 2005;21:754.
77. Deng L, Yu D. Deep learning: methods and applications. *Found Trends Signal Process.* 2014;7:197–387.
78. Young T, Hazarika D, Poria S, Cambria E. Recent trends in deep learning based natural language processing. arXiv:170802709 [cs]. 2017. <http://arxiv.org/abs/1708.02709>. Accessed 6 Jul 2018.
79. Salakhutdinov R, Mnih A, Hinton G. Restricted Boltzmann machines for collaborative filtering. In: Proceedings of the 24th international conference on machine learning. New York: ACM; 2007. p. 791–8. <https://doi.org/10.1145/1273496.1273596>.
80. Hinton G, Osindero S, Teh Y-W. A fast learning algorithm for deep belief nets. *Neural Comput.* 2006;18:1527–54.
81. Vincent P, Larochelle H, Lajoie I, Bengio Y, Manzagol P-A. Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion. *J Mach Learn Res.* 2010;11:3371–408.
82. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. 2012. p. 1097–1105. <http://papers.nips.cc/paper/4824-imagenet>. Accessed 9 Feb 2015.
83. Socher R, Lin CC, Manning C, Ng AY. Parsing natural scenes and natural language with recursive neural networks. In: Proceedings of the 28th international conference on machine learning (ICML-11). 2011. p. 129–136. http://machinelearning.wustl.edu/mlpapers/paper_files/ICML2011Socher_125.pdf. Accessed 9 Feb 2015.

84. Iftene M, Liu Q, Wang Y. Very high resolution images classification by fine tuning deep convolutional neural networks. In: Eighth International Conference on Digital Image Processing (ICDIP 2016). International Society for Optics and Photonics; 2016. p. 100332D. <https://doi.org/10.1117/12.2244339>.
85. Sager N, Friedman C, Chi E. The analysis and processing of clinical narrative. *Fortschr Med.* 1986;86:1101–5.
86. Sager N, Friedman C, Lyman MS. Medical language processing: computer management of narrative data. First Edition. Addison-Wesley; 1987.
87. Friedman C, Alderson PO, Austin JH, Cimino JJ, Johnson SB. A general natural-language text processor for clinical radiology. *J Am Med Inform Assoc.* 1994;1:161–74.
88. Gold S, Elhadad N, Zhu X, Cimino JJ, Hripcsak G. Extracting structured medication event information from discharge summaries. *AMIA Annu Symp Proc.* 2008;2008:237–41.
89. Xu H, Stenner SP, Doan S, Johnson KB, Waitman LR, Denny JC. MedEx: a medication information extraction system for clinical narratives. *J Am Med Inform Assoc.* 2010;17:19–24.
90. Haug PJ, Koehler S, Lau LM, Wang P, Rocha R, Huff SM. Experience with a mixed semantic/syntactic parser. In: Proceedings of the annual symposium on computer application in medical care; 1995. p. 284–8.
91. Fiszman M, Chapman WW, Aronsky D, Evans RS, Haug PJ. Automatic detection of acute bacterial pneumonia from chest X-ray reports. *J Am Med Inform Assoc.* 2000;7:593–604.
92. Agarwal S, Yu H. Biomedical negation scope detection with conditional random fields. *J Am Med Inform Assoc.* 2010;17:696–701.
93. Agarwal S, Yu H. Detecting hedge cues and their scope in biomedical literature with conditional random fields. *J Biomed Inform.* 2010;43(6):953–61. <https://doi.org/10.1016/j.jbi.2010.08.003>.
94. Vincze V, Szarvas G, Farkas R, Mora G, Csirik J. The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinforma.* 2008;9(11):S9.
95. Li Z, Liu F, Antieau L, Cao Y, Yu H. Lancet: a high precision medication event extraction system for clinical text. *J Am Med Inform Assoc.* 2010;17:563–7.
96. Rennie J. Boosting with decision stumps and binary features. *Relation.* 2003;10 1.33: 1666.
97. Cao Y, Liu F, Simpson P, Antieau L, Bennett A, Cimino JJ, et al. AskHERMES: an online question answering system for complex clinical questions. *J Biomed Inform.* 2011;44:277–88.
98. Cao Y, Cimino JJ, Ely J, Yu H. Automatically extracting information needs from complex clinical questions. *J Biomed Inform.* In Press, Uncorrected Proof. <https://doi.org/10.1016/j.jbi.2010.07.007>.
99. Liu F, Tur G, Hakkani-Tür D, Yu H. Towards spoken clinical question answering: evaluating and adapting automatic speech recognition systems for spoken clinical questions. *J Am Med Inform Assoc.* 2011;18:625–30.
100. Stolcke A, Anguera X, Boakyé K, Çetin Ö, A Janin Mandal A, et al. Further progress in meeting recognition: the ICSI-SRI spring 2005 speech-to-text evaluation system. 3869, LNCS, MLMI Workshop. 2005;78:463–75.
101. Wang Y, Wang L, Rastegar-Mojarad M, Moon S, Shen F, Afzal N, et al. Clinical information extraction applications: a literature review. *J Biomed Inform.* 2018;77:34–49.
102. Roberts K, Rink B, Harabagiu SM, Scheuermann RH, Toomay S, Browning T, et al. A machine learning approach for identifying anatomical locations of actionable findings in radiology reports. *AMIA Annu Symp Proc.* 2012;2012:779–88.
103. Li Q, Spooner SA, Kaiser M, Lingren N, Robbins J, Lingren T, et al. An end-to-end hybrid algorithm for automated medication discrepancy detection. *BMC Med Inform Decis Mak.* 2015;15 <https://doi.org/10.1186/s12911-015-0160-8>.
104. Sarker A, Gonzalez G. Portable automatic text classification for adverse drug reaction detection via multi-corpus training. *J Biomed Inform.* 2015;53:196–207.
105. Rochefort CM, Buckeridge DL, Forster AJ. Accuracy of using automated methods for detecting adverse events from electronic health record data: a research protocol. *Implement Sci.* 2015;10:5.
106. Yadav K, Sarioglu E, Smith M, Choi H-A. Automated outcome classification of emergency department computed tomography imaging reports. *Acad Emerg Med.* 2013;20:848–54.

107. Barrett N, Weber-Jahnke JH, Thai V. Engineering natural language processing solutions for structured information from clinical text: extracting sentinel events from palliative care consult letters. *Stud Health Technol Inform.* 2013;192:594–8.
108. Tang B, Cao H, Wang X, Chen Q, Xu H. Evaluating word representation features in biomedical named entity recognition tasks. *Biomed Res Int.* 2014;2014(240403):1–6.
109. Liu S, Tang B, Chen Q, Wang X. Effects of semantic features on machine learning-based drug name recognition systems: word embeddings vs. manually constructed dictionaries. *Information.* 2015;6:848–65.
110. De Vine L, Zucco G, Koopman B, Sitbon L, Bruza P. Medical semantic similarity with a neural language model. In: Proceedings of the 23rd ACM international conference on conference on information and knowledge management. ACM; 2014. p. 1819–1822. <http://dl.acm.org/citation.cfm?id=2661974>. Accessed 4 Jun 2016.
111. Wu Y, Xu J, Zhang Y, Xu H. Clinical abbreviation disambiguation using neural word embeddings. In: Proceedings of the 2015 workshop on biomedical natural language processing; 2015. p. 171–6.
112. Liu Y, Ge T, Mathews KS, Ji H, McGuinness DL. Exploiting task-oriented resources to learn word embeddings for clinical abbreviation expansion. In: Proceedings of the 2015 workshop on biomedical natural language processing; 2015. p. 92–7.
113. Henriksson A, Kvist M, Dalianis H, Duneld M. Identifying adverse drug event information in clinical notes with distributional semantic representations of context. *J Biomed Inform.* 2015;57:333–49.
114. Ghassemi MM, Mark RG, Nemati S. A visualization of evolving clinical sentiment using vector representations of clinical notes. In: 2015 Computing in cardiology conference (CinC). 2015. p. 629–32.
115. Choi E, Bahadori MT, Searles E, Coffey C, Sun J. Multi-layer representation learning for medical concepts. In: Proceedings of 22nd ACM SIGKDD conference on knowledge discovery and data mining. 2016. <http://arxiv.org/abs/1602.05568>. Accessed 10 Mar 2016.
116. Choi Y, Chiu CY-I, Sontag D. Learning low-dimensional representations of medical concepts. *AMIA Jt Summits Transl Sci Proc.* 2016;2016:41–50.
117. Jagannatha A, Yu H. Bidirectional RNN for medical event detection in electronic health records. San Diego; 2016. p. 473–82. <https://www.aclweb.org/anthology/N/N16/N16-1056.pdf>.
118. Jagannatha A, Yu H. Structured prediction models for RNN based sequence labeling in clinical text. 2016. <https://arxiv.org/abs/1608.00612>. Accessed 28 Aug 2016.
119. Munkhdalai T, Liu F, Yu H. Clinical relation extraction toward drug safety surveillance using electronic health record narratives: classical learning versus deep learning. *JMIR Public Health Surveill.* 2018;4:e29.
120. Li R, Yu H. A hybrid neural network model for joint prediction of medical presence and period assertions in clinical notes. In: AMIA fall symposium. 2017.
121. Choi E, Bahadori MT, Sun J. Doctor AI. Predicting clinical events via recurrent neural networks. arXiv:151105942 [cs]. 2015. <http://arxiv.org/abs/1511.05942>. Accessed 9 Mar 2016.
122. Cho K, van Merriënboer B, Gulcehre C, Bougares F, Schwenk H, Bengio Y. Learning phrase representations using rnn encoder-decoder for statistical machine translation. arXiv preprint arXiv:14061078. 2014.
123. Miotto R, Li L, Kidd BA, Dudley JT. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Sci Rep.* 2016;6:26094. <https://doi.org/10.1038/srep26094>.
124. Shin H-C, Lu L, Kim L, Seff A, Yao J, Summers RM. Interleaved text/image deep mining on a large-scale radiology database. In: 2015 IEEE conference on computer vision and pattern recognition (CVPR). 2015. p. 1090–9.
125. Wang X, Peng Y, Lu L, Lu Z, Summers RM. Tienet: text-image embedding network for common thorax disease classification and reporting in chest x-rays. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2018. p. 9049–58.
126. Gupta D, Saul M, Gilbertson J. Evaluation of a deidentification (De-Id) software engine to share pathology reports and clinical documents for research. *Am J Clin Pathol.* 2004;121:176–86.



Data Sharing and Reuse of Health Data for Research

18

Rebecca Daniels Kush and Amy Harris Nordo

Abstract

Facilitating the reuse and sharing of electronic health data for research is an important foundation for reengineering and streamlining research processes and will be critical to accelerating learning health cycles and broadening the knowledge that can be used to improve healthcare and patient health outcomes. In this chapter, data sharing refers to sharing data between partners and systems (not necessarily sharing of research results) in ways that preserve the meaning and integrity of the data. A range of ethical, legal, and technical considerations have thus far hindered the development and application of approaches for such reuse and data sharing, in general. However, standards adoption and technical capabilities are progressing, and incentives are now beginning to align to facilitate data sharing. Principles and values of data sharing and the responsible use of data and data standards have been published, and there is recognition of the value of “real-world data” (RWD) to generate additional evidence upon which to base clinical decisions. These will require broad adoption, adherence, communication, and collective support to positively transform research processes and informatics.

Participants in clinical research studies typically expect and want their data to be shared widely and appropriately such that we can all learn. Based on learning from research results, it is expected that patient care will be improved. This is the basis for learning health systems (LHS), in which research is clearly a vital

R. D. Kush, PhD (✉)
Catalysis, Austin, TX, USA

Elligo Health Research, Austin, TX, USA

Translational Research Informatics Center/FBRI, Kobe, Japan
e-mail: rkush@catalysisresearch.com

A. H. Nordo, MMCI, RN, CPHQ, LNC
Pfizer, Inc., Groton, MA, USA
e-mail: amy.nordo@pfizer.com

component. The knowledge gained from sharing the results of research can inform healthcare and clinical decisions to complete the learning cycle.

This chapter will describe the benefits and implementation considerations of reusing health data, particularly that from electronic health records (EHR), for clinical research, bio-surveillance, pharmacovigilance, outcome assessments, public health, quality reporting, and other research-related studies. Use cases are provided to illustrate the positive impact that data reuse and sharing will have for patients, clinicians, research sponsors, regulatory agencies, insurers, and all involved in LHS. Consensus-based principles for data sharing, technical aspects, and business requirements are also provided, along with specific examples of data sharing collaborations, initiatives, and tools. In the future, we hope that research will become embedded within health systems and that organizations will continue to embrace, harmonize, and broadly adopt standards and technologies to meet this challenge.

Keywords

Reuse · Secondary use · Real-world data · Real-world evidence · Learning health system · Clinical research · Interoperability · Data standards · Electronic health records · Translational science · eSource · FHIR

Introduction

The notion of reusing health data and sharing data, in general, has many different connotations, from providing a pathway to open science and minimizing duplication of efforts to breaching privacy and jeopardizing trust. Participants in clinical research studies typically expect and want their data to be shared widely and appropriately for the greater good. However, they want to give their informed consent and not have their contributions to research abused. As technological advances encourage exponential growth in the amount of data produced, data has been referenced as “the world’s most valuable resource” [1], and “owning” data has been equated to power. Conversely, high-quality clinical research data may be scarce, emphasizing the importance of achieving the greatest and best use for each piece of data provided by those who shared their time, energy, and often their blood and tissue samples; their data are precious.

There is a clear tension between assimilating vast amounts of data to identify “signals” or trends to inform public health awareness and action, and maximizing the value of each data point donated by patients in hopes of finding a cure for a specific condition. In practice, as this chapter will illustrate, the principles, standards, and technical approaches for reusing clinical data while preserving the meaning and integrity of the data are relevant to both these and other use cases. Valuable learning health systems will be accelerated and patient care improved when healthcare data can be more readily leveraged for research. The question is not why but how electronic health data should be reused in order to best honor the sacrifices that patients make to support research.

Best practices for the reuse of data and data sharing include appropriate planning before a research study is initiated, the application of appropriate standards, and following a process that minimizes transcription or redundant entry of data.

These best practices can decrease the time and resources necessary to reuse electronic health data for research, thus streamlining the research process thereby improving learning and clinical decisions, i.e., knowledge transfer of research to patient care [2].

Relevant Concepts and Terms

- *eSource Data (Electronic Source Data)*

eSource data is source data captured initially into a permanent electronic record (eSource document) used for the reconstruction and evaluation of a clinical study or a source data item included in an eCRF when direct entry is made. NOTE: permanent in the context of these definitions implies that any changes made to the electronic data are recorded via an audit trail. See also eSource document, source data, permanent data, data originator. (From body of FDA Final Guidance on eSource) [3].

- *Traceability*

“Traceability is the documenting of work and data flows for each element, from the point of origin to analysis of the data set, and has long been required in regulated research” [4]. Traceability (i.e., provenance) is a very important concept in clinical research, especially regulated research. If any changes are made on the path between source data and reporting such as in regulatory submissions, the changes must be appropriately identified along with the individual making the change, the date, and the reason for the change, i.e., ensuring an audit trail. The acronym ALCOA (attributable, legible, contemporaneous, original, accurate) has been used by the US Food and Drug Administration (FDA) to describe good documentation practices for source data.

- *Interoperability and Semantic Interoperability*

Interoperability is defined by the Healthcare Information Management Systems Society (HIMSS) as “the ability of different information technology systems and software applications to communicate, exchange data, and use the information that has been exchanged” [5]. Interoperability becomes more feasible, accurate, and meaningful through standards implementation. “Semantic interoperability” refers not only to the exchange of information but also the exchange of *meaning* such that the recipient of the information can readily understand and interpret the information accurately in the manner intended by the data generator and/or sender. Interpretation of the meaning of exchanged information frequently relies on having adequate metadata (i.e., data about the data) to interpret the meaning.

- *Data Standards*

According to the Clinical Data Interchange Standards Consortium (CDISC) Glossary [3], a *standard* refers to a criterion or specification established by authority or consensus for specifying conventions that support interchange of

common materials and information. Individual standards can be complementary, for example, when a transport standard “carries” a content standard, or multiple standards organizations can collaborate on a single comprehensive standard [6].

- *BRIDG*

The Biomedical Research Integrated Domain Group (BRIDG) Model, developed by a stakeholder group consisting of NIH/NCI, CDISC, FDA, and HL7, resulted in a single model to ‘bridge’ research and healthcare in addition to standards organizations <https://bridgmodel.nci.nih.gov/>. The scope of the BRIDG model is ‘protocol-driven research’. BRIDG is a single standard vetted through CDISC, HL7, and ISO standards organizations [7].

- *Electronic Case Report Form (eCRF)*

An electronic case report form (eCRF) is an electronic version of a case report form used to collect and store data elements for a specific clinical research study.

- *Secondary Use of Data*

Secondary use of data refers to the reuse of data collected for clinical purposes for additional uses other than direct delivery of healthcare. These may include all types of research, public health, safety surveillance, outbreak reporting, registries, quality measures, or even marketing or business uses.

- *De-identification, Anonymization, and Pseudonymization*

De-identification: as a result of HIPAA, the US Health and Human Services guidance states “the process by which identifiers are removed from the health information, mitigates privacy risks to individuals and thereby supports the secondary use of data for comparative effectiveness studies, policy assessment, life sciences research, and other endeavors” [8].

Anonymization: to remove identifying information from (something, such as computer data) so that the original source cannot be known; to make (something) anonymous [9].

Pseudonymization: to use a fictitious name for the research subject. In terms of data sharing, this can refer to the fact that a research participant is frequently given a “patient/subject number.” The sponsor of the research study receives the number associated with the research data, but only the clinicians treating that patient can link the research number back to the actual patient’s name and identifying information [9].

Benefits of Data Sharing and Reusing Health Data for Research

The benefits of data sharing often outweigh the risks, especially when the data sources are acknowledged and understood, informed consent is properly executed, the uses are valid, the sharing methodology accurately retains the meaning and integrity of the data, and the results are interpreted appropriately. There are several

very important applications and use cases for sharing clinical data for research. Table 18.1 provides examples of overarching benefits of data sharing and reuse in the area of clinical research.

Table 18.1 Benefits of reusing clinical data for research

Eliminating redundancy	Typically, EHR data is manually abstracted and then entered again into electronic clinical research forms (eCRF), registries, clinical trial management systems, and other places. The collection of data for research by the clinician simultaneously at the point of direct care and the use of interoperability standards to aid in the auto-population of data from the EHR to clinical research databases can eliminate redundant data entry of the same information. Redundant data entry adds burden to academic medical institutions that participate in more than one registry or research study. For any clinician who wishes to conduct research studies, this type of administrative burden is a primary deterrent.
Improving data quality	“Swivel chair” interoperability, a term coined by Landen Bain [10], refers to the transcription of data from the EHR to a registry, electronic data capture system (eCRF), or other database for clinical research. This is another example of redundancy, which provokes errors due to transcription. The reality of transcription errors has added to the need for data queries to confirm the quality of the data. Conversely, the use of interoperability standards to auto-populate clinical data into clinical research databases or eCRF can eliminate data transcription errors and improve quality.
Realizing learning health systems	A learning health system was defined, by the Institute of Medicine of the National Academy of Sciences (Institute of Medicine 2015), as a system in which “science, informatics, incentives, and culture are aligned for continuous improvement and innovation, with best practices seamlessly embedded in the delivery process and new knowledge captured as an integral by-product of the delivery experience.” [17] The reuse and analysis of data collected by healthcare systems for research purposes is the basis of generating new knowledge on successful healthcare delivery improvements and innovation.
Improving research through real-world evidence	The use of what is now referred to as “real-world data” (RWD) can inform clinical study designs such that they more closely reflect true patient care practices, potentially reduce the number of placebo patients necessary for a clinical research study, and augment the results of a clinical research study with information that comes from healthcare practices in addition to that collected specifically to follow a clinical research protocol. RWD and EHR data are also useful in assessing research protocol feasibility and identifying potential patients who meet eligibility requirements.
Informing patient choices	If and when healthcare data can be shared broadly in a manner that is comparable and readily understandable, patients can have the opportunity for choices that make sense to them, based upon their lifestyle, other complications, and prior patient experiences. Per Joseph H. Kanter, when faced with a life-threatening diagnosis, a simple question every patient must ask: “For each treatment option available, what are the survival or success rates—or alternatives—for someone like me?” [11]. Physicians no doubt wish to have such information at their fingertips as well.
Realizing personalized or precision medicine	Personalized medicine or precision medicine is “an emerging approach for disease treatment and prevention that takes into account individual variability in genes, environment, and lifestyle for each person” [12]. Turning this type of medicine into a broadly practiced reality relies on sharing of data, especially genomics information.

Requirements for the Use of eSource for Regulated Research

Regulators have been encouraging electronic data collection since 1997, authoring the 21CFR11 regulation, and the use of eSource technologies since the advent of eDiaries and mobile technologies. One FDA-encouraged initiative, which took place from 2004 to 2006, was called eSource Data Interchange Initiative. The product of this collaborative work is a document [13] that includes 12 requirements to follow to ensure that eSource implementations, such as the use of EHRs for research purposes, would adhere to regulations around the globe. These requirements were adopted by European Medicines Agency (EMA) in their guidance for field auditors [14] and encouraged the development of Retrieve Form for Data Capture (RFD) the eSource Data Interchange (eSDI).

These 12 requirements (Box 18.1) for eSource Data Interchange can be followed to ensure global regulatory compliance for different and innovating processes implementing eSource [15].

Box 18.1. Requirements for eSource Data Interchange

1. An instrument used to capture source data shall ensure that the data are captured as specified within the protocol.
2. Source data shall be accurate, legible, contemporaneous, original, attributable (ALCOA), and complete and consistent.
3. An audit trail shall be maintained as part of the source documents for the original creation and subsequent modification of all source data.
4. The storage of source documents shall provide for their ready retrieval.
5. The investigator shall maintain the original source document or a certified copy.
6. The sponsor shall not have exclusive control of a source document.
7. Source data shall only be modified with the knowledge or approval of the investigator.
8. Source documents and data shall be protected from destruction.
9. The source document shall allow for accurate copies to be made.
10. Source documents shall be protected against unauthorized access.
11. The location of source documents and the associated source data shall be clearly identified at all points within the capture process.
12. When source data are copied, the process used shall ensure that the copy is an exact copy preserving all of the data and metadata of the original.

Technical Considerations for Reuse of Health Data for Research

Ideally, the next stage for eSource is to enable a more direct electronic link from the EHR to the eCRF or other research data collection tools in order to make EHR data available more efficiently for clinical research. RFD is one methodology for enabling direct reuse of data from EHRs for research purposes. This integration

profile was developed jointly by CDISC and IHE and subsequently referenced through the Healthcare Information Technology (IT) Standards Panel (HITSP) in work done with the American National Standards Institute (ANSI) in an interoperability specification, along with CDISC Clinical Data Acquisition Standards Harmonization (CDASH) and Continuity of Care Document (CCD) [16, 17]. A proof-of-concept project, called STARBRITE [18], was conducted to analyze how to support research while maintaining natural clinic workflows. The STARBRITE project was based upon the eSDI requirements and demonstrated the feasibility of collecting health and research data simultaneously as a ‘single source’ without redundant data entry, paving the way for the development of the CDISC/IHE RFD profile.

RFD allows for secure interoperability between systems by providing an i-frame or “window” into the EHR so that the eCRF or other data collection form can (a) auto-populate using previously mapped EHR data elements and then (b) be surfaced within the EHR allowing the end user to toggle between sections of the EHR to manually complete fields in the eCRF that are not auto-populated. The location of launching RFD interoperability is customizable, allowing for flexibility in the design to fit within the clinician or researcher’s workflow and allow the opportunity for “concurrent data collection.” Data collected into the eCRF utilizing RFD is posted into the study database and not the EHR itself.

The success of RFD has been based on the use of a CDISC data collection standard called CDASH and the availability of patient data contained within the EHR being made available in a standard format. Initially in the USA, this has been through the Continuity of Care Document (CCD). In Japan, RFD has been implemented using Storage Standard for Medical Information Exchange (SS-MIX) [19]. Unfortunately, information in a CCD document has not proven to be an ideal solution for obtaining standard data from EHRs since implementations can vary, leading to different institutions having multiple CCDs and little harmonization across institutions in this regard [20]. The CCD was designed to share data relevant to a patient’s care across healthcare institutions, while research only requires *specific* data elements (specified in the research protocol). Therefore, redaction of certain data from the CCD is often necessary for RFD methodology to be a sound alternative for research. In particular, data collected for research must be de-identified, and clinicians and study sponsors must be vigilant about the protection of patient data. Currently in the USA, the RFD standard is under evaluation, and IHE has pivoted to the new mRFD (mobile RFD standard) that allows for use other than the CCD for data elements. In Europe, RFD was implemented in the TRANSFoRm project, which leveraged the BRIDG model and a specific ontology [21].

More recently, a disruptive innovation called Fast Healthcare Interoperability Resources (FHIR) has been gaining popularity. This standard was developed by Graham Grieve and adopted by Health Level Seven (HL7), acknowledging that HL7 V2, V3, and CDA were competing standards and that a fresh approach to healthcare standards was necessary [22]. FHIR® supports interoperability for approximately 80% of the data available within the EHR, either relying individually on common “resources” or through the creation of FHIR profiles. Although progress is being made, “research resources” are still in development and are not yet

harmonized with research standards and terminology required by regulators for data submitted in support of new drug approvals.

The availability of FHIR has encouraged additional support for leveraging EHR as eSource for research, but few comparative analysis research studies quantifying the impact of eSource have been published. In 2016, Duke University conducted an industrial grade comparative analysis pilot study on the current manual data collection process and a RFD-enabled eSource solution. The study evaluated an eSource solution with limited data auto-population (~2% of data fields in CRF) for flexibility, time, and data quality [23]. The results from the Nordo et al. study (2016) showed that this methodology of data transfer did not allow access to any data other than the data elements approved for the study and also that the native functionality of the eCRF was maintained. Further, the evaluation showed a 37% decrease in time for data collection using the eSource methodology and a decrease from 9% error rate on critical data elements (e.g., patient identifier numbers) for the manual data collection to 0% error rate for the eSource process. This early evaluation demonstrates the value of this eSource to efficiency of research, providing motivation for further development. The product has subsequently shifted toward utilizing FHIR and is part of a multi-stakeholder collaboration including regulatory agencies, academic medical centers, standards organizations, study sponsors, and vendors.

Several successful data sharing projects and networks are worthy of discussion. One success is the above referenced Duke University product. Additionally, University of California in collaboration with the FDA developed of a robust data collection tool based on the RFD standard across all their hospitals for breast cancer research [24]. The SS-MIX project in Japan, the TRANSFoRm project in Europe, and the IMI EHR4CR project are other examples. Commercial or vendor-specific products have been developed, but in order to ensure success of eSource approaches to reuse of EHR data for research, the industry needs a collaboratively developed end-to-end, open-source-spirited product that is based on standards (and thereby agnostic to the software system) and adopted globally. Broad adoption of common standards and semantics across EHR vendors would also be extremely beneficial toward realizing LHS. Certain standards for research have been developed and adopted globally and are now mandated by the U.S. FDA and Japan's Pharmaceuticals and Medical Devices Agency (PMDA). Unfortunately, the same cannot be said of standards and semantics for electronic health records, which are frequently customized by implementation. According to the first executive director of IMI, "In an era of increased transparency and integrative analyses of data from multiple origins, data standards are essential to ensure accuracy, reproducibility, and scientific integrity" [25].

Electronic population of data into the eCRF is one use case intended to streamline data collection in clinical research. EHR data is being reused for various purposes in addition to the completion of eCRF, including (but not limited to) safety surveillance (Sentinel), outcomes research (OMOP/OHDSI, PCORNet, i2b2), patient identification, and protocol feasibility (EHR4CR/i~HD) (see appendix table for additional information on these use cases and initiatives). These use cases are often implemented by using queries that generate information (data or aggregated "counts") from numerous EHRs, thus requiring each institution to provide the responses to the queries in the format requested by the network. Aggravating the problem of non-standard EHRs, the "common data models" (CDM) differ across these networks, thereby increasing the resources

needed for organizations to participate in multiple networks and subsequently decreasing their capacity to share data among them for research and other purposes. This has spawned a CDM Harmonization project, which is being funded by the Patient Centered Outcomes Research (PCOR) Trust Fund and led by FDA, to develop a semantic interoperability methodology using the BRIDG standard information model as an “intermediary” such that data received in any of the CDMs could be provided in a standard format for FDA or others to use. Results of this project will be published in the future.

General Considerations for Implementing eSource

The realization of the EHR data reuse that is needed to support LHS and the research community as a whole has been hindered by a culture of misaligned incentives and lack of trust, in addition to the technical and operational aspects of sharing data within and across different organizations. Most healthcare organizations have focused on using data for billing purposes and quality improvement within the organization, while reuse of the data, especially outside of the organization, for research typically has been a lower priority [26, 27]. Customization of EHRs to optimize a purely internal process can actually impede data sharing across centers. Critics of interoperability also express concerns on the quality of the EHR data for clinical research [28]. This disconnect among stakeholders as to the value of the data and the fear of change or misuse can make the process of using clinical data for research more difficult.

Holders of clinical data recognize the financial value of these data, and there is the fear of others using that data for nefarious purposes, among those being that research may benefit competitors or that so-called “rogue” analyses may produce inaccurate results. Institutions are vigilant about information security and therefore develop complex processes to access or share the data with detailed data sharing and data usage agreements for each individual use case.

The ethics around data sharing of patient data collected as part of the EHR are complicated and actually pose one of the key hurdles to overcome. Although individuals are generally willing to share their data for the “greater good,” situations have arisen where data have been used for purposes the “data donors” were not apprised of and in some cases have found offensive. Such abuse has resulted in the need for informed consent to include the reuse of EHR data, data use agreements, and data sharing agreements, which can take months or even years to execute.

The New York Times article “Where’d you go with my DNA?” illustrates the dangers of reusing data for different types of research without planning for impact on patients [29]. The Havasupai Indians in Arizona thought they were giving their blood for research on diabetes, a disease that affects many in their tribe, but later learned that their data was also used to study diseases that would stigmatize their tribe. Other such stories related to reusing patient materials without process to track have emerged, including the famous *The Immortal Life of Henrietta Lacks* [30]. One can easily imagine how a LHS conducting many studies, and mingling EHR and research data, could be vulnerable to a later investigator inadvertently using data collected under consent for one purpose to answer a different research question which could lead to harm. This issue is of concern to patients, and organizations need to develop safeguards and procedures to protect their patients that participate in research.

In addition to patient consent and data use agreements to address ownership and use of data, there are regulations, guidance documents, and “binding guidance” (i.e., requirements) published by regulatory authorities that address various aspects of data sharing [31]. Specifically, the FDA, EMA, China Food and Drug Administration (CFDA), and PMDA have published requirements around traceability and provenance of data that comes from research sites and is submitted to them for review when they approve new therapies. Understanding the regulations, guidance, and binding guidance from regulators in reference to the reuse of EHR data includes redefining long-held beliefs of roles and responsibilities of data stewardship. To fully implement eSource, industry research sponsors and partners will also have a learning curve to scale, as many data managers may lack awareness in the amount and quality of data available in the EHR and the percentage of EHR data used as source in the studies they manage. Data completeness is an important concern relevant to the use of eSource and EHR data for clinical research. There should be no expectation that the EHR can provide *all* of the necessary data elements for research; it is reasonable to expect that research will require data elements that are not in the EHR in order to answer unique and cutting edge research questions. It is for this reason that RFD and other techniques allow for entry of protocol-specific data.

In addition to education about the regulatory issues, other aspects of eSource must be considered for successful implementation. Semantics play a bigger role than is often realized until it is too late. Semantic variability around the data and the metadata can make it nearly impossible to interpret certain results if they are not collected in a standard way. Clearly defined data elements are required for research, and researchers and clinicians must make every effort to compare, contrast, and harmonize definitions of terms used in different health systems or research studies in order to conduct high quality research. Operationally, these definitions can vary across organizations. For example, the definition of a data element for “smoking status” could mean never smoked, smoking cessation within a defined time range, or some other definition. The definitions of terms used in research are of paramount importance, thus the need for clearly defined data definitions, controlled terminologies, and ontologies. Assumptions and misinterpretations by researchers of the definitions for data from EHR systems can impact the results of the study. In particular, the representational inadequacy (or the degree to which a data element differs from the desired concept [32]) of EHR data is a reasonable concern that can be mitigated by harmonization of data elements across sites. Only when definitions are harmonized will there be data integrity and semantic interoperability, i.e. exchange of data while preserving the meaning of that data.

Data quality, both perceived and real, is perhaps the greatest challenge for eSource to be widely used and trusted. The completeness and availability of historical patient data vary by organization and can impact the quality and completeness of data for research. Some organizations moved data from legacy systems into new EHR systems for all patient records, some moved only the data related to patients’ current treatment plans, and some used a date to differentiate the data contained in each system. The variability of legacy data contained within the current EHR systems is compounded by the flexible definitions and locations for storing the data

(e.g., ejection fraction data can be found in multiple places within an EHR, including the cath report, the ECHO report, and the flow sheet). The quality of these data to support research is also impacted by how the data from various sources are compiled (i.e., multiple EHRs at different institutions that care for the same patient or data shared in other formats), especially since the records of research subjects are de-identified, anonymized, or pseudonymized. Various methods proposed to deal with this data compilation while respecting confidentiality and patient privacy include the use of a unique patient identifier, the un-blinding of an individual to link disparate data for a patient, or block chain or other algorithm that allows patients' data to be appropriately matched. A consensus on the appropriate methodology has not yet been reached. Continued work to improve the semantics, completeness, representational adequacy, and compiling of data from multiple sources will greatly advance the use of EHR data for research.

The considerations and issues mentioned above have impacted further development and adoption of EHR-based eSource by research sponsors. There is understandable hesitation to invest in eSource systems and methodologies that do not meet FDA requirements and would render meaningless the research using those systems and data. This fear is compounded by a lack of clear understanding across the industry about regulatory requirements and expectations. Varying sources of information and guidelines from multiple different regulatory entities (e.g., FDA, EMA, ONC, IMI, AHRQ, NIH) create an appearance of lack of alignment, which leads to more confusion. Although the FDA is actively encouraging the use of new technologies, including eSource from mobile devices and wearables, and “real-world data” [33] for research, the “fear of regulatory repercussions” might slow the adoption of eSource. Clearly, in addition to the aforementioned considerations, communication, dissemination of results from eSource methodology assessments, education, and alignment will be needed for widespread adoption.

Best Practices and Methods of Data Sharing for Research

Broader adoption of eSource, reuse of EHR data, and data sharing will require the collaboration of all stakeholder groups. Whether data is shared among researchers, within a LHS, or externally, there are certain best practices and methods that apply. These include planning, implementing standards, and streamlining processes from beginning to end.

Planning

The importance of planning in accordance with well-documented principles for quality improvement, such as the Deming Wheel [34], is often underestimated. Failure to plan up front how data will be shared inevitably results in higher costs, longer timelines, and potential compromises in quality and integrity. Quality, time, and cost are key components of sound project management and are also cited in recommended total quality management and continuous process improvement

techniques (rapid cycle improvement) [34]. Anyone who has designed a data collection instrument for research can attest to the importance of evaluating during the design phase how the questions will be answered and what will be done with the data collected. Considering, at the start of a research study, what the data will look like when aggregated across patients into tables or analysis files will inform the data collection methods and can prevent misinterpretations before they occur, ultimately optimizing the number of data points and ensuring adequate metadata such that the results can be readily understood and interpreted. In a clinical research study, especially one that supports a regulatory submission, ensuring accuracy, traceability, and trustworthiness of each data element is an incentive not to collect unnecessary data, along with consideration for those participating in the research. Principles and recommendations from the Coordinated Research Infrastructure Building Enduring Life-Science Services (CORBEL) consensus document (see Box 18.2) have provided an excellent resource for consideration of all aspects of data sharing when planning clinical research studies [35]. In addition to the broader principles, this reference provides additional useful detail on this topic.

Box 18.2. Principles and Recommendations for Using Patient Data in Research

1. The provision of individual participant data should be promoted, incentivized, and resourced so that it becomes the norm in clinical research. Plans for data sharing should be described prospectively and be part of study development from the earliest stages.
2. Individual participant data sharing should be based on explicit broad consent by trial participants (or if applicable by their legal representatives) to the sharing and reuse of their data for scientific purposes.
3. Individual participant data made available for sharing should be prepared for that purpose, with de-identification of data sets to minimize the risk of reidentification. The de-identification steps that are applied should be recorded.
4. To promote interoperability and retain meaning within interpretation and analysis, shared data should, as far as possible, be structured, described, and formatted using widely recognized data and metadata standards.
5. Access to individual participant data and trial documents should be as open as possible and as closed as necessary, to protect participant privacy and reduce the risk of data misuse.
6. In the context of managed access, any citizen or group that has both a reasonable scientific question and the expertise to answer that question should be able to request access to individual participant data and trial documents.
7. The processing of data access requests should be explicit, reproducible, and transparent but, so far as possible, should minimize the additional bureaucratic burden on all concerned.

8. Besides the individual participant data sets, other clinical trial data objects should be made available for sharing (e.g., protocols, clinical study reports, statistical analysis plans, blank consent forms) to allow a full understanding of any data set.
9. Data and trial documents made available for sharing should be transferred to a suitable data repository to help ensure that the data objects are properly prepared, are available in the longer term, are stored securely, and are subject to rigorous governance.
10. Any data set or document made available for sharing should be associated with concise, publicly available, and consistently structured discovery metadata, describing not just the data object itself but also how it can be accessed. This is to maximize its discoverability by both humans and machines.

Adoption and Implementation of Data Standards from the Start

Standards enable the exchange of data in well-recognized formats so that there is semantic interoperability among technologies, platforms, and systems such that the meaning, integrity, and traceability of the data are preserved. The standards should be harmonized from data collection through aggregation and analysis to minimize mapping. Mapping of legacy data can certainly be done, and there are computer programs to do just this; however, this always adds time and cost to the research study and may compromise quality when data interpretation is required, especially by those who did not conduct the study or care for the patient. In the case of CDISC standards, implementing CDASH for data collection has been shown to substantially reduce the time required for study start-up (by 70–90%) and the overall non-patient participation time for a research study by ~60% [36]. CDASH also paves the way for generating Study Data Tabulation Model (SDTM) when patient data are aggregated into tables or for analysis at the end of the study.

Another best practice is collecting CDASH-structured data using the CDISC Operational Data Model (ODM), which is a transport standard that supports audit trails (traceability) and electronic signatures. ODM is used by many electronic data capture tools and can be used as an export format from EHRs [37, 38]. Many of the EDC systems support ODM; however, they may have immature APIs which impact the ability to interface with the FHIR resources. CDISC is updating the ODM standard to develop more robust APIs allowing the use of FHIR.

The Innovative Medicines Initiative (IMI) in Europe also encourages the use of standards. Their European Translational Information & Knowledge Management Services (eTRIKS) project [39] was designed “to create and run an open and sustainable research informatics and analytics platform for use by IMI (and other) projects with knowledge management needs.” Regarding standards, the eTRIKS team has deemed that these are “vital tools” that facilitate the loading of data into knowledge management platforms to compare with other data sets. They have

provided a *Standards Starter Pack* [40] with guidance to data managers and others working with data in research on data standards for genomic, clinical, and translational data management. They have frequently requested input on this starter pack and plan to update it regularly. The standards included in the starter pack are varied and include standards for “translational science” from genomics through clinical trials.

Process Analysis and Design

The planning and use of standards from the start is fundamental to a streamlined research process. Ideally, an optimal electronic clinical research study would have data entered once and only once for multiple purposes, eliminating reentry and the consequent opportunity to introduce transcription errors. Using EHR as electronic source data (eSource) for clinical research data has been a dream, sparsely realized, for decades. As previously noted, studies have shown that, as opposed to the current “swivel chair” interoperability with transcription of EHR data into research systems [10], the extraction of data from the health record for research can increase quality by eliminating transcription/reentry errors and thus reducing resources and time [23]. The methodology implemented for this purpose was also used to report adverse events, decreasing the time of reporting from ~35 min to less than 1 min [41]. Reuse of EHR data has proven useful in projects in Europe and Japan, particularly when a standard ontology for interpreting and storing disparate EHR data was leveraged [42, 43]. The process for a research study should be mapped and evaluated during the planning stage of the study, following these three recommended best practices.

Role of Research in Learning Health Systems and LHS Core Values

eSource and the ability to use health data for research are critical to enable LHS to improve the health of individuals and populations through more rapid cycles of learning from research to informing care decisions at the bedside. LHS accomplish this by generating information and knowledge from data captured and updated over time – as an ongoing and natural by-product of contributions by individuals, care delivery systems, public health programs, and clinical research, disseminating what is learned in timely and actionable forms that directly enable individuals, clinicians, and public health entities to separately and collaboratively make informed health decision [44]. A Learning Health Community (www.learninghealth.org), launched in May 2012 [45], has developed a set of LHS core values, presented below [45]. Endorsers of these values can be found on the LHC website. All of the values are relevant to Data Sharing; however, it should be noted that many of the issues covered in this chapter are inherent in the value around “Scientific Integrity”.

1. *Person-Focused*: The LHS will protect and improve the health of individuals by informing choices about health and healthcare. The LHS will do this by enabling strategies that engage individuals, families, groups, communities, and the general population, as well as the US healthcare system as a whole.
2. *Privacy*: The LHS will protect the privacy, confidentiality, and security of all data to enable responsible sharing of data, information, and knowledge, as well as to build trust among all stakeholders.
3. *Inclusiveness*: Every individual and organization committed to improving the health of individuals, communities, and diverse populations, who abides by the governance of the LHS, is invited and encouraged to participate.
4. *Transparency*: With a commitment to integrity, all aspects of LHS operations will be open and transparent to safeguard and deepen the trust of all stakeholders in the system, as well as to foster accountability.
5. *Accessibility*: All should benefit from the public good derived from the LHS. Therefore, the LHS should be available and should deliver value to all while encouraging and incentivizing broad and sustained participation.
6. *Adaptability*: The LHS will be designed to enable iterative, rapid adaptation and incremental evolution to meet current and future needs of stakeholders.
7. *Governance*: The LHS will have that governance which is necessary to support its sustainable operation, to set required standards, to build and maintain trust on the part of all stakeholders, and to stimulate ongoing innovation.
8. *Cooperative and Participatory Leadership*: The leadership of the LHS will be a multi-stakeholder collaboration across the public and private sectors including patients, consumers, caregivers, and families, in addition to other stakeholders. Diverse communities and populations will be represented. Bold leadership and strong user participation are essential keys to unlocking the potential of the LHS.
9. *Scientific Integrity*: The LHS and its participants will share a commitment to the most rigorous application of science to ensure the validity and credibility of findings and the open sharing and integration of new knowledge in a timely and responsible manner.
10. *Value*: The LHS will support learning activities that can serve to optimize both the quality and affordability of healthcare. The LHS will be efficient and seek to minimize financial, logistical, and other burdens associated with participation.

Conclusion

Facilitating data sharing and reuse of electronic health data for research is an important foundation for reengineering and streamlining research processes and will be critical to accelerating learning health cycles and broadening the knowledge that can be used to improve healthcare and patient health outcomes. A range of ethical, legal, and technical considerations have thus far hindered the development and application of approaches for such reuse and data sharing, in general. However, standards adoption and technical capabilities are progressing, and incentives are

now beginning to align to facilitate data sharing. Principles and values of data sharing and the responsible use of data and data standards have been published, and there is recognition of the value of “real-world data” (RWD) to generate additional evidence upon which to base clinical decisions. These will require broad adoption, adherence, communication, and collective support to positively transform research processes and informatics.

Appendix

Examples of Collaborations, Initiatives, and Tools Related to Data Sharing in Clinical Research

In this section, we describe some important (national and global) collaborations, initiatives, and tools related to reusing clinical data for purposes of streamlining research. This list is not intended to be exhaustive.

ARO Council and Global Network	The Academic Research Organization Council and Global Network brings together Japan, Taiwan, Singapore, South Korea, Europe, and the USA with strategic initiatives toward harmonization and standardization of data to streamline clinical research and accelerate academic innovation to overcome intractable diseases [46].
ASTER	The Adverse Drug Event Spontaneous Triggered Event Reporting (ASTER) study was a proof of concept for the model of using data from electronic health records to generate automated safety reports, replacing the current system of manual ADE reporting. The CDISC-IHE Retrieve Form for Data Capture (RFD) formed the basis for the data sharing from EHRs to directly populate MedWatch forms. The time to report an AE was reduced from 34 min to less than 1 min [41].
BRIDG Model	The Biomedical Research Integrated Domain Group (BRIDG) Model is an information model that represents the domain of protocol-driven research. It provides a shared view of the concepts of basic, preclinical, clinical, and translational research, including genomics. This information model is an ISO, CDISC, and HL7 standard. It supports development of data interchange standards and technology solutions that can enable semantic interoperability for biomedical and clinical research and bridges research and the healthcare arena. Currently there is work being done to develop HL7 FHIR research resources, which will be harmonized with the BRIDG model [47, 48].
CAMD	Coalition Against Major Diseases (CAMD) is an initiative of the Critical Path Institute (C-Path). “CAMD is a public-private partnership aimed at creating new tools and methods that can be applied to increase the efficiency of the development process of new treatments for Alzheimer’s disease (AD) and related neurodegenerative disorders with impaired cognition and function. CAMD has the following areas of focus: (1) qualification of objective biomarkers, including both biochemical and observational digital biosensor measures of health, (2) development of common data standards, (3) creation of integrated databases for clinical trials data, and (4) development of quantitative model-based tools for therapeutics development” [49].

CDISC	Clinical Data Interchange Standards Consortium (CDISC) is a standards development organization focused on developing global data standards for clinical research. Its mission is to develop and support global, platform-independent data standards that enable information system interoperability to improve medical research and related areas of healthcare. CDISC has successfully collaborated with multiple other stakeholder groups to initiate and develop industry-accepted standards, such as BRIDG. CDISC standards are now required for data in regulatory submissions to FDA and Japan's PMDA [50].
C-Path	The Critical Path Institute (C-Path) is a nonprofit, public-private partnership with the Food and Drug Administration (FDA). C-Path's aim is to accelerate the pace and reduce the costs of medical product development through the creation of new data standards, measurement standards, and method standards that aid in the scientific evaluation of the efficacy and safety of new therapies."CAMD is an example of one of C-Path's projects, many of which are executed through consortia [51].
CFAST	The Coalition For Accelerating Standards and Therapies (CFAST) was initiated by C-Path and CDISC to develop global therapeutic area standards that augment the foundational standards for clinical research. Contributing organizations included FDA, EMA, PMDA, NCI, IMI, and TransCelerate BioPharma [52]. The resulting standards are published in CDISC Therapeutic Area User Guides and through CDISC SHARE, with specifications cited in the FDA and PMDA Data Standards Catalogs.
Common Protocol Template	The Common Protocol Template (CPT), developed by TransCelerate Biopharma Inc. in collaboration with stakeholders, is a harmonized and streamlined approach to the format and content of clinical trial protocols. It aims to ease interpretation by the study sites and global regulatory authorities while enabling downstream automation of many clinical processes and alignment to industry data standards. It has now been harmonized with an FDA/NIH protocol template and will become an ICH project [53].
CORBEL	The CORBEL (Coordinated Research Infrastructures Building Enduring Life-Science Services) consortium is an 11 major new biological and medical research infrastructures (BMS RI) in Europe, which plan to create a platform for harmonized user access to technologies, samples, and data services required for biomedical research [54].
DataSphere	Project DataSphere LLC, which is not-for-profit initiative of the CEO Roundtable on Cancer's Life Sciences Consortium (LSC), has developed a free digital library/data laboratory to share and analyze patient-level comparator arm data from phase III cancer clinical trials. "Prostate Dream Cancer Challenge confirmed that an open-access model empowers global communities of scientists from diverse backgrounds and promotes crowd-sourced solutions to important clinical problems" [55, 56].
Duke University eSource study	A proof-of-concept study that aimed to quantify the benefits of the secondary use of EHR data for clinical research. Findings showed a 37% reduction in time, significant reduction in resource needs and zero data quality issues [23].
ECRIN	The European Clinical Research Infrastructure Network (ECRIN) provides investigators support from trial preparation to implementation navigating the fragmented health and legal systems of individual European countries in order to conduct multinational trials. ECRIN led the CORBEL project on consensus-based principles for sharing clinical trial data, results of which were published in December 2017 in the <i>British Medical Journal Open</i> [57].

(continued)

EHR4CR	“The EHR4CR project, funded by the Innovative Medicines Initiative (IMI) and the European Federation of Pharmaceutical Industries and Associations (EFPIA) in collaboration with 34 partners (academic and industrial) and 2 subcontractors is one of the largest public-private partnerships aiming at providing adaptable, reusable and scalable solutions (tools and services) for reusing data from Electronic Health Record systems for Clinical Research” [58].
ELIXIR	ELIXIR “unites Europe’s leading life science organizations in managing and safeguarding the increasing volume of data being generated by publicly funded research. It coordinates, integrates and sustains bioinformatics resources across its member states and enables users in academia and industry to access services that are vital for their research.” ECRIN and ELIXIR are both part of the CORBEL consortium [59].
Health Level Seven (HL7)	Health Level Seven is an international standards organization dedicated to the development and interoperability of health information through products such as V2 and Fast Healthcare Interoperability Resources (FHIR) [10].
Healthcare Link and IHE-CDISC integration profiles	Under the leadership of Rebecca Kush and Landen Bain, CDISC launched the Healthcare Link Initiative to create a means of better linking healthcare and clinical research through standards. As a part of Healthcare Link, Integrating the Health Enterprise (IHE) and CDISC created the Retrieve Form Data Capture (RFD) and Retrieve Protocol for Execution (RPE) standards that a majority of electronic health record systems were configured for as part of Meaningful Use (MU) requirements [16, 60]. BRIDG also supports the Healthcare Link philosophy.
i2b2	Informatics for Integrating Biology and the Bedside (i2b2) is an NIH-funded National Center for Biomedical Computing (NCBC) aimed to develop an informatics framework based on Massachusetts General Hospital’s Research Patient Data Registry (RPDR) [61].
IDDO	The Infectious Diseases Data Observatory (IDDO) builds upon the success of WorldWide Antimalarial Resistance Network (WWARN) to provide a global collaborative data platform for the benefit of clinical care and research of communicable diseases [62].
I~HD	The European Institute for Innovation Through Health Data (I~HD) arose out of the IMI’s Electronic Health Records for Clinical Research (EHR4CR), SemanticHealthNet, and other projects to become an organization of reference and does so through services such as the Interoperability Asset Register, an online service that contains documents, templates, clinical models, technical specifications, and software pertaining to the interoperability of health information [63].
IMI	Innovative Medicine Initiative is a public- private partnership between the European Union and European Federation of Pharmaceutical Industries and Associations (EFPIA) that has resulted in over 100 projects generating 60+ project tools and 2000+ publications [64].
LHC	The goal of the Learning Health Community (LHC) is to improve the health of the individual and population through rapid cycle improvements to a learning health system (LHS) from the information and knowledge gained from data collected from clinical research, individuals, population health, and care delivery. The LHC will leverage existing opportunities such as meaningful use and personal health records and strive to create a harmonization among stakeholders to facilitate data sharing for the good of the individual and the population, promising to empower personalized medicine [44, 65].

OHDSI	Observational Health Data Sciences and Informatics (OHDSI) strives to share observational healthcare data through common data models and development of tools for data analytics and visualization [66]. OHDSI arose from initial work to develop the OMOP model, which is no longer an active project. The OMOP Common data Model is maintained by OHDSI.
OneMind	OneMind is dedicated to disseminating donor funding for brain disease and injury research including the data standardization, curation, and mining necessary for regulatory approvals. Standardization of the data from two mega-studies conducted at separate NIH institutes (National Institute of Neurological Disorders and Stroke and National Institute of Mental Health) allows for the data to be merged into a “collaboratory” at the completion of the studies [67].
PCORI	The Patient-Centered Outcomes Research Institute funds comparative clinical effectiveness research in order to change clinical practice and improve patient outcomes. The PCORI program consists of five areas of focus: clinical effectiveness and decision science, healthcare delivery and disparities research, evaluation and analysis, engagement, and research infrastructure known as PCORnet [68].
SHARE	The Shared Health and Research Electronic (SHARE) library is a metadata repository and associated tools and services that enables users of CDISC to access the standards in various formats that are human- and machine-readable [27].
Sentinel	The Food and Drug Administration’s (FDA) Sentinel Initiative is a national electronic system that enables researchers to proactively monitor the safety of FDA-regulated medical products after they reach the market complementing the FDA’s Adverse Event Reporting System. This system compiles data from multiple sources such as claims data, registries, and EHRs using a distributed data model that allows the ability to maintain patient privacy and monitor the safety of regulated products [69].
SMART on FHIR	Harvard Medical School and Boston Children’s Hospital initiated an interoperability project in 2010, with a goal of “developing a platform to enable medical applications to be written once and run unmodified across different healthcare IT systems.” This was named Substitutable Medical Applications and Reusable Technologies (SMART). In 2013, the platform was modified to adopt the FHIR standard that was emerging at that time. The new platform was called “SMART on FHIR”.
TRANSFoRm	The Translational Research and Patient Safety in Europe (TRANSFoRm) project is the European learning health system initiative aimed to develop a digital infrastructure, method, model, and standards for three areas of focus of a LHS: use of biobank data sets develop genotype and phenotypes for epidemiological studies, embedding regulated clinical trials within the EHR with a focus on patient-reported outcome measures (PROM), and decision support tools for clinical care [70, 71].
Vivli	Designed to reduce barriers to data sharing in clinical research, Vivli, acting as an independent broker, created an independent data repository, cloud-based analytics platform and search engine, based on the gatekeeper model, where industry, academia, patient organizations, government, and not-for-profit organization’s researchers can share, access, and host data [72].

References

1. The Data Economy. 9039, London : s.n., 6–12 May 2017, The Economist, Vol. 423.
2. Kush RD. Science Translational Medicine. 2009. pp. 24–28, Vol. 1, pp. 1–4.
3. Clinical Data Interchange Standards Consortium. [Online] [Cited: February 18, 2018]. <https://www.cdisc.org/system/files/members/standard/foundational/glossary/CDISC%20Glossary%20v11.pdf>.
4. Wikipedia. Wikipedia the Free Encyclopedia. [Online] [Cited: February 18, 2018]. <https://en.wikipedia.org/wiki/Traceability>.
5. Healthcare Information and Management Systems Society HIMSS. Healthcare Information and Management Systems Society HIMSS. [Online] [Cited: February 18, 2018]. <http://www.himss.org/library/interoperability-standards/what-is-interoperability>.
6. Hammond WE, Jaffe C, Kush RD. Healthcare standards development—the value of nurturing collaboration. *J Am Health Inf Manag Assoc* (AHIMA). 2009;80:44–50.
7. <https://bridgmodel.nci.nih.gov/>.
8. Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. Washington DC: United States Health and Human Services Office of Civil Rights. https://www.hhs.gov/sites/default/files/ocr/privacy/hipaa/understanding/coveredentities/De-identification/hhs_deid_guidance.pdf.
9. Merriam Webster Dictionary. [Online] [Cited: February 18, 2018]. <https://www.merriam-webster.com/dictionary/anonymization>.
10. Conn J. ‘Swivel chair’ interoperability: FDA seeks solutions to mesh EHRs and drug research record systems. s.l. Modern Healthcare; 2015.
11. Kanter JH. Your life, your health: share your health data electronically: It may save your life, Library of Congress Control Number 2012904124. Joseph H Kanter; 2012. p. 3.
12. Precision Medicine Initiative. Online: <https://ghr.nlm.nih.gov/primer/precisionmedicine/initiative>.
13. [Online]. <https://www.cdisc.org/esdi-document>.
14. [Online]. http://www.ema.europa.eu/docs/en_GB/document_library/Regulatory_and_procedural_guideline/2010/08/WC500095754.pdf.
15. Electronic Source Data Interchange (eSDI) Group. Leveraging the CDISC standards to facilitate the use of electronic source data within clinical trials. s.l. Clinical Data Interchange Standards Consortium, 2006.
16. ITI Technical Committee. IHE IT infrastructure technical framework supplement retrieve form for data capture. s.l. IHE International;s 2010.
17. HITSP Inabling Healthcare Interoperability. [Online] [Cited: February 18, 2018]. http://hitsp.org/InteroperabilitySet_Details.aspx?MasterIS=false&InteroperabilityId=456&PrefixAlpha=1&APrefix=IS&PrefixNumeric=08&ShowISId=456.
18. Kush R, Alschuler L, Ruggeri R, Cassells S, Gupta N, Bain L, Claise K, Shah M, Nahm M. Implementing single source: the STARBRITE proof-of-concept study. *J Am Med Inform Assoc*. 2007;14:662–73.
19. Takenouchi K, Yuasa K, Shioya M, Kimura M, Watanabe H, Oki Y, Aki. Development of a new seamless data stream from EMR to EDC system using SS-MIX2 standards applied for observational research in diabetes mellitus. *Learn Health J*. 2018.
20. Ferranti JM, Musser RC, Kawamoto K, Hammond WE. The clinical document architecture and the continuity of care record: a critical analysis. *J Am Med Inform Assoc*. 2006;13(3):245–52. <https://doi.org/10.1197/jamia.M1963>.
21. Delaney BC, Curcin V, Andreasson A, Arvanitis TN, Bastiaens H, Corrigan D, Ethier J-F, Kostopoulou O, Kuchinke W, McGilchrist M, van Royen P, Wagner P. Translations medi-

- cine and patient safety in Europe: TRANSFoRm- Architecture for learning health system in Europe. *BioMed Res Int.* 2015;.
- 22. HL7. Fast Healthcare Interoperability Resources. Online: <https://www.hl7.org/fhir/>.
 - 23. Nordo A, et al. A comparative effectiveness study of eSource used for data capture for a clinical research registry. *Int J Med Inform.* 2017;103:89–94.
 - 24. Food and Drug Administration. [Cited: July 6, 2018]. <https://www.fda.gov/ScienceResearch/SpecialTopics/RegulatoryScience/ucm507090.htm>.
 - 25. Kush RD, Goldman M. Fostering responsible data sharing through standards. *N Engl J Med.* 2014;370:2163–4.
 - 26. Connecting health and care for the nation a shared nationwide interoperability roadmap. Washington, DC: Office Of National Coordinator Health Information Technology; 2015.
 - 27. Federal health IT strategic plan 2015–2020. Washington, DC: Office of the National Coordinator Department of Health and Human Services; 2015.
 - 28. Fridsma D, Payne T. AMIA letter in support of ONC pledge to improve interoperability. American Medical Informatics Association. [Online] [Cited: February 18, 2018]. <https://www.amia.org/sites/default/files/AMIA-Letter-of-Support-Stakeholder-Commitments-Pledge.pdf>.
 - 29. Harmon A. Where'd you go with my DNA? New York: New York Times; 2010.
 - 30. Skloot R. The immortal life of henrietta lacks. Crown publishers, ISBN 978-1-4000-5217-2.
 - 31. Committee on Strategies for Responsible, Board on Health Sciences Policy, Institute of Medicine of the National Academies. Sharing of clinical trial data: maximizing benefits, minimizing risk. Washington DC: The National Academies Press; 2015. <http://www.nap.edu>
 - 32. Zozus M, et al. Assessing data quality for healthcare systems data used in clinical research. Washington, DC: NIH Collaboratory Health Care Systems Research Collaboratory.
 - 33. Sherman RE, Anderson SA, Dal Pan GJ, Gray GW, Gross T, Hunter NL, LaVange L, Marinac-Dabic D, Marks PW, Robb MA, Shuren J, Temple R, Woodcock J, Yue LQ Califf RM. Real-world evidence — what Is it and what can it tell us? *N Engl J Med.* 2016;375(23). <https://doi.org/10.1056/NEJMsb1609216>.
 - 34. Pelletier L, Beaudin C. Q Solutions: Essential Resources for Healthcare Quality Professionals. National Association for Healthcare Quality. Second Edition.
 - 35. Ohmann C, et al. Sharing and reuse of individual participant data from clinical trials: principles and recommendations. *BMJ Open.* 2017;7:e018647. s.l.
 - 36. Rozwell C, Kush R, Helton E. Saving time and money. *Appl Clin Trials.* 2007;16(6):70–4.
 - 37. Douga M ODM. s.l. Clinical data interchange standards consortium.
 - 38. Hume S, et al. Current applications and future directions for the CDISC operational. *J Biomed Inform.* 2016;60:352–62.
 - 39. eTRIKS. [Online] [Cited: February 18, 2018]. <https://www.etriks.org/>.
 - 40. <https://www.etriks.org/standards-starter-pack/>.
 - 41. Brajovic S, et al. Quality assessment of spontaneous triggered adverse event reports received by the Food and Drug Administration. s.l. *Pharmacopidemiol Drug Saf.* 2012;21. <http://www.asterstudy.com/>.
 - 42. Fadly A, Rance B, Lucas N, et al. Integrating clinical research with the healthcare enterprise: from the RE-USE project to the EHR4CR platform. *J Biomed Inform.* 2011;44:S94–S102.
 - 43. Takenouchi K. Healthcare link project in Japan: development of a new seamless data stream from EHR to EDC system using SS-MIX2 storages. Chicago: DIA; 2017.
 - 44. Friedman C, Rubin J, Brown J, et al. Toward a science of learning systems: a research agenda for the high-functioning Learning Health System. *J Am Med Inform Assoc.* 2015;22(1):43–50. <https://doi.org/10.1136/ajmajnl-2014-00297>.
 - 45. Learning Health Community. [Online] [Cited: February 18, 2018]. <http://www.learninghealth.org/history/>.

46. Academic Research Organization (ARO). s.l. Clinical Data Interchange Standards Consortium; 2017. https://www.tri-kobe.org/koho/PressRelease/2016/1st_ARO_WS_flyer.pdf.
47. BRIDG. Clinical data standards interchange consortium. [Online]. <https://www.cdisc.org/standards/domain-information-module/bridg>.
48. Becnel LB, Hastak S, Ver Hoef W, Milius RP, Slack M, Wold D, Glickman ML, Brodsky B, Jaffe C, Kush R, Helton E. BRIDG: a domain information model for translational and clinical protocol-driven research. *J Am Med Inform.* 2017;24:882–90.
49. Coalition Against Major Diseases (CAMD). Critical Path Institute. [Online] [Cited: February 19, 2018]. <https://c-path.org/programs/camd/>.
50. CDISC. Clinical data interchange standards consortium. [Online] [Cited: February 19, 2018]. <https://www.cdisc.org/>.
51. Critical Path institute. Critical Path Institute. [Online]. U.S. Food and Drug Administration. [Cited: February 19, 2018]. <https://c-path.org/about/>.
52. Coalition For Accelerating Standards and Therapies. Critical Path Institute. [Online] [Cited: February 19, 2018]. <https://c-path.org/programs/cfast/>.
53. Common Protocol Template. TransCelerate Biopharma Inc. [Online] [Cited: February 19, 2018]. <http://www.transceleratebiopharmainc.com/assets/common-protocol-template/>.
54. CORBEL – Coordinated Research Infrastructures Building Enduring Life-science Services. elixir. [Online] [Cited: February 19 , 2018]. <https://www.elixir-europe.org/about/eu-projects/corbel>.
55. Bertagnoli M, et al. Advantages of a truly open-access data-sharing model. *N Engl J Med.* 2017;376:1178–81. <https://doi.org/10.1056/NEJMsb1702054>.
56. Project Data Sphere. [Online] [Cited: February 12, 2018]. <https://www.projectdatasphere.org/projectdatasphere/html/PressRelease/LAUNCH>.
57. European Clinical Research infrastructure Network. [Online] [Cited: February 10, 2018]. <http://www.ecrin.org/>.
58. De Moor G, Sundgren M, Kalra D, Schmidt A, Dugas M, Claerhout B, Karakoyun T, Ohmann C, Lastic P, Ammour N, Kush R, Dupont D, Cuggia M, Daniel C, Thienpont G, Coorevits P. Using electronic health records for clinical research: the case of the EHR4CR project. *J Biomed Inform.* 2014; <https://doi.org/10.1016/j.jbi.2014.10.006>.
59. ELIXIR [Online] [Cited: February 19, 2018]. <https://www.elixir-europe.org/>.
60. EHR incentives and Certifications. Health IT.gov. [Online] [Cited: February 18, 2018]. <https://www.healthit.gov/providers-professionals/meaningful-use-definition-objectives>.
61. Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, Kohane I. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *J Am Med Inform.* 2010;17:124–30.
62. Infectious Diseases Data Observatory. [Online] [Cited: February 10, 2018]. <https://www.iddo.org/tools-and-resources>.
63. i-HD. The European Institute for Innovation through Health Data i-HD. [Online] [Cited: February 08, 2018]. <http://www.i-hd.eu/index.cfm/about/description-and-scope/>.
64. Innovative Medicines Initiative. IMI. [Online] [Cited: February 10, 2018]. <http://www.imi.europa.eu/projects-results/catalogue-project-tools>.
65. [Online] [Cited: February 09, 2018]. Toward a science of learning systems: a research agenda for the high-functioning Learning Health System. <https://www.ncbi.nlm.nih.gov/pubmed/25342177>.
66. Observational Health Data Sciences and Informatics. [Online] [Cited: February 11, 2018]. <https://ohdsi.org/>.
67. One Mind. [Online] [Cited: February 09, 2018]. <https://onemind.org/>.
68. Patient Centered Outcomes Research Institute. [Online] [Cited: February 10, 2018]. <https://www.pcori.org/>.

69. FDA's Sentinel Initiative. U.S. Food and Drug Administration. [Online] [Cited: February 10, 2018]. <https://www.fda.gov/Safety/FDAsSentinelInitiative/ucm2007250.htm>.
70. Delaney BC, Curcin V, Andreasson A, et al. Translational medicine and patient safety in Europe: TRANSFoRm—architecture for the learning health system in Europe. *Biomed Res Int.* 2015; <https://doi.org/10.1155/2015/961526>. s.l.
71. TRANSFoRm. i-HD [Online] [Cited: February 10, 2018]. <http://www.transformproject.eu/>.
72. Vivli Center for Global Research Data. [Online] [Cited: February 10, 2018]. <http://vivli.org/>.



Developing and Promoting Data Standards for Clinical Research

19

Rachel L. Richesson, Cecil O. Lynch, and W. Ed Hammond

Abstract

This chapter describes the importance of data standards in clinical research, particularly for streamlining regulatory oversight and enabling research that is conducted using electronic health record systems in “real-world settings.” Standards are needed to exchange data between partners with preserved meaning and to enable accurate analytics, a core aim of research. There are different types of standards and numerous organizations – national, international, and global – that develop them. The coordination and harmonization of these efforts will be necessary to fully realize an efficient clinical research system that is synergistic with healthcare systems in the USA and abroad. We highlight important collaborations that are influencing the development and use of clinical and research standards to solve significant and outstanding scientific, societal, and business challenges of biomedical research and population health.

Keywords

Clinical research data standards · Standards development · Data exchange
Healthcare informatics · Clinical research informatics

R. L. Richesson, PhD, MPH, FACMI (✉)
Duke University School of Nursing, Durham, NC, USA
e-mail: rachel.richesson@dm.duke.edu

C. O. Lynch, MD, MS
Accenture Digital, San Francisco, CA, USA
e-mail: cecil.o.lynch@accenture.com

W. Ed Hammond, PhD, FACMI, FIMIA, FHL7, FIAHSI
Duke Center for Health Informatics, CTSI, Durham, NC, USA
e-mail: William.Hammond@duke.edu

Clinical Research: Escalating Efficiencies with Data Standards

Calls for the transformation and reengineering of our national clinical research system have been interminable for decades. The research studies required to test the efficacy and effectiveness of new treatments take time to design and accrue patients. The studies are incredibly expensive to implement, and the majority are unable to enroll enough patients to be completed. The review and approval process for investigational new drugs through the US Food and Drug Administration (FDA) is slow, in large part because studies use different variables and measures, requiring reviewers to develop a custom process for each submission and review. Comparing the safety or efficacy between multiple drugs is challenging to impossible as different studies use different endpoints – even within the same disease area. The collection of adverse events is passive and therefore incomplete. As a consequence, the system is insufficient to monitor the performance and safety of drugs and devices used by patients in the real world. There generally are no follow-up or population-based studies to assess the long-term impact of drugs or devices on patients whose clinical profiles, lifestyle factors, and compliance are markedly different than the eligibility criteria of the trials that led to market approval. In short, research is expensive and time-consuming and has limited generalizability. At best the current national clinical research system is full of missed opportunities, and at worst it is downright wasteful and possibly dangerous.

Standards play an important role in addressing the problems described above, and both industry and regulators have demonstrated engagement and support for creating and using standards in drug development and drug safety assurance activities. The Clinical Data Interchange Standards Consortium (CDISC) formed in 1997 as a collaboration of biopharmaceutical, technology, and regulatory partners to develop and promote standards that were desperately needed to speed the compilation of study data required for regulatory submissions and to improve the time for review and decisions by the FDA, the European Medicines Agency (EMA), and other international regulatory bodies. The FDA steadfastly supports CDISC and its standards mission as a means to improve the efficiency of the FDA to regulate drugs and devices in order to safeguard public health.

CDISC has successfully created a number of standards to support the reporting and sharing of the results and supporting data from clinical trials, and these standards are mandated by the FDA and widely adopted by pharmaceutical companies. There are reports that these standards have created measurable efficiencies for the companies (in terms of faster study start-up and compilation of submissions to the FDA) and for the FDA (in terms of streamlined reviews). Despite these impacts, research remains expensive, and it still can take a decade or more to finalize the human studies required for regulatory approval. Standards are still needed to optimize many of the tasks and workflows in the design, conduct, and analysis of clinical trials. Additionally, there is a need to understand and use clinical data standards to support the integration of research into healthcare delivery systems.

The narrative in Box 19.1 presents a vision for a well-functioning clinical research system, unencumbered by the inefficiencies we now see. Underlying these improvements is the idea of data exchange between different clinical research information systems (see Chap. 9), supported by data standards.

Box 19.1 Vision for an Efficient Clinical Research System

- Fundamental clinical research tasks – e.g., identifying participants for trials, obtaining informed consent, and locating existing biological samples – would be relatively effortless.
- Data for research protocols could be conveniently collected at the time of clinical visits. The data specific to research could be collected as specified in the research protocol, and protocol events could be directed without significant additional cost.
- Data collection systems would be configured to ensure that data are collected by the most appropriate person or device, at the proper time, according to protocol specifications, thereby enhancing the quality and consistency of the data.
- Standardized data elements could be easily searched and retrieved from a common location, and used for other studies, reducing the time to design and build case report forms.
- Data managers preparing the data could reuse tools that translate standardized data elements into analytic variables, automatically generating data dictionaries and preparing coding for study analysis.
- FDA reviewers could quickly reproduce and verify study analyses without having to customize the review process for each new submission and could use similar processes and code for multiple submissions in the same area.
- Adverse events could be identified anytime patients present to any provider, in the USA or internationally, and seamlessly reported to regulatory and public health authorities.

Standardized representation of eligibility criteria for trials could be automatically matched to historical data to determine with confidence whether there are sufficient numbers of eligible patients, reducing the likelihood of trial failures related to accrual. These same structured eligibility criteria could be matched to patient data and identify new subjects as they become eligible. The need to reenter clinical data in research systems and duplicate clinical tests could be reduced. When new data are needed, the standardized questions, already developed, could be rapidly added to clinical trial management systems. When new sites need to be added to a study, the same data collection tools, with standardized data elements, could be easily reused. Because these standardized data elements can be quickly transformed into analytic variables, regulatory reviewers could reuse existing code for replication of analyses and conceivably review safety of new products in days rather than months. Adverse event reporting, using a standard clinical vocabulary such as SNOMED CT, could be highly automated, and regulators could more easily identify risks for products already marketed. These systems in turn would build confidence in the system, perhaps leading to expanded use of conditional approvals (with post-market monitoring) for future products. It would be faster to design, plan, implement, conduct, and analyze, submit, and review new investigations. Further, the development and use of standards for eligibility, patient features, data collection,

and adverse events would increase innovation in clinical trial management systems, as new companies could focus development resources on products and functions that enhance workflow for research rather than on the representation and collection of data.

Clinical Research Standards Developers and Drivers and Stakeholders

A number of successful data standards efforts are supporting the activities presented in the vision for an efficient clinical research system described above. For adverse events, the International Conference of Harmonization (ICH, a collaboration of the regulatory authorities of Europe, Japan, and the USA) has developed a set of data elements (the E2B data model standard) for transmitting individual case safety reports, which will enable the development of standardized electronic regulatory data reporting applications by various vendors [1].

The first CDISC standards focused on creating specifications for standardizing data sets for submission to the FDA. The Study Data Tabulation Model (SDTM) specifies required and optional variables, associated controlled terminology (i.e., code lists or data values) and formats for tabulation, analysis dataset creation, and the actual data submission. CDISC later developed the Clinical Data Acquisition Standards Harmonization (CDASH) standards to standardize data on the front end (i.e., on the case report at the time of the collection). The CDISC organization hosts and maintains an ongoing inventory of data definitions and provides a library of case report forms using CDASH data elements for its members. Both SDTM and CDASH utilize controlled terminology lists (e.g., body site, laboratory tests, units of measure) developed by CDISC. CDASH is optimized for data capture and SDTM for submitting the research data. As one might expect, there is a tremendous overlap in content between CDASH and SDTM – at least 60–80% depending upon the direction of mapping. When used together, CDASH and SDTM enable the standardization and formatting of the data sets submitted to FDA by pharmaceutical companies. The CDASH and SDTM data elements can be retrieved free of charge from the Cancer Data Standards Repository (caDSR), a public resource hosted by the National Cancer Institute.

Another important CDISC contribution has been the development of *therapeutic area standards* to represent data that pertains to specific disease areas. These standard data elements and models are designed to ensure that regulatory submissions within a given disease area have consistency in the names of variables and terms and, more importantly, with study endpoints. The development of therapeutic area standards by clinical domain specialists and subsequent adoption by research sponsors will generate new efficiencies for regulatory and safety reviewers in specified disease areas. CDISC has published user guides for a number of therapeutic areas, including Alzheimer's, asthma, diabetes, and many others [2]. The therapeutic area standards were developed in response to a 2011 list of 54 prioritized disease and therapeutic areas (compiled the FDA's Center for Drug Evaluation and Research

and Center for Biologics Evaluation and Research) for which standardized data elements, terminologies, and data structures were needed to enable automation of important analyses of clinical study data to support more efficient and effective regulatory decision-making.

CDISC has worked to provide integrated standards that can link or bridge different parts of the clinical research workflow, and this paradigm is important for continued improvement and efficiency of biomedical and clinical research – from product development, study design, evaluation, and safety and regulatory approval. CDISC has built its own data exchange specification, called the CDISC Operational Data Model (ODM). This standard is designed to enable interoperability and has enabled a broad range of use cases, including study planning, data collection, electronic data capture from EHRs, data tabulation and analysis, and study archival.

Despite the positive impact that the aforementioned research-specific data, information, and transmission standards have made, even greater research efficiencies might be achieved if one thinks about clinical research on a grander scale – as a system that complements healthcare delivery and works synergistically with healthcare information systems to identify and provide (research-based) solutions to population health problems. In the next section, we present an enhanced vision of the future, which provides a rationale to explore a broader-range healthcare data standards and the standards development organizations that create them.

Advancing Research by Fully Integrating with Health Systems: Relevance of Health Data Standards to Clinical Research

A broad vision of clinical research functions that are deeply integrated into healthcare delivery which can be used to illustrate the value of leveraging national health information standards and infrastructure to support research that addresses health of populations is shown in Box 19.2.

Box 19.2 A Vision of Efficient Clinical Research Integrated with Health Information Systems

- All components of vision of efficient research are described in Box 19.1.
- The generation of “real-world evidence” from clinical data is seamless, requiring minimal if any additional data collection.
- The data collected in EHRs could be used for clinical research, reducing the number of data points collected specifically for the research.
- Issues of provenance (i.e., who recorded, changed, or authenticated each piece of data) can be reliably assured.
- Adverse events can be managed within the context of comprehensive patient care records, so that health providers are aware of research participation and trial interventions in their patients.

- Clinical data from any provider – in the USA or internationally – can be used for recruitment, study data, or monitoring adverse events.
- Researchers, regulators, and patient safety advocates can be notified of emergent health issues (i.e., possible adverse events, safety issues, or confounding factors) that are reported to any healthcare organization.
- Data elements for new data collection can be identified from curated libraries by searching metadata and quickly implemented into EHR systems to address clinical, business, and research questions.
- Data from a number of sources can be used to identify current and important problems of patients and providers, which can be identified as research priorities at local, regional, and national levels.
- Questions about comparative effectiveness of different treatments, or evaluation of new treatments and interventions, can quickly be answered by randomizing patients or providers to different treatments.
- Studies – using combination of EHR data and new data – can be rapidly configured and implemented to address local questions.
- Customized communications – including recruitment invitations, progress updates, and research results – could be sent to patients and their providers, sensitive to the context of the patient’s health and visit schedule.
- Patients can be informed of research opportunities and willingly participate if they perceive them to be important and coordinated with their care and providers.
- Consent and participation in the study would be simple and easy.
- New data elements can be rapidly configured when needed to collect data to address emergent clinical questions about optimal treatments and patient safety.
- Patient data is available to facilitate the continuity, safety, and quality of patient care, as well as to support clinical trials, data mining, public health reporting, reimbursement, audit, and performance measures, with little additional effort.

In a world with integrated and interoperable health information systems, public health reporting would be timely, accurate, and complete. The experience of real people would influence the topics and funding for new research, and subsequent discoveries could be put into practice and evaluated quickly and continuously. Quality of care could be explored, in real or near-real time, within and across organizations and national boundaries, and the science around treatment innovation, implementation, and quality improvement would continue to evolve. Learning health systems, where practice informs research (including identifying patient and population health needs and research priorities) and research informs practice, could truly be the norm – in healthcare organizations large and small.

All of these scenarios require the secure sharing of patient data, which depends upon the notion of interoperability – the ability of different information systems and

software applications to communicate, exchange data, and use the information that has been exchanged. Standards, i.e., the specifications for the collection, exchange, and security of clinical and research data, are essential for interoperability and hence to the vision of integrated clinical and research systems that can work together to advance biomedical knowledge and continuously improve population health.

Types of Healthcare Standards

The different activities related to the collection, storage, transfer, and use of data in healthcare and research provide a framework for organizing standards by their function, as shown in Fig. 19.1. The broad functions for standards (depicted as blue pentagons) are presented for the general steps in any data collection, analytics, or exchange project. These steps include the *planning* of a data collection, analytics, or exchange activity, the definition of *data structures* (e.g., the formatting and representation of the data), the process of *collection* (or ingestion) of the data, the *preparation and transformation* of the data to address specific needs of the project, the *exchange* (or transfer) and *storage* of the data, as well as the use and presentation of the data in *EHR applications* or *query specifications*. These steps or

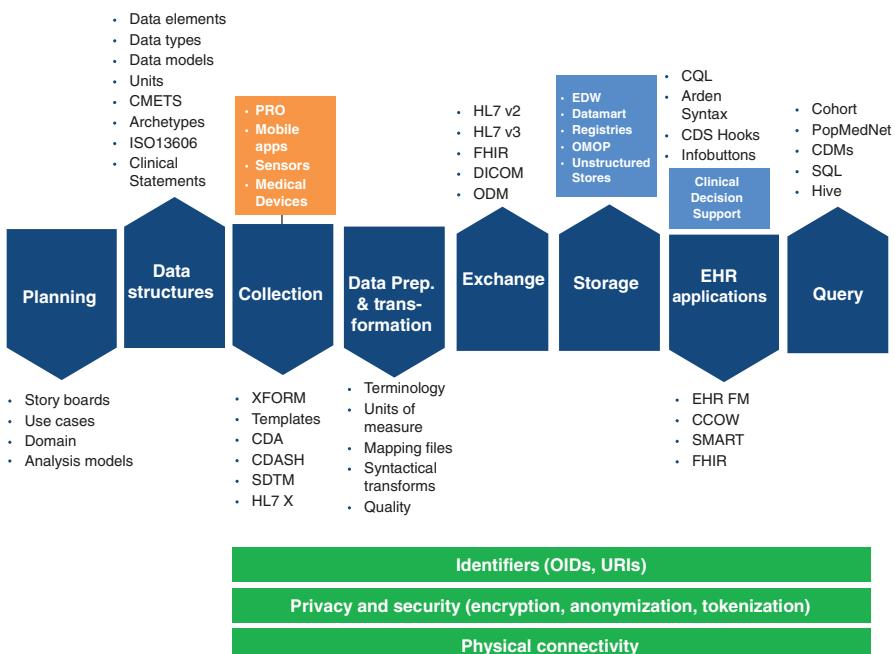


Fig. 19.1 Standards specifications by function for projects that collect or use health data. (This figure is intended to provide an overview of the large number of standards that exists for each step in a data collection, analytics, or exchange project. The standards listed are important, but not exhaustive, and are defined in the Appendix)

activities can be thought of as the building blocks of a data collection, analytics, or exchange project, and interoperability, from end-to-end, requires them to work together seamlessly.

Figure 19.1 also highlights the different types of standards used to support data-relevant projects (i.e., a data collection, analytics, or exchange projects). At the most basic level are physical connectivity, privacy and security standards, and identifiers. These highly technical specifications are essential for any kind of data collection or exchange. Standards, including use case and data analysis models, can assist in *planning* and development of architectures for data collection and exchange. Tools such as the Business Process Modeling Notation (BPMN) and Unified Modeling Language (UML), coupled with appropriate domain knowledge, can be used to create standard representation of the processes, data, and functional requirements that underpin the design and construction of information systems. These same standards can be used to ensure that different applications within a subject domain represent terms and entities consistently.

After the critical planning phase, a series of processes are involved in getting the data to the part where it is ready to be used in applications. In the *data structure* step, researchers must make decisions on the specification for data elements, data types, etc. The *data collection* (or ingestion) step can bundle a number of data collection items into packages (e.g., templates, disease-related data) and also provides standards that guide how the data collection interfaces are designed. Once collected, the data can be transmitted to other systems to be used to support other applications. The *preparation and standardization* of data activities are the most critical in terms of preserving intended meaning; these activities include terminology binding to information models. A host of *data exchange* standards (discussed in detail in the next section) specify the syntactical and content requirements so that data can be transmitted between systems without losing its meaning (to varying degrees of success). Moving further to the right on Fig. 19.1, we see a number of standards relevant to the *storage of data*, ideally in its most finely grained form and with appropriate modifiers (e.g., location and timing for an observed heart murmur) for maximum utility, in various containers such as enterprise data warehouses (EDW), data marts or registries, or common data model specifications. Generally, the data storage architectures for EHR systems are proprietary, but there are standards-based approaches, such as the CEN EN 13606 standard, or an EHR architecture that is a set of CDAs or CCR documents.

Data is most frequently used and presented independent of its collection, and so a number of standards support the presentation and use of data in *EHR applications*. Most notably are the use of clinical data as inputs for clinical decision support (CDS) logic and functions (building on standards such as Arden Syntax and CDS Hooks) and use of data for bolt-on applications (e.g., SMART on FHIR apps). Finally, there are standards related to data *queries* – for example the use of computable phenotype definitions for identifying research cohorts or the value sets developed by Centers for Medicaid and Medicare Services (CMS) and hosted on the NLM Value Set Authority Center (VSAC) to support quality measurement reporting.

The data exchange and preparation and exchange activities of a data-intensive project each have had rapid evolution of standards and are particularly relevant to clinical research. In particular, these standards support the ability to access, transmit, and use EHR data for research purposes and to integrate research-specific data elements into EHR systems for collection at the point of care or research. These warrant a deeper discussion and will be described in the following two sections.

Data Exchange Standards: The Evolution of FHIR

There are multiple standards for the exchange of data between applications, as seen on Fig. 19.1. These have been developed by many different SDOs and are mostly focused on specific domains. For example, the Digital Imaging and Communications in Medicine (DICOM) standard which is used universally for exchanging images and the National Council for Pharmacy Drug Program (NCPDP) have created a set of standards for e-prescribing and reimbursement for drug prescriptions. CDISC developed the ODM for exchanging and archiving clinical and translational research data and associated reference data and audit information.

The most common form of a general health data exchange standard is called a *messaging standard*. The most popular standard for data exchange used in the USA today is the HL7 version 2.x standard (HL7 v2), developed by Health Level Seven (HL7) in 1987. Created at a time of limited bandwidth and computing power, the HL7 v2 standard uses defined messages composed of functional segments, which in turn are composed of data fields, composed of data elements. Data elements are defined by position within the fields, separated by a hierarchical set of delimiters. In the late 1990s, HL7 introduced a more robust and sophisticated model-based exchange standard, version 3 (HL7 v3), which enables interoperability through the use of a Reference Information Model (RIM). The HL7 v3 standards are fundamentally different than version 2 in that they focused on the process of building applications rather than the syntax as in HL7 v2. While some model-based aspects of HL7 v3 were a success (such as the CDA), the use of HL7 v3 for data exchange was not well adopted, and users unremittingly complained about the complexity.

FHIR

In response to the end-of-life of HL7 v2 and the slow uptake and complexity of HL7 v3, HL7 convened a task force called “Fresh Look” and sought to step back with no constraints and look at how a standard could be developed with modern technologies and developer-friendly methodologies. There were no requirements to reuse any of the existing standards, but it was recognized that there were certainly components of the prior versions of HL7 that would accelerate the development of any new artifacts. In 2011, the modeling and methodology workgroup approved the RFH (Resources For Health) project which is recognized as the birth of Fast Health Interoperability Resources (FHIR). The first normative version of FHIR was

published in 2018, and there has been a wide pre-adoption of the draft standard by EHR vendors and industry, resulting in a number of demonstrations using FHIR in clinical applications. Its simplicity for developers and focus on using existing data (with little modeling) have facilitated the rapid development of FHIR-based applications directed toward real clinical information needs, and demonstrations of these applications in turn create new adopters for the applications and an escalating interest in the FHIR specification and the innovations it will likely enable.

FHIR has been very well-received by the informatics and healthcare community, and there is currently a strong momentum and tone of optimism around FHIR. It is worth noting that the number of FHIR adopters is greater than for any previous HL7 standard. The audience for healthcare standards is bigger than ever before, and the number of partnerships (commercial and public) that are forming around FHIR standards is unprecedented. FHIR provides a viable pathway to enable the missions laid out by visionary initiatives for health, including precision medicine initiatives and the Cancer Moonshot.

The basic unit of FHIR is a resource, a fully encapsulated contextual healthcare element. This is akin to the HL7 v3 CMET or common message element type. Also borrowed from HL7 v3 is the terminology model with slight modifications. Each FHIR resource can be expressed in a number of different technology implementations including XML, UML, JSON, and an RDF format (turtle syntax). While a transport mechanism is not dictated, the most common method of implementation is over a RESTful API using HTTPS transport. Each resource is published with HL7 v2 and HL7 v3 mappings where they exist.

The impact of FHIR for research informatics cannot be underestimated. It offers a means to integrate directly within EHR systems using SMART (Substitutable Medical Applications, Reusable Technologies) on FHIR. FHIR provides a mechanism to dynamically pull data elements of interest from EHR systems for research projects. FHIR includes a consent resource that is robust enough to develop research-oriented consent models with granular levels of options for participation and also provides the security models necessary to provide confidence in transmission of protected health data. FHIR resources can support a wide range of clinical observations, device data used to collect that information, and a full model for genomics metadata to aid in the new directions of research. Combined with SMART functions, FHIR could support the acquisition of patient-reported and patient-generated data and combine it with information acquired from their EHR.

The current challenge with FHIR now, and into the next decade, will be the addition of resources to address spectrum of research needs and to ensure that they are standardized. To some extent, this explosion of resources is managed in two ways in FHIR. First is the terminology-driven nature of resources that allows a level of abstraction. This is primarily managed in the “category” element that enables one to designate the classification of type of observation (e.g., lab, imaging, vital signs) and the “code” element that allows a description of the type of classified observation (e.g., the specific LOINC code for a lab order). The second mechanism is through the use of extensions that any resource can have. These extensions will allow

nuanced addition of elements to a resource. For instance the observation-time-offset extension to the Observation resource allows the recoding of milliseconds offset of time in sequential sampling. There is also some contextual information that can be used to further specialize an observation such as the reference to another resource that defines what the observation is based on (basedOn element) such as a ProcedureRequest or CarePlan.

Data Preparation and Transformation: Terminology Binding

This stage is perhaps the most important and greatest challenge. Collection of data is not the endpoint of research but rather the beginning. The analysis and dissemination of the results of the analysis are the end goal. To enable reliable, publishable results, the data must be made ready for analytics. There are two main functions executed in the preparation phase. First is the syntactical normalization which involves the conversion to a single data format and data model. This also involves the normalization of units of measure to a common representation. The second functional process of “data preparation” involves the tagging of concepts with terminology fit for the domain (e.g., LOINC for lab, SNOMED CT for clinical observations) and the mapping between terminologies (e.g., ICD-10 to SNOMED) so that reliable comparison of data collected across sites or over time from the same site can occur. It may also require the “roll-up” of similar leaf concepts to a parent so that features are reduced, such as using the parent “demyelinating CNS disease” to group “multiple sclerosis” and “subcortical leukoencephalopathy” for the purpose of studying the effect of anti-lipid agents of all forms of demyelination.

As background, it is important to understand that terminologies and coding systems used in healthcare information systems have very different structures and features and are often large and complex. They are not merely data dictionaries or flat enumerated lists of values. They have dimensionality, implicit and explicit semantics, and data formats associated with them. They come from different organizations with different curation policies and update schedules. They are typically designed to work in one context, and their curation environments likely reflect different commitments to the use of standard in that context. Some are designed for strict contexts (e.g., ICD) and others for many contexts (e.g., LOINC and SNOMED CT). Overlaps are common. For example, SNOMED CT covers medications although other controlled terminologies do as well. SNOMED CT also covers laboratory tests, as does LOINC. Several countries use different parts of SNOMED CT (e.g., laboratory test names and medications) where the USA does not. LOINC is moving toward standardized patient assessments and data elements. Because there are so many standards in use, mapping has been proposed as a way toward interoperability. However, the very heterogeneous structures, scope, and features of healthcare terminologies make mapping a very difficult activity that is inherently vulnerable to loss of meaning (Box 19.3).

Box 19.3 Mapping

Despite emerging and promising cooperative efforts, there are still, unfortunately, many overlapping and competing standards addressing all aspects of healthcare and data. The most common approach has been to allow the coexistence of overlapping standards by supporting mapping efforts between the standards.

Mapping is the process of finding a concept in a target terminology that “best matches” a particular concept in the source terminology, although what a “best match” means can range from exact synonymy to mere relatedness, depending upon the context of use. The process for creating cross-terminology mappings itself is time-consuming and labor-intensive, and there are potential problems with the mapping approach, including information loss and ambiguity [3].

Mappings are by definition context specific and are not an ideal or easy solution due to a lack of a uniform standard. Mapping between two standards will always result in some loss of information. (If they mapped perfectly, why have two standards?) Further, it is impossible to keep two independent standards synchronized. Ongoing maintenance is essential and can consume considerable resources [4]. How best to handle versioning in mappings is still a largely unresolved issue [3].

In addition, the mapping approach is much more difficult to support when including different data models that underlie various medical systems. In general, mapping should be considered a work-around and not a solution. With that said, if mapping will be used, then existing mappings should be used. The Unified Medical Language System (UMLS) in the USA (globally available) has facilitated the mapping of various terminologies and coding systems [5].

In addition to mapping between terminologies, there are also needs to map between different information or data models. These maps must address both data fields and code names that can be used in different combinations for different data models. If one data model collects race and ethnicity, and another only collects race, then the analysis across sites is limited to the race variable and also to the smallest set of codes available. Again, there is often information loss as analytics can only be performed at the “lowest common denominator.”

A number of different common data models have been used to bring together heterogeneous data for research networks or multi-site studies. Examples include the PCORnet Common Data Model (CDM) [6] used in the Patient Centered Outcomes Research Network (PCORnet) and the Observational Medical Outcomes Partnership (OMOP) CDM[7], used in a number of research networks including the Observational Health Data Sciences and Informatics (OHDSI), which is heavily used by pharmaceutical companies in clinical research. Local data models are mapped to the CDM, and in some cases the CDMs are mapped to each other. Analysts using any of these data must be aware of missing data and information loss that might occur in these transformations [8].

The situation becomes even more complicated when one considers that terminologies do not operate in isolation. Terminologies alone are insufficient to precisely communicate clinical or scientific meaning; they must be bound to clinical data models to fully represent the semantic context, and there are many approaches (and few standards) to do this. The easiest way to think about these models is as collection of data elements that can take on a range of pre-defined values with agreed-upon meanings. For example, “family history of cancer” could be represented as an (data) element in a clinical data model, with values of yes/no, present/absent, or perhaps different types of cancers. The same concept “family history of cancer” could also be represented entirely in the terminology (assuming a sufficiently robust clinical terminology such as SNOMED CT), or the concept could be modeled in different ways – e.g., the data element could be “family history of [conditions],” and “cancer” (including type and location) could be one value (or code) of many codes for various conditions. In reality, there are multiple approaches for system designers to semantically model clinical information using terminologies and clinical models [9]. Creation of clinical models and terminology bindings for a domain is a difficult, tedious, and time-consuming exercise that involves negotiation between multiple stakeholders. This complexity of terminology binding and the relative shortage of qualified terminologists make this semantic normalization an appropriate target for machine learning to accomplish some of the lower-level tasks of terminology binding.

The Clinical Information Modeling Initiative (CIMI) has been under development for more than 20 years and has recently been adopted by HL7 as an official working group. This CIMI group is creating a shared repository of detailed clinical information models for multiple application contexts, including EHR data storage and retrieval using standard APIs for decision logic, clinical trials data, and quality measures.

International Landscape and Coordination

The steps (planning, data structure, collection, preparation and transformation, exchange, storage, applications, and query) described above for data projects can be thought of as the building blocks of a health information system; and interoperability, from end-to-end, requires them to work together seamlessly. This requires communication, coordination, and cooperation between the people and organizations that develop, maintain, and promote those standards. Figure 19.2 presents most of the organizations and initiatives – national and international – that develop standards related to healthcare and health transactions. To realize the vision of interoperability and integrated research and health systems, many organizations – with many different needs – must work together to realize the vision we described in Box 19.2.

The dominant discussion forums for advancing applied clinical research data standards are CDISC and the HL7 Biomedical Research and Regulation (BR&R) working group. HL7, a not-for-profit volunteer organization, has the broadest scope

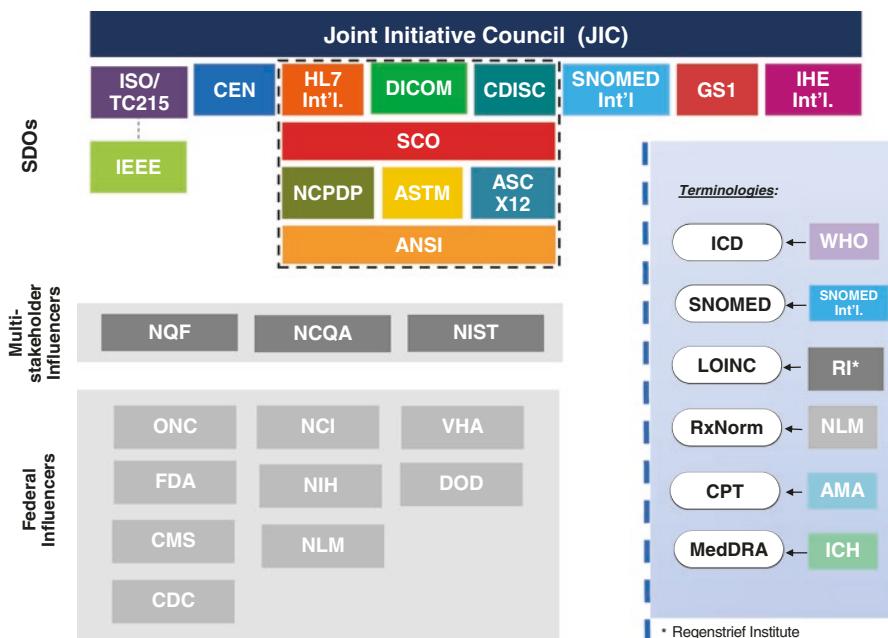


Fig. 19.2 The international standards landscape (acronyms defined in Appendix)

to produce standards for clinical and administrative data in all health settings and is an American National Standards Institute (ANSI)-accredited standards developing organization (SDO) [2]. Like all ANSI-accredited SDOs, HL7 adheres to a strict and well-defined set of operating procedures that ensures consensus, openness, and balance of interest.

The HL7 BR&R (formerly named the Regulated Clinical Research Information Management, RCRIM) working group and CDISC worked together to build a formal conceptual model of the research space called the Biomedical Research Integrated Domain Group (BRIDG). This model was developed in 2005 to link the CDISC data reporting models with the HL7 RIM. The BRIDG model provides a comprehensive conceptual model of the clinical research domain as a basis for harmonization across information model standards. The BRIDG Model supports many National Cancer Institute (NCI) research projects and is increasingly becoming recognized as a means to bring together different systems.

HL7 has worked with ISO and others to formalize and ballot standards that have been endorsed by the FDA. Examples include the Individual Case Safety Report [6], the Structured Product Labeling, annotated ECG, and Common Product Models. HL7 has produced standards for the exchange of genetic testing results and family history (pedigree) data, and many others are in development. The Digitize Action Collaborative is a cross-industry group established by the Institute of Medicine Genomics Roundtable to accelerate the goals of increasing clinical genetic IT support.

The many types of standards presented in this chapter have been created by a number of US and international standards bodies – sometimes working independently, sometimes working together, sometimes working competitively, sometimes working harmoniously. Often, there is an assumption of some master architect that has – if not a legal authority – a master conceptual model of how the pieces (data systems, data models, activities, and terminologies) of the health and research enterprises should fit together. This has not been the case with healthcare information systems to date, and collaborations such as the Standards Coordinating Organization (SCO) and the Joint Initiative Council (JIC) are designed to address this coordination. The need for EHR and patient data to flow across international borders makes it obvious that standards must be internationally used and hence require coordination and involvement from many countries. Given the scope of the task and the number of organizations and stakeholders involved, the challenges for meaningful standards are tremendous, but facing them is inevitable.

Figure 19.2 illustrates a number of standards developing organizations (SDOs) that exist and are creating standards and are meant to stimulate an appreciation for the number and types of organizations that will need to collectively work together to realize all of the components in the vision of truly integrated and interoperable research and clinical systems. The figure illustrates several different kinds of organizations. The standards developing organizations are international (CDISC, CEN, DICOM, GS1, HL7 International, IEEE, IHE International, ISO TC 215, and SNOMED International) and USA-based (ASC X12, ASTM E31, and NCPDP). The Joint Initiative Council is an international collaborative that encourages single, joint international standards. The SDO Charter Organization (SCO) is a similar-purposed US body promoting harmonization among US SDOs. HL7 and CDISC participate in both groups. IEEE and DICOM are both international SDOs, but only DICOM formally participates in the JIC or SCO. Both have a relationship with ISO and work effectively with the other SDOs. ANSI is a US standards regulating body; it does not create standards but through a set of rules and balloting processes approves standards as US standards. ANSI is also the US representative to ISO. ANSI also has been identified as the permanent certification body for the certification of EHR systems. The groups on the right maintain controlled terminologies that are both international (SNOMED, MedDRA, ICD) and domestic (LOINC, RxNorm, CPT) in scope. The other boxes represent US federal influencers as part of the Office of the National Coordinator (ONC), which drives programs toward nationwide EHR adoption and coordination. The National Institute for Standards and Technology (NIST), as part of the American Recovery and Reinvestment Act of 2009, has assumed a role in identifying and testing standards. Optimistically for the vision presented in Box 19.2, there is movement toward harmonization and cooperation among the different groups.

The European Standards body Comité Européen de Normalisation (CEN) created a standard EN 13606 (now ISO 13606 standard) that defines a data structure called *archetypes*. Archetypes are reusable clinical models of content and process, developed to provide a standard shared model of important clinical data as well as standard requirement for terminology. OpenEHR, an open-source organization

based in Australia, has created a number of archetypes that are increasingly being used worldwide. In a very separate organizational effort and distinctively different modeling approach, HL7 CIMI and ISO are creating *detailed clinical models* – data structures that also model discrete set of precise clinical knowledge for use in a variety of contexts, such as XML or JSON syntax. HL7 also creates standards for Common Message Element Terms (CMETS) and templates for a variety of uses. The Integrating the Healthcare Enterprise (IHE) has created structured documents in XDS for imaging diagnostic reports. A new relationship, called Gemini, between HIE and HL7 committed to working together on common causes.

Obviously, there is overlap in activities, and we are moving toward an era of increased communication. There are collaborative agreements between many of the organizations that show promise to reduce the overlap between terminologies and enable them to coevolve. Examples include coordination between LOINC and SNOMED CT and between SNOMED CT and ICD. As a harmonization effort between two SDOs, HL7 took the content of the ASTM Continuity of Care Record (CCR) standard for the exchange of patient summary data and implemented it in the HL7 Clinical Document Architecture (CDA) standard. This product, called the Continuity of Care Document (CCD), is essentially an implementation guide using the HL7 CDA standard.

The organizations represented in Fig. 19.2 represent most, but certainly not all, standards organizations in the picture. Undoubtedly, there are scores of professional societies and ad hoc groups defining content standards, and there are initiatives, such as the FDA Critical Path Initiative, that demand aggregation and sharing of data, integration of functionality, multiple uses of data without redundant, independent collection of data, and an overall perspective of the individual – independent of the clinical domain or disease – that can only be accomplished by an engaging and interoperable suite of standards.

Standards Influencers: Collaborative Initiatives Driving Efficiencies in Clinical Research

Developing complex standard is one thing, but applying them to address real data exchange needs of real clinical and business problems is quite another. In addition to the developers and sponsors of standards shown in the figure, there are a number of organizations and initiatives that are not official standards developers or sponsors but influence standards nonetheless. Most of these are collaborations of stakeholders that are frustrated with the current clinical research system and are banding together to share resources and advocacy toward a common solution. Several important examples are presented below:

The *Critical Path Institute (C-Path)* is a nonprofit, public-private partnership with FDA (created under the auspices of the FDA's Critical Path Initiative program in 2005) designed to accelerate medical product development through the creation of new data standards, measurement standards, and methods standards that aid in the scientific evaluation of the efficacy and safety of new therapies. This initiative is promoting collaboration for shared resources in the precompetitive space.

TransCelerate is a consortium of the largest biopharma and biomedical research companies with a goal of sharing noncompetitive and preclinical data to better the industry, to reduce the costs of drug discovery and production, and to promote a standard way to manage data and investigation. Data is contributed by each company to a standards-based data platform, based on CDISC models, with common and private areas for each company. TransCelerate is supporting pilot studies that use FHIR resources for the collection of patient-generated data and demonstrating the capabilities of SMART on FHIR for acquiring information directly from EHR.

The *Clinical Trials Transformation Initiative (CTTI)* is a public-private partnership established by the FDA and Duke University in 2007 to develop and drive adoption of practices that will increase the quality and efficiency of clinical trials. CTTI includes over 80 organizations from across the clinical trial enterprise. Members include representatives of government agencies (FDA, CMS, Office of Human Research Protections, NIH), industry representatives (pharmaceutical, biotech, device, and clinical research organizations), patient advocacy groups, professional societies, investigator groups, academic institutions, and other interested parties.

The *Integrating the Healthcare Enterprise (IHE)* is a current multi-organization initiative developed to address the global coordination of standards. IHE is led by the Healthcare Information and Management Systems Society (HIMSS) and Radiological Society of North America (RSNA) and includes dozens of EHR vendors to define profiles using suites of standards to advance end-to-end interoperability. In many countries, the government identifies or mandates which standards are used for what purposes. In the international scene, these profiles will require global governance. Multinational coordination processes are being developed in the European Union.

Argonaut is a private sector industry group assembled to promote the use of HL7 FHIR in healthcare through the specification of APIs and the production of implementation guides and security specifications. Argonaut has broad sponsorship and participation from large healthcare organization, technology vendors, EHR vendors, and academic institutions.

Da Vinci is a payer/provider consortium where Da Vinci stakeholders are industry leaders and health IT technical experts who are working together to accelerate the adoption of HL7 FHIR as the standard to support and integrate value-based care data exchange across communities.

The above initiatives provide just a few examples of collaboration around business-driven uses cases to implement and refine standards. They recognize the need for implementation and harmonization of multiple standards to solve important research and healthcare problems. These collaborations are funding projects to jumpstart demonstrations of interoperability and standards harmonization, but ultimately, the development and promulgation of standards will have to be supported by the government, business, or other activities. The continued cooperation and development of collaboratives like those listed here will be critical to develop standards in emerging areas such as mobile devices and sensors and wearables and social media or other data sources.

Standards Maintenance and Access

Standards are dynamic and need to be maintained. The maintenance process for any standard should be well documented and thoughtfully designed to allow the standard to evolve with the field and stay relevant and useful [10]. Commercial developers that incorporate standards into products must be permitted to receive the return on the investment before changes are introduced. If the currently implemented standard meets the need, it is unlikely that user will spend more money just to be up-to-date, hence the reason multiple versions of a standard are in use at one time.

Clinical research is complicated by the need to pick the best standards for the intended purpose and mapping between standards. Most recently, that issue has been further challenged by the fact that some standards are open source and are generally available without membership or a global license. An example is the International Patient Summary Implementation Guide (IG). Ideally, that IG would specify what data representation terminology would be. SNOMED CT would be a likely choice. Unfortunately, since some countries do not have a SNOMED license, that choice cannot be represented in the standard. Another case that limits collaboration is between ISO and HL7, as the HL7 standards are now open source and ISO standards have a cost. The challenge is to create a business model that will accommodate both strategies.

There is tension between making a standard freely available but also provide a quality standard with comprehensive, timeline, and useful documentation and information for new and experienced users. Open-source or free standards encourage use, but quality standards require resources to build. New and creative models for incentivizing the coordination and integration of healthcare and research standards are badly needed and represent a wide open area for informatics and clinical research experts.

Conclusion

The continued development and adoption of standards will be vital to achieve efficient clinical research processes that are integrated with healthcare systems and optimized to advance biomedical knowledge and its application to improve human health. Standards are needed for interoperable systems that can exchange data while preserving meaning and also are essential to enable accurate analytics, a core aim of research.

Like the Great Wall of China, the achievement of standardized and interoperable health and research information systems will take a shared vision, collaboration, and coordination. A consensus vision for efficient biomedical research can help mobilize coordinated standards to support the integration of clinical research and health information infrastructures. Progress toward this goal and the incremental steps to get there is an exciting aspect of clinical research informatics and will be for years to come.

Appendix 19.1: Standards Developing Organizations and Standards

Organizations and Initiatives

Accredited Standards Committee (ASC X12) – Develops electronic data interchange (EDI) standards and related documents for national and global markets. With more than 315 X12 EDI standards and a growing collection of X12 XML schemas, ASC X12 enhances business processes, reduces costs, and expands organizational reach. ASC X12's diverse member base includes 3000+ standards experts representing over 340 companies from multiple business domains, including communications, finance, government, insurance, supply chain, and transportation. Chartered in 1979 by the American National Standards Institute. <http://www.X12.org>.

American Health Information Management Association (AHIMA) – An association of health information management (HIM) professionals committed to advancing the HIM profession in an increasingly electronic and global environment through leadership in advocacy, education, certification, and professional education. AHIMA's more than 61,000 members are dedicated to the effective management of personal health information to support quality healthcare. Founded in 1928. <http://www.ahima.org>.

American Medical Association (AMA) – A voluntary association of physicians in the USA. It promotes the art and science of medicine and the betterment of public health. The American Medical Association helps doctors help patients by uniting physicians nationwide to work on the most important professional and public health issues. Founded in 1847. <http://www.ama-assn.org>.

American National Standards Institute (ANSI) – A not-for-profit organization that oversees the creation, promulgation, and use of thousands of norms and guidelines that directly impact businesses in nearly every sector, including acoustical devices, construction equipment, dairy and livestock production, energy distribution, and healthcare. ANSI is also actively engaged in accrediting programs that assess conformance to standards – including globally recognized cross-sector programs such as the ISO 9000 (quality) and ISO 14000 (environmental) management systems. ANSI is also the US representative to the ISO. Founded in 1918. <http://www.ansi.org>.

American Society for Testing and Materials (ASTM) – A globally recognized leader in the development and delivery of international voluntary consensus standards. Today, some 12,000 ASTM standards are used around the world to improve product quality, enhance safety, facilitate market access and trade, and build consumer confidence. Formed in 1898 by chemists and engineers from the Pennsylvania Railroad. <http://www.astm.org>.

European Committee for Standardization or Comité Européen de Normalisation (CEN) – A major provider of European standards and technical specifications. It is the only recognized European organization according to Directive 98/34/EC for the planning, drafting, and adoption of European standards in all areas of economic activity with the exception of electrotechnology

(CENELEC) and telecommunication (ETSI). The Vienna Agreement – signed by CEN in 1991 with ISO (International Organization for Standardization), its international counterpart – ensures technical cooperation by correspondence, mutual representation at meetings and coordination meetings, and adoption of the same text, as both an ISO standard and a European standard. Founded in 1961. <http://www.cen.eu/cen/pages/default.aspx>.

European Committee for Electrotechnical Standardization (CENELEC) – A nonprofit Belgian organization, CENELEC is responsible for standardization in the electrotechnical engineering field. CENELEC prepares voluntary standards, which help facilitate trade between countries, create new markets, cut compliance costs, and support the development of a European Single Market. Created in 1973. <http://www.cenelec.eu/index.html>.

European Telecommunications Standards Institute (ETSI) – A not-for-profit organization that produces globally applicable standards for information and communications technology. Their approach is one of openness and knowledge accessibility within standardization. Created in 1988. <http://www.etsi.org/web-site/homepage.aspx>.

Clinical Data Interchange Standards Consortium (CDISC) – A global, open, multidisciplinary, nonprofit organization that has established standards to support the acquisition, exchange, submission, and archive of clinical research data and metadata. *The CDISC mission is to develop and support global, platform-independent data standards that enable information system interoperability to improve medical research and related areas of healthcare.* CDISC standards are vendor neutral, platform independent, and freely available via the CDISC website. Began as a volunteer group in 1997. <http://www.cdisc.org/>.

Digital Imaging and Communications in Medicine (DICOM) – A joint committee formed from the American College of Radiology (ACR) and the National Electrical Manufacturers Association (NEMA) to create a standard method for the transmission of medical images and their associated information. The DICOM Standards Committee exists to create and maintain international standards for communication of biomedical diagnostic and therapeutic information in disciplines that use digital images and associated data. The actual *DICOM Standard* (currently in version 3.0) defines an upper layer protocol (ULP) that is used over TCP/IP (independent of the physical network), messages, services, information objects, and an association negotiation mechanism. These definitions ensure that any two implementations of a compatible set of services and information objects can effectively communicate. Committee formed in 1983. DICOM Standard versions released in 1995, 1988, and 1993. <http://medical.nema.org/>.

GS1 – An international not-for-profit association with member organizations in over 100 countries. GS1 is dedicated to the design and implementation of global standards and solutions to improve the efficiency and visibility of supply and demand chains globally and across sectors. The GS1 system of standards is the most widely used supply chain standards system in the world. Founded in 1977. <http://www.gs1.org/>.

Healthcare Information and Management Systems Society (HIMSS) – A cause-based, not-for-profit organization exclusively focused on providing global leadership for the optimal use of information technology and management systems for the betterment of healthcare. Its mission is to lead healthcare transformation through the effective use of health information technology. It was founded in 1961. <http://www.himss.org>.

Health Level Seven International (HL7) – A not-for-profit, ANSI-accredited standards developing organization dedicated to providing a comprehensive framework and related standards for the exchange, integration, sharing, and retrieval of electronic health information that supports clinical practice and the management, delivery, and evaluation of health services. HL7's 2300+ members include approximately 500 corporate members who represent more than 90% of the information system vendors serving healthcare. Founded in 1987. <http://www.hl7.org>.

Institute of Electrical and Electronics Engineers (IEEE) – The world's largest technical professional society and an association dedicated to advancing innovation and technological excellence for the benefit of humanity. It is designed to serve professionals involved in all aspects of the electrical, electronic, and computing fields and related areas of science and technology that underlie modern civilization. IEEE was established in 1963 as a merger of the Institute of Radio Engineers (founded in 1912) and the American Institute of Electrical Engineers (founded in 1884). <http://www.ieee.org>.

Integrating the Healthcare Enterprise (IHE) – An initiative by healthcare professionals and industry to improve the way computer systems in healthcare share information. IHE promotes the coordinated use of established standards such as DICOM and HL7 to address clinical need and support optimal patient care. Systems developed in accordance with IHE communicate with one another better, are easier to implement, and enable care providers to use information more effectively. <http://www.ihe.net>.

International Conference on Harmonisation (ICH) – ICH's mission is to make recommendations toward achieving greater harmonization in the interpretation and application of technical guidelines and requirements for pharmaceutical product registration, thereby reducing or obviating duplication of testing carried out during the research and development of new human medicines. Founded in 1990. <http://www.ich.org>.

International Health Terminology Standards Development Organisation (IHTSDO) – A not-for-profit association that develops and promotes the use of SNOMED CT to support safe and effective health information exchange. SNOMED CT is a clinical terminology and is considered to be the most comprehensive, multilingual healthcare terminology in the world. Formed in 2006. <http://www.ihtsdo.org/>. As of 2018, the organization is now called SNOMED International.

International Organization for Standardization (ISO) – The world's largest developer and publisher of international standards. Its network consists of 162 countries, coordinated by a general secretariat in Geneva, Switzerland. It is a non-governmental multinational organization that forms a bridge between public and private sectors. Founded in 1947. <http://www.iso.org/iso/home.html>.

Joint Initiative Council (JIC) – A harmonization process between standards development organizations (SDOs) to enable common, timely health informatics standards by addressing and resolving issues of gaps, overlaps, and counterproductive standardization efforts, particularly between ISO TC215 and HL7. The Council consists of leaders and appointed liaison members of the participating SDOs and strategically oversees the *Joint Initiative on SDO Global Health Informatics Standardization*. Currently, the participating SDOs are ISO/TC 215, HL7, CEN/TC 251, CDISC, IHTSDO, and GS1. The Charter was signed in 2007. <http://www.jointinitiativecouncil.org/>.

National Council for Prescription Drug Programs (NCPDP) – is a not-for-profit, ANSI-accredited standards development organization representing the pharmacy services industry. <http://www.ncpdp.org>.

National Quality Forum (NQF) – A nonprofit organization with a mission to improve the quality of American healthcare by building consensus on national priorities and goals for performance improvement and working in partnership to achieve them, endorsing national consensus standards for measuring and publicly reporting on performance, and promoting the attainment of national goals through education and outreach programs. NQF's membership includes a wide variety of healthcare stakeholders, including consumer organizations, public and private purchasers, physicians, nurses, hospitals, accrediting and certifying bodies, supporting industries, and healthcare research and quality improvement organizations. The NQF was established in 1999 in response to the recommendation of the Advisory Commission on Consumer Protection and Quality in the Health Care Industry, which concluded that an organization was needed to promote and ensure patient protections and healthcare quality through measurement and public reporting. <http://www.qualityforum.org>.

OpenEHR – An international, not-for-profit foundation working toward developing an interoperable, lifelong electronic health record. To this end, it is developing open specification, open-source software, and knowledge resources. It also participates in international standards development. <http://www.openehr.org>.

Professional societies, for example, the American College of Cardiology (ACC) – The American College of Cardiology is a nonprofit medical association of 39,000 members to advocate for quality cardiovascular care through education, research, development, and applications of standards and guidelines. It also works to influence healthcare policies. Established in 1949. <http://www.cardiosource.org/acc>.

Radiological Society of North America (RSNA) – The mission of the Radiological Society of North America is to promote and develop the highest standards of radiology and related sciences through education and research. The Society seeks to provide radiologists and allied health scientists with educational programs and materials of the highest quality and to constantly improve the content and value of these educational activities. The Society seeks to promote research in all aspects of radiology and related sciences, including basic clinical research in the promotion of quality healthcare. Founded in 1916 as the Western Roentgen Society, it was given its present name in 1919. <http://www.rsna.org>.

SDO Charter Organization (SCO) – Provides an environment that facilitates effective coordination and collaboration on US national healthcare informatics standards development. Among its purposes are to facilitate the coordination of conventions for enhanced interoperability among diverse standards development organizations in the areas of health data acquisition, processing, and handling systems and to communicate and coordinate when appropriate with the US Technical Advisory Group (US TAG) in order to facilitate a unified representation of US standards (this is not intended to supersede any member's existing coordination with the US TAG). Established in 2008. <http://scosummit.com/>; http://www.ncpdp.org/resources_sco.aspx.

SNOMED International – A not-for-profit association that develops and promotes use of SNOMED CT to support safe and effective health information exchange. SNOMED CT is a clinical terminology and is considered to be the most comprehensive, multilingual healthcare terminology in the world. The organization has over 29 member countries. It was founded in 2006 as the International Health Terminology Standards Development Organisation, IHTSDO.

World Health Organization (WHO) – WHO is the directing and coordinating authority for health within the UN system. It is responsible for providing leadership on global health matters, shaping the health research agenda, setting norms and standards, articulating evidence-based policy options, providing technical support to countries, and monitoring and assessing health trends. Established in 1948. <http://www.who.int/en>.

US Government Organizations Developing and Naming Standards

Centers for Disease Control and Prevention (CDC) – One of the major operating components of the Department of Health and Human Services. Its mission is to collaborate to create the expertise, information, and tools that people and communities need to protect their health – through health promotion, prevention of disease, injury and disability, and preparedness for new health threats. It began on July 1, 1946 as the Communicable Disease Center. <http://www.cdc.gov>.

Centers for Medicare and Medicaid Services (CMS) – Part of the Department of Health and Human Services, this agency is responsible for Medicare health plans, Medicare financial management, Medicare fee for service operations, Medicaid and children's health, survey and certification, and quality improvement. Founded in 1965. <http://www.cms.gov>.

Department of Defense (DOD) – The mission of the DOD is to provide the military forces needed to deter war and to protect the security of our country. Defense.gov supports the overall mission of the Department of Defense by providing official, timely, and accurate information about defense policies, organizations, functions, and operations, including the planning and provision of healthcare, health monitoring, and medical research, training, and education. Also, Defense.gov is the single, unified starting point for finding military information online. Created in

1789 as the War Department, in 1949 it became known as the Department of Defense. <http://www.defense.gov>.

The US Department of Health and Human Services (HHS) – The principal government agency for supervising the health of American citizens and providing essential human services, particularly for vulnerable populations. Representing almost a quarter of all federal outlays, it administers more grant dollars than all other federal agencies combined, including the Medicare and Medicaid healthcare insurance programs. HHS programs are directed by the Office of the Secretary and administered by 11 operating divisions, including 8 agencies in the US Public Health Service and 3 human services agencies. The department includes more than 300 programs, which provide health services, support equitable treatment of recipients nationwide, and enable national health and data collection. Originally founded in 1953 as the Department of Health, Education, and Welfare (HEW), it was officially renamed in 1979. <http://www.hhs.gov/>.

Department of Homeland Security (DHS) – With the passage of the Homeland Security Act by Congress in November 2002, the Department of Homeland Security formally came into being as a stand-alone, cabinet-level department to further coordinate and unify national homeland security efforts, opening its doors on March 1, 2003. The DHS has five departmental missions: to prevent terrorism and enhance security, to secure and manage our borders, to enforce and administer US immigration laws, to safeguard and secure cyberspace, and to ensure resilience to disasters. <http://www.dhs.gov>.

Federal Health Architecture (FHA) – An E-Government Line of Business initiative managed by the United States' Office of the National Coordinator for Health IT. FHA was formed to coordinate health IT activities among the more than 20 federal agencies that provide health and healthcare services to citizens. FHA and its federal partners are helping build a federal health information technology environment that is interoperable with private sector systems and supports the president's plan to enable better point-of-service care, increased efficiency, and improved overall health in the US population. <http://www.hhs.gov/fedhealtharch>.

Food and Drug Administration (FDA) – An agency within the US Department of Health and Human Services, it is responsible for protecting the public health by assuring the safety, effectiveness, and security of human and veterinary drugs, vaccines, and other biological products, medical devices, the nation's food supply, cosmetics, dietary supplements, and products that give off radiation. Though FDA can trace its history back to the appointment of chemist Lewis Caleb Beck to the Agricultural Division in the Patent Office in 1848, its origins as a federal consumer protection agency began with the passage of the 1906 Pure Food and Drugs Act. This law was the culmination of about 100 bills over a quarter-century that aimed to rein in long-standing, serious abuses in the consumer product marketplace. <http://www.fda.gov>.

National Cancer Institute (NCI) – The National Cancer Institute (NCI) is part of the National Institutes of Health (NIH), which is one of 11 agencies that compose the Department of Health and Human Services (HHS). The NCI, established under the National Cancer Institute Act of 1937, is the federal government's principal

agency for cancer research and training. The National Cancer Act of 1971 broadened the scope and responsibilities of the NCI and created the National Cancer Program. Over the years, legislative amendments have maintained the NCI authorities and responsibilities and added new information dissemination mandates as well as a requirement to assess the incorporation of state-of-the-art cancer treatments into clinical practice. <http://www.cancer.gov>.

National Institute for Standards and Technology (NIST) – A nonregulatory federal agency within the US Department of Commerce. Its focus is on promoting innovation and industrial competitiveness by advancing measurement science, standards, and technology in ways that enhance economic security and improve our quality of life. The NIST also managed the Advanced Technology Program between 1990 and 2007 to support US businesses, higher education institutions, and other research organizations in promoting innovation through high-risk, high-reward research in areas of critical national need. Founded in 1901. <http://www.nist.gov>.

National Institute of Neurological Disorders and Stroke (NINDS) – Part of the NIH, NINDS conducts and supports research on brain and nervous system disorders. It also supports training of future neuroscientists. Created by Congress in 1950. <http://www.ninds.nih.gov>.

National Institutes of Health (NIH) – A division of the US Department of Health and Human Services and the primary agency of the US government responsible for biomedical and health-related research. The purpose of NIH research is to acquire new knowledge to help prevent, detect, diagnose, and treat disease and disability by conducting and supporting innovative research, training of research investigators, and fostering communication of medical and health sciences information. The NIH is divided into “extramural” divisions, responsible for the funding of biomedical research outside of NIH, and “intramural” divisions to conduct research. It is headed by the Office of the Director and consists of 27 separate institutes and offices. It was initially founded in 1887 as the Laboratory of Hygiene but was reorganized in 1930 as the NIH. <http://www.nih.gov>.

The US National Library of Medicine (NLM) – Located in the National Institutes of Health, a division of the US Department of Health and Human Services. The NLM is the world’s most extensive medical library with medical and scientific collections which are comprised of books, journals, technical reports, manuscripts, microfilms, and images. It also develops electronic information services, including the free-access PubMed database and the MEDLINE publication database. The NLM provides service scientists, health professionals, historians, and the general public both nationally and globally. Originally founded in 1836 as the Library of the Office of the Surgeon General of the Army, it has been restructured multiple times before finally reaching its current configuration in 1956. <http://www.nlm.nih.gov>.

Office of the National Coordinator for Health Information Technology (ONC) – Located within the US Department of Health and Human Services as a division of the Office of the Secretary. It is the nationwide coordinator for the implementation of new advances in health information technology to allow electronic use and exchange of information to improve healthcare. Prior to 2018, the ONC made recommendations on standards, implementation specifications, and certification

criteria through two federal advisory committees, the *Health IT Policy Committee (HITPC)* and the *Health IT Standards Committee (HITSC)*. The HITPC developed a policy framework for the development and adoption of a nationwide health information infrastructure, including standards for the exchange of patient medical information. The HITSC developed a schedule for the annual assessment of the HITPC's recommendations and advised on testing of standards and implementation specifications by the National Institute for Standards and Technology (NIST). The position of national coordinator was created through an executive order in 2004 and legislatively mandated in the Health Information Technology for Economic and Clinical Health Act (HITECH Act) of 2009. The *Health Information Technology Advisory Committee (HITAC)* was established in the 21st Century Cures Act and will recommend policies, standards, implementation specifications, and certification criteria, relating to the implementation of an infrastructure that will advance the electronic access, exchange, and use of health information. HITAC unifies the roles of, and replaces, the HITPC and the HITSC. <http://healthit.hhs.gov/>.

Veterans Health Administration (VHA) – Component of the US Department of Veterans Affairs that implements the medical assistance program through the administration and operation of numerous outpatient clinics, hospitals, medical centers, and long-term care facilities. The first VHA hospital dates back to 1778. <http://www.va.gov/health/default.asp>.

Controlled Terminologies (Standards)

Current Procedural Terminology (CPT) – A registered trademark of the American Medical Association (AMA), CPT codes are used in medical billing to describe medical, surgical, and diagnostic services and are designed to communicate uniform information about medical services and procedures for administrative, financial, and analytic purposes. <http://www.ama-assn.org>.

International Classification of Diseases (ICD) – The classification used to code and classify mortality data from death certificates. The International Classification of Diseases, Clinical Modification is used to code and classify morbidity data from the inpatient and outpatient records, physician offices, and most National Center for Health Statistics (NCHS) surveys. In 1893, a French physician, Jacques Bertillon, introduced the Bertillon Classification of Causes of Death at the International Statistical Institute in Chicago. A number of countries adopted Dr. Bertillon's system, and in 1898, the American Public Health Association (APHA) recommended that the registrars of Canada, Mexico, and the USA also adopt it. Since 1959, the US Public Health Service published several versions of this classification system which is the standard to code diagnostic and operative procedural data for official morbidity and mortality statistics in the USA. It is currently in its tenth edition. <http://www.cdc.gov/nchs/icd.htm>.

Logical Observation Identifiers Names and Codes (LOINC) – A universal code system for identifying laboratory and clinical observations. Mapping local terms to LOINC makes it possible to exchange and pool data from many

independent systems for clinical care, research, outcomes management, and lots of other purposes. Initiated in 1994 and maintained by the Regenstrief Institute. <http://loinc.org>.

Medical Dictionary for Regulatory Activities (MedDRA) – A terminology that applies to all phases of drug development, excluding animal toxicology. It also applies to the health effects and malfunction of medical devices. It was developed by the International Conference on Harmonisation (ICH) and is owned by the International Federation of Pharmaceutical Manufacturers and Associations (IFPMA) acting as trustee for the ICH Steering Committee. MedDRA is used to report adverse event data from clinical trials and for postmarketing reports and pharmacovigilance. <http://meddramsso.com/index.asp>.

RxNorm – Provides normalized names for clinical drugs and links its names to many of the drug vocabularies commonly used in pharmacy management and drug interaction software, including those of First DataBank, Micromedex, Medi-Span, Gold Standard Alchemy, and Multum. By providing links between these vocabularies, RxNorm can mediate messages between systems not using the same software and vocabulary. RxNorm now includes the National Drug File – Reference Terminology (NDF-RT) from the Veterans Health Administration. NDF-RT is a terminology used to code clinical drug properties, including mechanism of action, physiologic effect, and therapeutic category. <http://www.nlm.nih.gov/research/umls/rxnorm>.

Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT) – A comprehensive clinical terminology, originally created by the College of American Pathologists (CAP) and, as of April 2007, owned, maintained, and distributed by SNOMED International (formerly the International Health Terminology Standards Development Organisation, IHTSDO), a not-for-profit association. http://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html.

Resources

NIH Common Data Element (CDE) Resource Portal – NIH encourages the use of common data elements (CDEs) in clinical research, patient registries, and other human subject research in order to improve data quality and opportunities for comparison and combination of data from multiple studies and with electronic health records. This portal provides access to information about NIH-supported CDEs, as well as tools and resources to assist investigators developing protocols for data collection. <https://www.nlm.nih.gov/cde/>

Cancer Data Standards Registry and Repository (caDSR) – Database and a set of APIs (application programming interfaces) and tools to create, edit, control, deploy, and find common data elements (CDEs) for use by metadata consumers and information about the UML models and forms containing CDEs for use in software development for research applications. Developed by National Cancer Institute for Biomedical Informatics and Information Technology. <https://cabig.nci.nih.gov/concepts/caDSR>.

National Center for Biomedical Ontology (NCBO Bioportal) – An open repository of biomedical ontologies. Supports ontologies in OBO, OWL, RDF, Rich Release Format (RRF), Protégé Frames, and LexGrid XML. The goal of the NCBO is to support biomedical researchers by providing online tools and a Web portal, enabling them to access, review, and integrate disparate ontological resources in all aspects of biomedical investigation and clinical practice. Funded by the US NIH and National Centers for Biomedical Computing. Created in 2007. <http://www.bioontology.org>.

National Drug File-Reference Terminology (NDF-RT) – An extension of the VHA National Drug File (NDF). It organizes the drug list into a formal representation and can be considered as a knowledge base or ontology for classifying drugs and medication products. NDF-RT is used for modeling drug characteristics including ingredients, chemical structure, dose form, physiologic effect, mechanism of action, pharmacokinetics, and related diseases. http://bioportal.bioontology.org/ontologies/40402?p=terms#40402?p=summary&_suid=426.

Unified Medical Language System (UMLS) – A set of files and software that brings together many health and biomedical vocabularies and standards to enable interoperability between computer systems. UMLS can be used to enhance or develop applications, such as electronic health records, classification tools, dictionaries, and language translators. The UMLS has three tools, which are called the Knowledge Sources:

- *Metathesaurus*: Terms and codes from many vocabularies, including CPT®, ICD-10-CM, LOINC®, MeSH®, RxNorm, and SNOMED CT®
- *Semantic network*: Broad categories (semantic types) and their relationships (semantic relations)
- *SPECIALIST Lexicon and Lexical Tools*: Natural language processing tools

Created in 1986. <http://www.nlm.nih.gov/research/umls>.

References

1. ICH. Information paper. Step 3 Release E2B(R3). Revision of electronic submission of individual case safety reports: status and regional requirements update. Geneva; 2011.
2. CDISC. Therapeutic area standards. 2018. [cited 2018 July 1]. Available from: <https://www.cdisc.org/standards/therapeutic-areas>.
3. Richesson RL, Fung KW, Krischer JP. Heterogeneous but “standard” coding systems for adverse events: issues in achieving interoperability between apples and oranges. *Contemp Clin Trials*. 2008;29(5):635–45.
4. Hammond WE, et al. Integration of a computer-based patient record system into the primary care setting. *Comput Nurs*. 1997;15(2 Suppl):S61–8.
5. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc AMIA Symp*. 2001:17–21.
6. PCORnet. PCORnet common data model (CDM). Why, what, and how? 2015 [cited 2015 Aug 30]. Available from: <http://www.pcornet.org/pcornet-common-data-model/>.

7. OHSDI. OMOP common data model. 2015. [cited 2015 August 30]. Available from: <http://www.ohdsi.org/data-standardization/the-common-data-model/>.
8. Rijnbeek PR. Converting to a common data model: what is lost in translation?: commentary on “fidelity assessment of a clinical practice research datalink conversion to the OMOP common data model”. *Drug Saf.* 2014;37(11):893–6.
9. Chute CG. Medical concept representation. In: Chen H, et al., editors. *Medical informatics. Knowledge management and data mining in biomedicine*: Springer, New York, U.S; 2005. p. 163–82.
10. Oliver DE, et al. Representation of change in controlled medical terminologies. *Artif Intell Med.* 1999;15(1):53–76.



Back to the Future: The Evolution of Pharmacovigilance in the Age of Digital Healthcare

20

Michael A. Ibara and Rachel L. Richesson

Abstract

Pharmacovigilance originated in an attempt to better understand the safety of drugs so that we can protect individual patients and consumers. Over time the development of the field has been heavily influenced by the need for the pharmaceutical industry to fulfill regulatory requirements, with the unintended result of losing track of the individual patient. With the onset of digitized healthcare data, we have an opportunity to reunite the industrial and personal in pharmacovigilance. Informatics can help with this by focusing future work on a pharmacovigilance research agenda.

Keywords

Pharmacovigilance · Informatics · Adverse drug events · Postmarketing surveillance · Pharmacoepidemiology · Quantitative signal detection · Risk management plans

Introduction

This chapter seeks to provide a foundation for future work in pharmacovigilance for the informatician involved in clinical research. It will not attempt to provide an overview of the field of pharmacovigilance, as this has been covered extensively elsewhere (see below) [15, 19]. The focus here will be on key developments in

M. A. Ibara, PharmD (✉)
Elligo Health Research, Princeton, NJ, USA
e-mail: michael.ibara@elligodirect.com

R. L. Richesson, PhD, MPH, FACMI
Division of Clinical Systems and Analytics, Duke University School of Nursing,
Durham, NC, USA

Duke Center for Health Informatics, Durham, NC, USA
e-mail: rachel.richesson@dm.duke.edu

pharmacovigilance and related areas as a result of the growing digitization of health-care data. We will propose an informatics research agenda meant to move the field forward and provide for a more holistic consideration of patient safety.

Pharmacovigilance is defined by the World Health Organization (WHO) [61] as “the science and activities relating to the detection, assessment, understanding and prevention of adverse effects or any other drug-related problem.” Pharmacovigilance is a central practice for understanding and assuring drug safety. For an excellent history of the development of pharmacovigilance as a discipline and the general applicability of informatics, see the previous edition of this chapter [19] in which the authors provide a superb primer for those wishing to gain a better understanding of the topic. A full treatment of the historical, regulatory, industrial, statistical, and medical aspects of the field can be found in several excellent reference works on the topic, especially *Stephens' Detection and Evaluation of Adverse Drug Reactions: Principles and Practice* and *Mann's Pharmacovigilance* [2, 55].

Background

Pharmacovigilance originated as an attempt to better understand the safety of drugs in order to protect individual patients and improve medicine. But while today pharmacovigilance plays a key and vital role in the research and public health arena, to the uninitiated it can seem bureaucratic, arcane, and arbitrary. This is due mainly to the myriad influences on the field from medicine, public health, industry, and regulation, as well as from broad interest in the topic by patients and practitioners, academic and industry researchers, and regulatory and legal bodies – all groups who have a stake in the endeavor. Over time, pharmacovigilance has taken on the shape of these combined influences, and their often disparate demands have led to a balkanization of the original pharmacovigilance landscape. Today, what a biopharmaceutical industry professional would describe as the daily work of pharmacovigilance would be unrecognizable to the layman or even to healthcare researchers in safety not otherwise engaged with industry.

For a number of years, the most significant forces of differentiation in pharmacovigilance were (and remain) the regulatory and legal requirements to which drug and device manufacturers must comply (hence the often-quoted statement by industry professionals that “compliance” is their first priority). And while healthcare practitioners are subjected to significant regulations and laws as well, a difference in focus and content means that “drug safety” in a healthcare or academic research setting has come to mean something quite different from the industrial use of the term. As the field developed over the last 50 years, the patient was seen as the recipient of any learning and good practices in research on safety of drugs and medical devices but was only taken seriously as a participant at the level of their individual healthcare provider. Both industrial and academic researchers saw the patient more as a source rather than a collaborator in their own health and well-being.

The result of these trends is that, today, pharmacovigilance looks very much like the rest of healthcare: siloed and having difficulty in interoperating with other healthcare components. With separate standards, processes, systems, and data

stores, various practitioners of “drug safety” work on their individual agendas, not noticing or acknowledging that they share (or could share) the same data with researchers in other fields of pharmacovigilance. But today, the increasing digitization of healthcare data is challenging this compartmentalization as it becomes possible to have a single data source serve a host of downstream practitioners and researchers, as well as the empowered patient.

What is less obvious, but we argue even more significant, is that the digitization of healthcare data creates the possibility for a return to the original aspirations of the field – where we can recapture the original goals of pharmacovigilance and reunite the individual, population, academic, and industrial pursuits to an extent that benefits all stakeholders, but most especially which allows us to realize one of the original goals of pharmacovigilance: to protect the individual while contributing to greater understanding at a population level. Practitioners in academic, medical, and industrial settings are finding themselves more often than not pursuing and working with the same data from the same sources. It is encouraging to imagine that they will also work on research topics that will help to reunify the field of pharmacovigilance and move it forward.

To support the thesis that the digitization of healthcare data creates opportunities to unify the field of pharmacovigilance, it is helpful to use an approach that has been applied widely to other industries undergoing digitization of their core content but to date has not been used to understand pharmacovigilance. To this end, we will examine the development of the field through the narrowly focused lens of Coasian economics.

Coasian Transactions: The Development and Evolution of PV/Drug Safety

In 1991 Ronald Harry Coase, a British economist and author, won the Nobel Memorial Prize in Economic Sciences in part for work outlined in his paper “The Nature of the Firm” (1937), where Coase introduces the concept of transaction costs to explain the nature of firms and how they behave in the marketplace [44]. Coase’s ideas were later applied to explain Internet economics [37, 42]. When transaction costs are relatively expensive, it is economical to house everything in a single firm (a single vertical) as that facilitates coordination and handoffs. But, as the transactions costs of handling data become cheaper (due to digitization), according to Coase we should expect new business models to develop as the cost of working horizontally (across different companies) becomes cheaper than that of working in verticals.

The application of this theory to pharmacovigilance is through the transaction costs associated with adverse events (AEs). Because a large component of pharmacovigilance is concerned with understanding how drugs (and devices) may cause AEs, the field seeks to identify, collect, process, analyze, and distribute this information. We can think of the steps in this process as the transaction costs in pharmacovigilance. Prior to digitization and the Internet, transaction costs to obtain information on AEs were relatively quite expensive. If a patient happened to mention a problem to their doctor, the doctor would need to interrupt their workflow to find and fill out a paper form and then somehow get that form to the FDA in the USA or appropriate regulator in another country.

Needless to say this has never been a strong avenue for AE reporting. In the past, pharmaceutical manufacturers were the only organizations able to deploy enough resources through site monitoring, call centers, and education to reliably collect AEs, and they were also the only organizations able to gather enough professionals to process, analyze, and distribute the information. Hence, the verticals (pharma companies) managed the transaction costs of AEs by housing the operations internally. However, as healthcare data is digitized, the transaction costs associated with finding, collecting, processing, analyzing, and distributing AEs decrease dramatically. All of the online data collection techniques (online forms, mobile reporting, scraping websites, etc.) can be applied here. And we can much more easily collect AE-specific data directly from individuals.

When you examine the developments in pharmacovigilance over the last 10 years, this is, in fact, what we see. One of the first signs of a coming change in the business model of pharmacovigilance was iGuard, a consumer-facing prescription drug-risk monitoring service that used digitized AE data and consumer-reported data to attempt to provide drug-risk monitoring services to consumers [47]. As new online services for doctors, patients, and consumers came online, the ability to digitally capture online postings that related to AEs and drugs became straightforward and quite inexpensive.

In 2015 FDA and the online patient community PatientsLikeMe signed a research agreement to explore how patient-reported data can provide useful AE and drug safety insights [40]. This was made possible by the fact that PatientsLikeMe has an online system for patient reporting of AEs and houses other patient information in their online system – the data is fully digitized. This agreement can be seen as the culmination of work begun several years earlier focused on the digitization of pertinent safety data from the PatientsLikeMe community.

These are two examples which signaled the beginning of a shift from an environment where AEs were hard to find and process, and so were scarce, to one in which, because the transaction costs of AEs were negligible, they could be discovered, processed, and distributed at a pace never before seen. Today, the number of possible sources for AEs continues to grow – from social media, electronic health records, registries, mobile devices, sensors, etc. There is no reason to expect this trend will not continue. It is generally acknowledged that we are in a world of growing sources and data, but it is less often recognized that this entails a shift in our approach to what was once an expensive and scarce resource. The fact that healthcare data and the sources for safety-related data are now abundant requires us to reexamine the way in which we've thought about the pharmacovigilance practices, systems, and regulations that have been developed at time when it was costly to obtain safety information and AEs were scarce. As Herbert Simon said, "A design representation suitable to a world in which the scarce factor is information may be exactly the wrong one for a world in which the scarce factor is attention" [58] p.144.

The Coasian development of pharmacovigilance can be outlined as follows:

1. Historically, pharmacovigilance was largely developed by vertical organizations having the resources to find, collect, and process safety information – drug and device manufacturers.

2. These organizations were the de facto owners of safety information and responsible for it (focus of regulations) because they were the only organizations able to afford the transaction costs.
3. As healthcare data has become digitized, there has been a dramatic lowering of the “transaction cost” of finding, collecting, and reporting safety information.
4. The movement of AE transaction costs toward zero means that the economic incentives to maintain vertical organizations for pharmacovigilance will no longer be present.
5. With AE transaction able to be horizontally (across different organizations), this creates an environment where new business models and opportunities are encouraged.

In pharmacovigilance these developments have brought us to the challenge presciently defined by Simon – the need to design our systems based on an abundance of safety data rather than a scarcity. If we view pharmacovigilance through a Coasian lens, we see that not only what we call adverse events but also related healthcare data which may impact our assessments, or which can be used in novel ways to improve our ability to practice pharmacovigilance, will continue to increase in number, and at an increasing rate, for the foreseeable future. The examples provided here show a clear momentum toward looking beyond traditional pathways to discover tools, techniques, and concepts that will allow us to take advantage of the growing body of data at our disposal and to improve the practice of pharmacovigilance.

The challenge for us is to unify (or reunify) the very different professional guilds that have developed as previously described. While it is tempting to imagine that new techniques or methods will simply wipe away traditional practices, this is rarely the case even scientific revolutions [31] let alone a field that has the complexities of healthcare entwined with the economics of industry and regulatory concerns. While it is beyond the scope of this chapter, an examination of the potential gains in health and economic terms to be achieved from a unification of the field across these areas is motivation enough to hold this out as a goal.

What follows is a proposed research agenda which concentrates on a few areas (1) that provide common ground among researchers, industry professionals, and regulators; (2) in which technological advances are beginning to provide significant advances; and (3) in which research informaticists can provide major contributions and guidance.

Research Program/Agenda

A Note on Machine Learning

Over the last few years, as computing power has reached sufficient levels and research has matured, there has been an explosion in the application of machine learning techniques to many areas in healthcare and pharmaceutical research [8, 10, 26, 64].

Such is the meteoric rise in the use of machine learning and algorithmic computation across healthcare and research that research topics 3, 4, and 5 here are largely concerned with the impact in these areas, whereas just a few years ago, they would be mentioned in passing.

It is no longer possible to approach a research agenda for pharmacovigilance without careful consideration of how these techniques and technologies are changing what is possible. But while their impact is considered here in light of their impact on the field, this chapter makes no attempt to evaluate specific techniques in machine learning or artificial intelligence, except as they apply to the specific research topics listed.

Topic 1: The Operational Definition of an Adverse Event

The regulatory definition of an adverse event (AE)¹ is well-established, with the term coming into common use in the 1930s and being refined in the 1960s and 1970s, at the same time that formal pharmacovigilance systems began to be established [55]. There has been a refinement of the term since then, but the general definition has remained fairly stable. For our purposes, what is important to note is that the definition of an AE was conceived at a time when the Internet, social media, big data, and the promise of large amounts of digital healthcare data were nascent or nonexistent. The most important effect this has had on the definition of an AE is to cast it in terms of a paper metaphor – we picture in our minds collecting AEs onto forms, and we think of the various elements of the form, the amount of information to be collected, and the location of what type of information should go together, all in terms of a piece of paper. The insidious use of this metaphor encourages a habitual mode of thought which, having been ossified in regulatory definitions, is hard to escape. And while the metaphor has been extended significantly, initially to cover copies and facsimiles and later to include the concept of electronic data stores, the impact of the Internet and the wholesale digitization of healthcare data have stretched the paper metaphor to its limit. It is past time for a reexamination of the fundamental definitions of the field.

The need to update our concepts in regards to how we define AEs becomes evident when we seek to operationalize the definition of an AE in order to implement it into systems and use it for research. The classic operational definition derived originally from regulatory use is that a valid adverse event report has “four elements”: an identifiable patient, an identifiable reporter, a suspect drug, and a serious adverse event or fatal outcome [41]. Over time the requirements for a regulatory report (which were created to help busy doctors understand what to report on a piece of paper) have become conflated with the definition of an AE, to the point where we might define a report that is missing these elements as irrelevant. But when we understand that the “4 elements” are simply an operational definition meant to assist

¹Those familiar with the use of the term “ADR” (adverse drug reaction) vs “AE” (adverse event) should note that this discussion does not attempt to differentiate between those stricter definitions. Here the term “AE” is meant to be used in a general sense of a reported or noticed problem or concern.

doctors in reporting, we can see that, given the digitization of healthcare data today, there is a need for a new operational definition.

An example illustrates the difficulties that arise from the mismatch of our concepts and the digital reality today in healthcare. In 2010 a pilot study demonstrated for the first time that it was possible to collect AEs at the point of care directly from an electronic health record, with minimal impact on clinicians, and to have those events sent electronically to FDA, in a matter of minutes after the initial recognition of the event [32]. At the time this study was performed, one of the authors engaged in fierce debate with industry colleagues over the fact that the individual physician's name was masked on the report (although the medical institution was known) and therefore the report was not a "qualified" AE (personal communication). This arcane argument took place as a result of an outdated operational definition for an AE, so that even though we could infer the existence of an individual physician given the design and operation of the electronic health record, the exact requirement of an "identifiable reporter" could be interpreted to mean the report was disqualified.

Healthcare research has no such operational definition for what constitutes an AE, and while this allows for a more rational approach to collecting medically relevant information, it means that there can be no direct sharing of approaches or interpretation of findings between the different sectors. And the reason such operational definitions are required by regulators and industry is that there are massive efforts which span companies and continents, which require some semblance of uniformity if the attempts to perform pharmacovigilance are to yield useful results.

Given that both sectors have an interest in AEs, it would be of great benefit if a more inclusive, subtle, and encompassing operational definition of an AE could be developed. Informaticians seeking to make progress here could begin with sound medical concepts to define the broadest category of adverse events. Clearly this work should be built on existing useful clinical models and ontologies (a topic discussed later), but an understanding of the regulatory definitions will be important as well. The goal would be to create a continuum of definitions based on informatics rather than the incongruous set of definitions that exist today. In this way we can imagine that AEs of "regulatory interest" would be a subset of a larger group of medical interest.

It could be argued that this distinction exists today – AEs collected as a matter of course in healthcare are examined to see if they meet regulatory criteria, and if so, they are classed as such. The problem with this approach is that using the outdated "four elements" to define AEs of regulatory interest ignores a significant number of medically interesting events. The time has come to rework the operational definition to better align with what qualifies today as an AE from work being done by researchers in healthcare.

Topic 2: Expanding and Formalizing the Data Model

Similar to the operational definition of an AE, the data model used to report AEs was developed from a need by regulators to have industry be able to report, in a consistent manner, AE reports. The original document of the 1996 document from the International Council of Harmonization (ICH) that addressed the "Data Elements

for Transmission of Individual Case Safety Reports” was designated “E2” (the ICH designation for pharmacovigilance documents) and “B” referred to the particular document that defined data elements [49]. Hence, when referring to “E2B,” we are referring to the underlying data model for an AE.

The E2B data model is well-developed and used internationally, which is an advantage. But as is the case with the operational definition of an AE, E2B had its origins long before big data, the Internet, and the dramatic increase in digitized healthcare data. With the most recent version (E2BR3), the overall standard is based upon a HL7 ICSR model that is capable of supporting the exchange of messages for a wide range of product types (e.g., human medicinal products, veterinary products, medical devices). This is an excellent move toward more functionality within the regulatory reporting realm, but whereas this works well to allow submission of AEs to regulators, from an informatics perspective, looking to the future support of research across healthcare, this is lacking.

Contrast this with the type of large-scale research done today using very large and disparate datasets. This work has driven the creation of common data models which often include adverse events. A good example of this is the Observational Medical Outcomes Pilot (OMOP) common data model (CDM) [38] produced by OHDSI (Observational Health Data Sciences and Informatics). The OMOP CDM was created to use in the systematic analysis of disparate observational databases, and to this end it has a common format and common terminologies, vocabularies, and coding schemes.

Use of this approach in pharmacovigilance is what Koutkias and Jaulent have called the “computational approach” [29], in this case specifically for signal detection. The authors argue that pharmacovigilance should exploit all possible sources of information that may impact drug and device safety, and they do an excellent job of reviewing the sources, tools, and approaches. Most importantly, they suggest that semantic technologies are the right approach to this new pursuit of using diverse data sources in a unified fashion.

One semantic technology increasingly popular in clinical informatics is ontologies – explicit, formal specifications of terms or concepts in a domain and the relationships among them [14]. An early introduction of ontologies to the field of pharmacovigilance came in 2006 when Henegar et al. looked at formalizing MedDRA, the standardized medical terminology used for international regulatory purposes, one of which is to report AEs [17]. What Henegar discovered with MedDRA is illustrative of many models and terminologies in use with pharmacovigilance – there were no formal definitions of terms in MedDRA, and this meant that no formal description logic could be applied to reason against data described with this terminology. The lack of formal logic and rigorous concept representation meant that inference was not possible based on semantic content.

For many years, those engaged in pharmacovigilance research in industry were well aware of the lack of a semantic layer, but it was considered simply an artifact of the way in which data was collected. Groupings and counts of terms in MedDRA were gathered, and what then followed was a long and arduous process of in effect manually applying the semantic layer back to the data. Ontologies have been demonstrated to significantly improve this situation and allow us to imagine the ability to combine large and disparate sources of data and properly infer from them [17, 29, 39, 46].

The challenge today is that there is still relatively sparse communication between the regulatory-facing tools used in pharmacovigilance and those being borrowed from computational biology and other disciplines allowing us to expand the data sources and techniques used in researching the safety of medical products. The Salus study [66] took on the challenge of harmonizing data models and terminologies in an effort not typical in signal verification studies. This approach holds great promise and engenders a significant amount of research, but Salus was unusual in that the authors sought to harmonize the work with regulatory requirements. To achieve this, in addition to creating a rich ontology to work with the EHR, they mapped certain elements onto the previously described reporting standard, E2B (R2). And while this was an effective demonstration that it is possible to unify the healthcare, industry, and regulatory needs in pharmacovigilance (by seeking a logical lower-level ontological representation), the fact that now a major revision to E2B (R3) is coming into effect and demonstrates the continued balkanized nature of the field.

Work by informaticists is needed to unify and maintain the representations needed in pharmacovigilance, and settling on a set of key ontologies would be a dramatic step forward and would enable better utilization of diverse sources of data, more economical translation of data for industrial research, and more accurate, better quality communication of this information for regulatory purposes.

Topic 3: Terminologies

Since the beginning of medical and industrial research, terminologies have been developed in an attempt to categorize and standardize work. And it has long been recognized that the problem of semantics, or the meaning of terms in medicine and healthcare research, cannot be fully divorced from the terminologies used to describe things [9, 50]. Along with heterogeneous data models, lack of consistency in various terminologies and how they're applied has been a challenge even before described succinctly by Cimino and is understood as a lynchpin to using EHRs for big data research [48].

Recently, the work being done in machine learning, ontologies, and computational methods is shedding new light on ways to tame the terminology issues, such that it is now imaginable that the problem of inconsistency could be solved by a logically rigorous ontology which binds terminologies to data models [11]. As a discussion of ontologies preceded this section, here we highlight work being done in machine learning which impacts challenges with terminologies.

For the last several years, researchers have looked at computer-assisted ways to extract AEs from text (specifically from narratives in AE reports) [30], but more recently new levels of sophistication in handling terminology as part of the process has been demonstrated. Jiang et al. evaluated using machine-learning-based approaches to extract clinical entities from hospital discharge summaries written in free text [24]. Clinical entities included medical problems, tests, and treatments. While this work did not specifically address identification of AEs, the clinical and conceptual challenges are the same, and indeed in some cases, medical problems are adverse events.

Of interest was their finding that traditional mapping of text to controlled vocabularies (time-consuming work that often reflects individual preference) could be helped by accurate boundary detection by machine learning systems which do Named Entity Recognition (NER) tasks (find and classify words and phrases into semantic classes). They hypothesize this system could help recognize unknown words based on context and so could supplement traditional dictionary-based NLP systems. The implication here is that the task of finding and accurately coding adverse events (among other medical concepts) could be significantly standardized and automated via the methods described.

For pharmacovigilance, this would have a direct application not only in finding AEs in discharge summaries, but in recognizing AEs from patient diaries and notes, where an expression that refers to an AE may have no recognition in a dictionary-based system (e.g., “this stuff split my head into” – where the vernacular refers to a drug-induced headache, but the terms and the misuse of “into” vs “in two” makes machine recognition challenging).

The development of a machine-learning approach demands better-defined, more logically consistent datasets, and this has spurred work which will change the traditional challenges associated with terminologies. Borrowing from a bioinformatics and systems biology approach, Cai et al. created ADReCS – the Adverse Drug Reaction Classification System [7]. ADReCS is an ontology of AE terms built with MedDRA and UMLS with hierarchical classification and digital identifiers. This means that direct computation on ADR terms can be achieved using the system, a significant step for the efficient use of machine learning technologies. We can imagine a future where this system or ones similar are expanded and mapped to other ontologies built in a similar manner, allowing for an approach to pharmacovigilance that is unlike anything in the past. As we reach this stage of computational maturity in pharmacovigilance, it will create a very significant driver for the biopharmaceutical industry, which spends a great deal on gathering data from disparate sources to test drug safety hypotheses and to standardize and recode that data into common formats that can be submitted to regulators. As systems like ADReCS become the norm, many of the inefficiencies the industry now faces will begin to disappear.

As with ontologies, work is needed to expand the most promising systems and to find the most universal and effective representations of terminologies that can migrate successfully from healthcare to industry to regulators with no loss of meaning and will decreased manual effort.

Topic 4: Discovery/Curation of AEs

Research on the discovery of AEs is being done in every possible source – electronic health records, social media, registries, large databases, real-world data from insurance claims, and other sources [35]. In 2012, Harpaz et al. set the stage for the use of novel methodologies using large datasets with their review of current work [16]. The authors made several salient points regarding the new research methods, including the fact that (1) combining data from heterogeneous sources requires the development of new and reproducible methods; (2) standardized (and simulated)

datasets will grow in importance to allow rapid testing of new methods; and (3) standards in PV must be developed to evaluate algorithmic approaches applied to the data. In 2013 Jiang et al. began work on ADEpedia 2.0, which built on their previous AE knowledge base derived from drug product labels; in keeping with the direction laid out by Harpaz, in 2.0 the authors began to enrich the database with data from UMLS (Unified Medical Language System) and EHR data, with a goal to create a standardized source of AE knowledge [25]. Banda et al. continued this approach, standardizing the FDA's FAERS (FDA Adverse Event Reporting System) database [3]. They provided a curated database removing duplicate records, mapping the data to standardized vocabularies with drug names mapped to RxNorm concepts and outcomes mapped to SNOMED-CT concepts, and created a set of summary statistics about drug-outcome relationships for general consumption. While not involved directly with machine learning, this approach pointed the way toward further machine-based approaches by providing all source code for the work, so that it could be used and updated as needed, and by mapping outcomes and indications to SNOMED-CT, this allows for direct linkage to other ontologies.

Since that time, an explosion of work has taken place in all three areas identified by Harpaz, emphasizing the discovery of AEs using machine learning combined with statistical techniques [5, 6, 18, 20, 23, 67].

The study by Bean et al. serves to illustrate a new way of approaching discovery of AEs in the postmarketing phase – one that doesn't wait for a series of reports to emerge, rather it takes advantage of what until recently were infrequently connected sources of data to discover previously unknown AEs due to specific drugs and to validate this via EHRs. The authors constructed a knowledge graph with four primary sources of data: drugs, protein targets, indications, and adverse reactions that predicted AEs from public data. They then used this to develop a machine learning algorithm and deployed that algorithm on an EHR. The algorithm was fed by an NLP pipeline developed to parse free text in the EHR. This work is similar to work on prediction of AEs using structure-activity relationships [12], gene expression [60], and protein drug targets [45]. In this work we can see a computational biological approach which can view with the current biology-based approach that has paid dividends but dominated PV for decades.

In 2017 Voss et al. moved the field forward significantly with their work to automatically aggregate disparate sources of data into a single repository [59] that allows a machine learning approach to selecting positive and negative controls for pharmacovigilance research design testing. As previous work demonstrated, creating a reference database for pharmacovigilance using manual or even semi-manual methods, is extremely time and resource intensive. The authors built on previous work (described in Banda) and added the relationship between a drug and a health outcome of interest (HOI). They performed a quantitative assessment of how well the evidence base could discriminate between known positive drug-condition causal relationships and drugs known to be not associated with a condition, thus allowing the automated creation of an assessment for pharmacovigilance research study designs that allows comparisons across designs with a significant savings in time and increase in standardization. The authors worked through methods for accepting data from various sources at various granular levels, for example, mapping the source at either an ingredient level of a drug

or at a clinical drug level and subsequently aggregating evidence to individual ingredients to allow analysis across the dataset.

While work in AE discovery occurs at every level and is often the primary topic in other research covered under other research topics such as terminologies, the topic of curation is not one typically addressed except in individual efforts that are not reproducible and rarely maintained due to the intense effort required. This shift toward automated curation across various data sources will prove to be an important stimulus to the computational approach in pharmacovigilance, allowing a much more rapid and standardized testing of research designs. In the near future, we can expect more reference sets which can be used to train machine learning algorithms and test large-scale analysis methods.

Just as the creation of high-quality curated datasets in machine learning is driving forward progress across many fields [63], we can expect the same to occur in pharmacovigilance as work continues. It is insightful to review the dramatic effect that a massive, well-curated (automatically generated) dataset can have on accuracy of machine learning algorithms in the example of ImageNet [21, 62], a database of over 14 million image URLs that are labeled to provide a curated set. Prior to establishing this dataset, progress in visual object recognition was steady but slow. In 2012, using a deep convolutional neural network trained on ImageNet, researchers bested other networks by over 40% to the next best [13]. This massive, curated dataset is widely attributed as one of the primary drivers of the deep learning revolution.

The moral of this story for pharmacovigilance is that a focus on the creation of large, curated, automatically created test datasets has the potential to move a computational approach to pharmacovigilance forward just as quickly if not more quickly than the best analytical methods. This is certainly an area for future informatics research.

Topic 5: Delayed Toxicity and Complex Causal Assessments

The tragic discovery of delayed hepatotoxicity caused by fialuridine is required reading in any pharmacovigilance or clinical research education. It is important to understand just how difficult it was at that time to attribute observed toxicities to the drug, given how they initially presented in patients and the presence of similar symptoms due to underlying disease or caused by an initial therapeutic response. These challenges, coupled with the piecemeal accumulation of information over a period of time, made it difficult to form a conviction that fialuridine caused a fatal toxicity – although, as some argue, evidence was clearly present [4, 57]. The 1995 Institute of Medicine report on the review of events leading up to the tragic deaths of five patients in a 1993 clinical trial of fialuridine for hepatitis B concluded that overall, clinical researchers involved in various trials acted correctly and made the best decisions possible given the available information. Looking at the set of trials that were done over a period of several years, however, one cannot help but be struck by the series of “clues” pointing to fialuridine, and how, when taken together, they provide a strong signal that the drug was implicated [22].

In our current post-behavioral economics atmosphere, it may be easier for us to appreciate how we could fail to recognize a problem of delayed toxicity in a drug:

humans are superb at pattern recognition over a relatively short timeframe, but our skill degrades rapidly as cause is separated in time from effect and obscured by other possible causes. In pharmacovigilance, one has a feeling of inadequacy when it comes to sorting out the possible links between drugs and toxicities, except in the most obvious and common cases. The investigations into fialuridine-delayed toxicity produced better regulation and reasonable research recommendations [22], but beyond these improvements, not much has been gained in our ability to recognize delayed toxicity in drugs from complex situations.

A less dramatic but conceptually similar challenge faces anyone seeking to sort out what drugs may be contributing to a patient's clinical signs and symptoms when they have underlying disease and are on a multiple drug regimen. The classic questions regarding "dechallenge/rechallenge" (whether a sign or symptom stopped once drug was stopped, and returned after drug was restarted) and the time course of drug dose vs appearance of symptoms are well-designed but often unanswerable in a real-world situation. Oncology trials come to mind as a particularly challenging environment in which to attribute cause to individual drugs.

These scenarios are not unique to pharmacovigilance. They share the same basic external challenges – incomplete information, competing causes, extended over-time, and internal challenges – idiosyncratic human perception, and bias with pursuits as diverse as cognitive psychology and behavioral economics [54] or the study of policy impacts [43].

Computational approaches to these questions hold out promise to provide the most significant advancement in years for pharmacovigilance, by transferring the burden of recognition to computers working with large datasets using sound methods. Most of the work reviewed earlier in the recognition of AEs applies here as well. Huang et al. systems pharmacology approach of combining clinical observation with molecular biology [20] can be seen as template for research in predicting toxicities in drugs and arming researchers with information that will enhance the design as well as the monitoring of trials using drugs with increasingly complex mechanisms of action. Recent similar work indicates that a systems pharmacology or computable biology approach holds out great promise in predicting toxicities at an earlier stage than previously imagined [1, 27, 28, 65].

Combining data across disciplines in a computable framework is a fertile area of research, especially as it applies to predicting toxicities in a real-world setting. The contribution of informatics to this work can have a tangible and concrete impact in improving safety for patients. Arming clinical researchers and pharmacovigilance professionals with these methods holds out hope that another fialuridine tragedy would be avoided today.

Topic 6: Risk Profiling of the Individual

The concept of precision medicine that medical care can be tailored – especially in a genomic and molecular sense – to select groups of patients is now commonplace and being realized in the design of clinical trials and healthcare policy in addition to medical practice. In pharmacovigilance, however, there is a need for better

understanding of how the concepts of precision medicine can be incorporated into goals and practice. This section simply poses some basic open questions that informaticians can help to address in order to improve the theoretical basis of pharmacovigilance. But while the ideas here are to some degree speculative, the authors believe they should be taken seriously, as they are at the heart of pharmacovigilance itself.

A simple coined term “Precision Pharmacovigilance” is enough to raise questions and spark ideas about how the discoveries in medicine and biology can be more directly taken up in the study of drug and device safety. But a broader (and more provocative) research question to ask is: Is it possible to provide to an individual a “risk profile” as it relates to their particular drug and/or device regimen? One aspect of the question relates to the degree to which we can simply follow the discoveries in precision medicine and practice pharmacovigilance along the way – e.g., looking at AEs in certain genetic subgroups while undergoing treatment with immune modulators. It could be said that in this respect, there’s nothing new here; pharmacovigilance has looked at subgroups of patients for some time [33] and continues to bear fruit [53].

But recent research in methods dealing with large-scale longitudinal observational databases [34, 52] allows us to imagine a scenario different from that of looking at the AEs related to certain subtypes of patients – what if we could predict the risk of being a certain person (age, race, genetic makeup), and taking a certain set of drugs (let’s say a regimen of five separate drugs), and living in a certain area of the world, and having a particular occupation..? Can we reach the point where we can tell you that for you as an individual, you have a 60% chance of a significant toxicity if you fit the above profile? The question serves less to examine how much data would it take to provide an exact answer and more to challenge us to decide how feasible it is to pursue this goal. Can pharmacovigilance aspire to studying and predicting risk not only for patient subtypes but for situational circumstances?

At this early stage of discovery and application in big data, machine learning, and improving methods, it is important to keep an open mind about what pharmacovigilance can become. Being able to speak directly to select groups of patients who are living in specific circumstances as regards their drug therapy was an original motivation for pharmacovigilance, and we believe should continue to inspire research.

Conclusion

For many years, pharmacovigilance developed in lockstep with general medical and clinical research, focusing on average effects in the “average” patient. Any focus on specificity came in the form of concerns regarding individual drugs, reinforced by the regulators’ need to approve specific compounds made by specific manufacturers. In recent years, however, the shift to precision medicine and away from the idea that the goal is treat an average population has left pharmacovigilance caught out, with the need to reexamine its methods and aspirations. Too often

today conversations in pharmacovigilance proceed as though the approach to drug safety is immutable and can be applied mutatis mutandis to new “topics” such as precision medicine. And while it is true that the pursuit of a better understanding of the safety of drugs remains the same, it is not true that the assumptions formed before the era of the Internet, big data, and machine learning are adequate to carry the discipline forward. As mentioned earlier, Simon’s admonition that “A design representation suitable to a world in which the scarce factor is information may be exactly the wrong one for a world in which the scarce factor is attention”, [58] can be seen as an indictment of our reliance on outdated concepts in pharmacovigilance today.

There are, however, a group of individuals who are pressing ahead with a new vision of pharmacovigilance and who are beginning to provide the outlines of the future of the field. Tatonetti makes a well-reasoned argument for integrating observational data with laboratory experiments in model systems to create a new pharmacovigilance process [56]. His three-step process involves detection, now performed by mining large observational data sources; corroboration, using techniques similar to those described earlier to test the plausibility of hypotheses in EHR data combined with other types of data such as molecular or physical chemistry; and finally validation, where a model system is found to test the strongest of the hypotheses.

Schuemie et al. propose a new method for observational studies that moves away from the one-off model and provides the ability to look at treatments at scale [51]. They describe methods to compare the full set of depression treatments for a set of outcomes, which produced 17,718 hazard ratios – clearly demonstrating what the authors call “high-throughput observational studies.”

Natsiavas et al. address issues with an outdated process of communication of possible signals in pharmacovigilance that is built on the pen and paper paradigm, missing out on opportunities for automated data sharing and reuse and thus creating the chance for much more rapid, accurate dissemination of information with better collaboration, signal validation, and characterization [36]. The authors created “OpenPVSignal” which is an ontology to specifically support the communication of PV signals. The authors specify concepts and relationships to allow publishing signal information with accompanying data and allowing computation on the data. The benefits of wide adoption of such a system cannot be underestimated, and the authors are to be commended for their attempt to address a problem so entrenched in daily practice that we do not notice how much better practice could be if we were to envision it as operating in a modern Internet-enabled and computable manner.

In pharmacovigilance, it is clear that the flow of digitized data being generated in healthcare will continue to drive together what over the years has become separate islands of practice. It is up to us to decide if we will use this confluence to build a new and vibrant foundation for the field.

Data is a precious thing and will last longer than the systems themselves.
–Tim Berners-Lee

References

1. Ai H, Chen W, Zhang L, Huang L, Yin Z, Hu H, Zhao Q, Zhao J, Liu H. Predicting drug-induced liver injury using ensemble learning methods and molecular fingerprints. *Toxicol Sci.* 2018; <https://doi.org/10.1093/toxsci/kfy121>.
2. Andrews EB, Moore N. Mann's pharmacovigilance. 3rd ed. Chichester: Wiley-Blackwell; 2014.
3. Banda JM, Lee E, Vanguri RS, Tatonetti NP, Ryan PB, Shah NH. A curated and standardized adverse drug event resource to accelerate drug safety research. *Sci Data.* 2016;3:160026.
4. Bari A. Severe toxicity of Fialuridine (FIAU). *N Engl J Med.* 1996;334(17):1135; author reply 1137–38.
5. Bean DM, Honghan W, Iqbal E, Dzahini O, Ibrahim ZM, Broadbent M, Stewart R, Dobson RJB. Knowledge graph prediction of unknown adverse drug reactions and validation in electronic health records. *Sci Rep.* 2017;7(1):16416.
6. Boland MR, Jacunski A, Lorberbaum T, Romano JD, Moskovitch R, Tatonetti NP. Systems biology approaches for identifying adverse drug reactions and elucidating their underlying biological mechanisms. *Wiley Interdiscip Rev Syst Biol Med.* 2016;8(2):104–22.
7. Cai M-C, Xu Q, Pan Y-J, Pan W, Ji N, Li Y-B, Jin H-J, Liu K, Ji Z-L. ADReCS: an ontology database for aiding standardization and hierarchical classification of adverse drug reaction terms. *Nucleic Acids Res.* 2015;43(D1):D907–13.
8. Chilcott M. How data analytics and artificial intelligence are changing the pharmaceutical industry. *Forbes Mag.* May 10, 2018. 2018. <https://www.forbes.com/sites/forbestechcouncil/2018/05/10/how-data-analytics-and-artificial-intelligence-are-changing-the-pharmaceutical-industry/>.
9. Cimino JJ, Clayton PD, Hripcsak G, Johnson SB. Knowledge-based approaches to the maintenance of a large controlled medical terminology. *J Am Med Inform Assoc JAMIA.* 1994;1(1):35–50.
10. Dua S, Rajendra Acharya U, Dua P. Machine learning in healthcare informatics. Intelligent Systems Reference Library. 2014. <https://link.springer.com/book/10.1007%2F978-3-642-40017-9>.
11. Ethier J-F, Dameron O, Curcin V, McGilchrist MM, Verheij RA, Arvanitis TN, Tawee A, Delaney BC, Burgun A. A unified structural/terminological interoperability framework based on LexEVS: application to TRANSFoRm. *J Am Med Inform Assoc: JAMIA.* 2013;20(5):986–94.
12. Frid AA, Matthews EJ. Prediction of drug-related cardiac adverse effects in humans – B: use of QSAR programs for early detection of drug-induced cardiac toxicities. *Regul Toxicol Pharmacol: RTP.* 2010;56(3):276–89.
13. Gershgorn D. The data that transformed AI research—and possibly the world. *Quartz.* Quartz. July 26, 2017. 2017 <https://qz.com/1034972/the-data-that-changed-the-direction-of-ai-research-and-possibly-the-world/>.
14. Gruber TR. A translation approach to portable ontology specifications. *Knowl Acquis.* 1993;5(2):199–220.
15. Härmäk L, van Grootheest AC. Pharmacovigilance: methods, recent developments and future perspectives. *Eur J Clin Pharmacol.* 2008;64(8):743–52. <https://doi.org/10.1007/s00228-008-0475-9>. Epub 2008 Jun 4. <https://www.ncbi.nlm.nih.gov/pubmed/18523760>.
16. Harpaz R, DuMouchel W, Shah NH, Madigan D, Ryan P, Friedman C. Novel data-mining methodologies for adverse drug event discovery and analysis. *Clin Pharmacol Ther.* 2012;91(6):1010–21.
17. Henegar C, Bousquet C, Louët AL-L, Degoulet P, Jaulent M-C. Building an ontology of adverse drug reactions for automated signal generation in pharmacovigilance. *Comput Biol Med.* 2006;36(7):748–67.
18. Ho T-B, Le L, Thai DT, Taewijit S. Data-driven approach to detect and predict adverse drug reactions. *Curr Pharm Des.* 2016;22(23):3498–526.

19. https://link.springer.com/chapter/10.1007/978-1-84882-448-5_19.
20. Huang L-C, Wu X, Chen JY. Predicting adverse side effects of drugs. *BMC Genomics.* 2011;12(5):S11.
21. ImageNet Large Scale Visual Recognition Competition (ILSVRC). n.d. Accessed 2 Jul 2018. <http://www.image-net.org/challenges/LSVRC/>.
22. Institute of Medicine (US) Committee to Review the Fialuridine (FIAU/FIAC) Clinical Trials. In: Manning FJ, Swartz M, editors. *Review of the fialuridine (FIAU) clinical trials.* Washington, DC: National Academies Press (US); 1995.
23. Jamal S, Goyal S, Shanker A, Grover A. Predicting neurological adverse drug reactions based on biological, chemical and phenotypic properties of drugs using machine learning models. *Sci Rep.* 2017;7(1):872.
24. Jiang M, Chen Y, Mei L, Trent Rosenbloom S, Mani S, Denny JC, Hua X. A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries. *J Am Med Inform Assoc: JAMIA.* 2011;18(5):601–6.
25. Jiang G, Liu H, Solbrig HR, Chute CG. ADEpedia 2.0: integration of normalized adverse drug events (ADEs) knowledge from the UMLS. In: AMIA joint summits on translational science proceedings. AMIA joint summits on translational science 2013 (March); 2013. p. 100–4.
26. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, Wang Y, Dong Q, Shen H, Wang Y. Artificial intelligence in healthcare: past, present and future. *Stroke Vasc Neurol.* 2017;2:230–43. September, svn – 2017-000101.
27. Kim E, Nam H. Prediction models for drug-induced hepatotoxicity by using weighted molecular fingerprints. *BMC Bioinforma.* 2017;18(7):227.
28. Kotsampasakou E, Montanari F, Ecker GF. Predicting drug-induced liver injury: the importance of data curation. *Toxicology.* 2017;389:139–45.
29. Koutkias VG, Jaulent M-C. Computational approaches for pharmacovigilance signal detection: toward integrated and semantically-enriched frameworks. *Drug Saf: Int J Med Toxicol Drug Experience.* 2015;38(3):219–32.
30. Kovacevic A, Dehghan A, Filannino M, Keane JA, Nenadic G. Combining rules and machine learning for extraction of temporal expressions and events from clinical narratives. *J Am Med Inform Assoc: JAMIA.* 2013;20(5):859–66.
31. Kuhn TS. The structure of scientific revolutions. Chicago: University of Chicago Press; 2012. pu3430623_3430810. April 2012. <http://www.press.uchicago.edu/ucp/books/book/chicago/S/bo13179781.html>.
32. Linder JA, Haas JS, Iyer A, Labuzetta MA, Ibara M, Celeste M, Getty G, Bates DW. Secondary use of electronic health record data: spontaneous triggered adverse drug event reporting. *Pharmacoepidemiol Drug Saf.* 2010;19(12):1211–5. <https://doi.org/10.1002/pds.2027>.
33. Lynch T, Price A. The effect of cytochrome P450 metabolism on drug response, interactions, and adverse effects. *Am Fam Physician.* 2007;76(3):391–6.
34. Moghaddass R. The factorized self-controlled case series method: an approach for estimating the effects of many drugs on many outcomes. n.d.
35. Murff HJ, Patel VL, Hripcak G, Bates DW. Detecting adverse events for patient safety research: a review of current methodologies. *J Biomed Inform.* 2003;36(1–2):131–43.
36. Natsiavas P, Boyce RD, Jaulent M-C, Koutkias V. OpenPVSignal: advancing information search, sharing and reuse on pharmacovigilance signals via FAIR principles and semantic web technologies. *Front Pharmacol.* 2018;9:609.
37. Naughton J. How a 1930s theory explains the economics of the internet. *The Guardian.* September 7, 2013. 2013. <http://www.theguardian.com/technology/2013/sep/08/1930s-theory-explains-economics-internet>.
38. OMOP Common Data Model – OHDSI. n.d. Accessed 8 Mar 2018. <https://www.ohdsi.org/data-standardization/the-common-data-model/>.
39. Pacaci A, Gonul S, Anil Sinaci A, Yuksel M, Erturkmen GBL. A semantic transformation methodology for the secondary use of observational healthcare data in postmarketing safety studies. *Front Pharmacol.* 2018;9:435.

40. PatientsLikeMe and the FDA Sign Research Collaboration AgreementPatientsLikeMe. n.d. Accessed 28 June 2018. <http://news.patientslikeme.com/press-release/patientslikeme-and-fda-sign-research-collaboration-agreement>.
41. [PDF]Guidance for Industry Postmarketing Adverse Event Reporting ... – FDA. n.d. <https://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm071982.pdf>.
42. [PDF]How the Internet Promotes Development – World Bank Documents. n.d. http://documents.worldbank.org/curated/en/896971468194972881/310436360_20160263021502/additional/102725-PUB-Replacement-PUBLIC.pdf.
43. [PDF]NoNIE Guidance on Impact Evaluation – World Bank Group. n.d. http://siteresources.worldbank.org/EXTOED/Resources/nonie_guidance.pdf.
44. [PDF]The Nature of the Firm R. H. Coase *Economica*, New Series, Vol. 4, No. n.d. <https://www.colorado.edu/ibs/es/alston/econ4504/readings/The%20Nature%20of%20the%20Firm%20by%20Coase.pdf>.
45. Pérez-Nueno VI, Souchet M, Karaboga AS, Ritchie DW. GESSE: predicting drug side effects from drug–target relationships. *J Chem Inf Model.* 2015;55(9):1804–23.
46. Personeni G, Bresso E, Devignes M-D, Dumontier M, Smail-Tabbone M, Coulet A. Discovering associations between adverse drug events using pattern structures and ontologies. *J Biomed Semant.* 2017;8(1):29.
47. Quintiles Launches Patient Website iGuard for Drug Safety Service – CenterWatch News Online. CenterWatch news online. September 13, 2007. 2007. <https://www.centerwatch.com/news-online/2007/09/13/quintiles-launches-patient-website-iguard-for-drug-safety-service/>.
48. Reich C, Ryan PB, Stang PE, Rocca M. Evaluation of alternative standardized terminologies for medical conditions within a network of observational healthcare databases. *J Biomed Inform.* 2012;45(4):689–96.
49. Research, Center for Drug Evaluation and. Guidances (drugs) – E2B(R3) electronic transmission of individual case safety reports implementation guide – data elements and message specification; and appendix to the implementation guide – backwards and forwards compatibility. n.d. <https://www.fda.gov/drugs/guidancecomplianceregulatoryinformation/guidances/ucm274966.htm>.
50. Schroll JB, Maund E, Gøtzsche PC. Challenges in coding adverse events in clinical trials: a systematic review. *PLoS One.* 2012;7(7):e41174.
51. Schuemie MJ, Ryan PB, Hripcak G, Madigan D, Suchard MA. A systematic approach to improving the reliability and scale of evidence from health care data. 2018. arXiv [stat.AP]. arXiv. <http://arxiv.org/abs/1803.10791>.
52. Shaddox TR, Ryan PB, Schuemie MJ, Madigan D, Suchard MA. Hierarchical models for multiple, rare outcomes using massive observational healthcare databases. *Stat Anal Data Min.* 2016;9(4):260–8.
53. St Sauver JL, Olson JE, Roger VL, Nicholson WT, Black JL 3rd, Takahashi PY, Caraballo PJ, et al. CYP2D6 phenotypes are associated with adverse outcomes related to opioid medications. *Pharmacogenomics Personalized Med.* 2017;10:217–27.
54. Stiensmeier-Pelster J, Heckhausen H. Causal attribution of behavior and achievement. In: Heckhausen J, Heckhausen H, editors. Motivation and action. Cham: Springer International Publishing; 2018. p. 623–78.
55. Talbot J, Aronson JK, editors. Stephens' detection and evaluation of adverse drug reactions: principles and practice. 6th ed. Chichester: Wiley; 2011.
56. Tatonetti NP. The next generation of drug safety science: coupling detection, corroboration, and validation to discover novel drug effects and drug-drug interactions. *Clin Pharmacol Ther.* 2018;103(2):177–9.
57. The Cure That Killed | DiscoverMagazine.com. Discover Magazine. n.d. Accessed 4 Jul 2018. <http://discovermagazine.com/1994/mar/thecurethatkille345>.
58. The MIT Press, editor. The sciences of the artificial. 3rd ed. The MIT Press; n.d. Accessed 29 June 2018. <https://mitpress.mit.edu/books/sciences-artificial-third-edition>.

59. Voss EA, Boyce RD, Ryan PB, van der Lei J, Rijnbeek PR, Schuemie MJ. Accuracy of an automated knowledge base for identifying drug adverse reactions. *J Biomed Inform.* 2017;66:72–81.
60. Wang Z, Clark NR, Ma'ayan A. Drug-induced adverse events prediction with the LINCS L1000 data. *Bioinformatics.* 2016;32(15):2338–45.
61. WHO. http://www.who.int/medicines/areas/quality_safety/safety_efficacy/pharmvigi/en/.
62. Wikipedia contributors. ImageNet. Wikipedia, the free encyclopedia. June 21, 2018. 2018a. <https://en.wikipedia.org/w/index.php?title=ImageNet&oldid=846928201>.
63. Wikipedia contributors. List of datasets for machine learning research. Wikipedia, the free encyclopedia. July 1, 2018. 2018b. https://en.wikipedia.org/w/index.php?title=List_of_datasets_for_machine_learning_research&oldid=848338519.
64. WuXi Global Forum Team. Artificial intelligence already revolutionizing pharma. January. 2018. <http://www.pharmexec.com/artificial-intelligence-already-revolutionizing-pharma>.
65. Yang H, Sun L, Li W, Liu G, Tang Y. In silico prediction of chemical toxicity for drug design using machine learning methods and structural alerts. *Front Chem.* 2018;6:30.
66. Yuksel M, Gonul S, Erturkmen GBL, Sinaci AA, Invernizzi P, Faccinetti S, Migliavacca A, Bergvall T, Depraetere K, De Roo J. An interoperability platform enabling reuse of electronic health records for signal verification studies. *Biomed Res Int.* 2016;2016:1–18. <https://doi.org/10.1155/2016/6741418>.
67. Zhang W, Liu F, Luo L, Zhang J. Predicting drug side effects by multi-label learning and ensemble learning. *BMC Bioinforma.* 2015;16:365.



Clinical Trial Registries, Results Databases, and Research Data Repositories

21

Karmela Krleža-Jerić

Abstract

Trial registration, results disclosure, and sharing of analyzable data are considered powerful tools for achieving higher levels of transparency and accountability for clinical trials. The emphasis on disseminating knowledge and growing demands for transparency in clinical research are contributing to a major paradigm shift in health research. In this new paradigm, knowledge will be generated from the *culmination* of all existing knowledge – not just from bits and parts of previous knowledge, as is largely the case now. The full transparency of clinical research is a powerful strategy to diminish publication bias, increase accountability, avoid unnecessary duplication of research (and thus avoid research waste), efficiently advance research, provide more reliable evidence for diagnostic and therapeutic interventions, and regain public trust. Transparency of clinical trials, at a minimum, means sharing information about the design, conduct, results, and analyzable data. Not only must the information itself be explicitly documented, but an access location or medium for distribution must be provided. In the case of clinical trials, the public disclosure of data is realized by posting cleaned and anonymized data in well-defined, freely accessible clinical trial registries and results databases. Making cleaned, anonymized individual participant data sets analyzable is still a challenge.

Basic electronic tools that enable sharing clinical trial information include registries hosting protocol data, results databases hosting aggregate data, and research data repositories hosting reusable/analyzable data sets and other research-related information. These tools are at different levels of development

K. Krleža-Jerić, MD, MSc, DSc (✉)

IMPACT Observatory, Mediterranean Institute for Life Sciences- MedILS,
Split & Cochrane Croatia, University of Split School of Medicine, Split, Croatia

Electronic Health Information Laboratory -EHIL, CHEO, Ottawa, ON, Canada
e-mail: karmela@krleza.com

and plagued with heterogeneity as international standards exist only for trial registration. The lack of standards related to publishing data in repositories makes it difficult for researchers to decide where to publish and search for data from completed studies.

Keywords

Transparency in clinical research · Trial registries · International standards · Results databases · Protocol-Results-Data · Cleaned · Anonymized individual participant data (IPD) · Analyzable data · Research data repositories · Reuse of data-Open data · User perspectives

Background

The movement toward open science and open data (i.e., making raw data from research available for analysis) is slowly beginning to penetrate clinical trials [1]. For clinical trials, any discussion of raw data refers specifically to the cleaned and anonymized individual participant data (IPD). However, consumers of these data ultimately need analyzable data sets, which include IPD, metadata, and adjacent (or supporting) documents.

The clinical trial enterprise is international, and therefore the development of clinical trial registries, results databases, and research data repositories should be at an international level and with open access. Such international standards should be flexible to allow elaboration of required fields and addition of more fields as needed.

There are three broad types of clinical trial data that can be shared publicly or openly: protocol, results and findings, and raw data sets [2]. More precisely, these include:

- (a) The registration of selected protocol elements in trial registries which might be complemented by publication of full protocols in journals.
- (b) The public disclosure of summary results (aggregate data) in databases, usually developed by clinical trial registries; these are usually beyond publications in peer-reviewed journals.
- (c) The public availability of analyzable data sets; these data sets are based on cleaned, anonymized individual participant data (IPD) and adjacent trial documentation.

There are several modes or mechanisms of finding and accessing IPD-based analyzable data sets for secondary analysis (often called pooled or meta-analysis of IPDs). These include (a) direct researcher-to-researcher contact (reviewer contacting initial data producers), (b) initiatives and projects that play intermediary role, and (c) publicly accessible repositories.

- (a) *Direct researcher-to-researcher contact*: The reviewer gets the data directly from the original data creator by contacting him or her. The reviewer identifies studies mainly by following the literature and/or by visiting trial registries.
- (b) *Intermediary contact in which the researcher requests data from special initiatives or projects including Clinicalstydatarequest, [3] Yoda [4], the Project DataSphere [5] and recently launched Vivli [6, 7]*: The reviewer applies for data to an independent panel, a sort of peer-reviewed panel that is formed by a group of data pharmaceutical industry providers or producers (generally the pharmaceutical industry at present). The panel is usually independent international panel. Increasingly, government agencies are also moving to this direction, such as the European Medicine Agency (EMA) [8].
- (c) *Open-access, publicly accessible research data repositories* (in further text repositories). They might be either domain repositories that specialize in hosting clinical trial data or general repositories that host clinical trial data in addition to hosting raw data from several or all research areas. There are currently several such open-access general research data repositories in public domain that host CT data.

In this chapter, we focus on registries, databases, and repositories.

Rationale

Trial registration, results disclosure, and making analyzable IPD-based data publicly available all share the same underlying rationale. All three are based on the principles of making the most out of clinical research, diminishing research waste, and enhancing knowledge creation. Trial registration, results disclosure, and data sharing are considered powerful tools for achieving higher levels of transparency and accountability of clinical trials [9]. Increasing emphasis on knowledge sharing and growing demands for transparency in clinical research are contributing to a major paradigm shift in health research that is well underway. In this new paradigm, knowledge will be generated from the *culmination* of all existing knowledge – not just from bits and parts of previous knowledge, as is largely the case now [10].

A stepwise process of opening clinical trial data began with the registration of protocol elements, but it was clear from the very beginning that without results disclosure, the registration would be an empty promise. Later on, it became well understood that transparency would be not be achieved without results *and* data disclosure. Actually, one could argue that results disclosure includes publication in a journal, posting summary results in open-access Internet-based database or registry, and publishing analyzable data sets in research data repository.

We are firmly in the era of evidence-informed decision-making in health for both individuals and populations at all levels – local, regional, national, and global. This decision-making is multifaceted, from the individual patient via physician to health administrators and policy-makers [10]. Registration of protocol items, publication

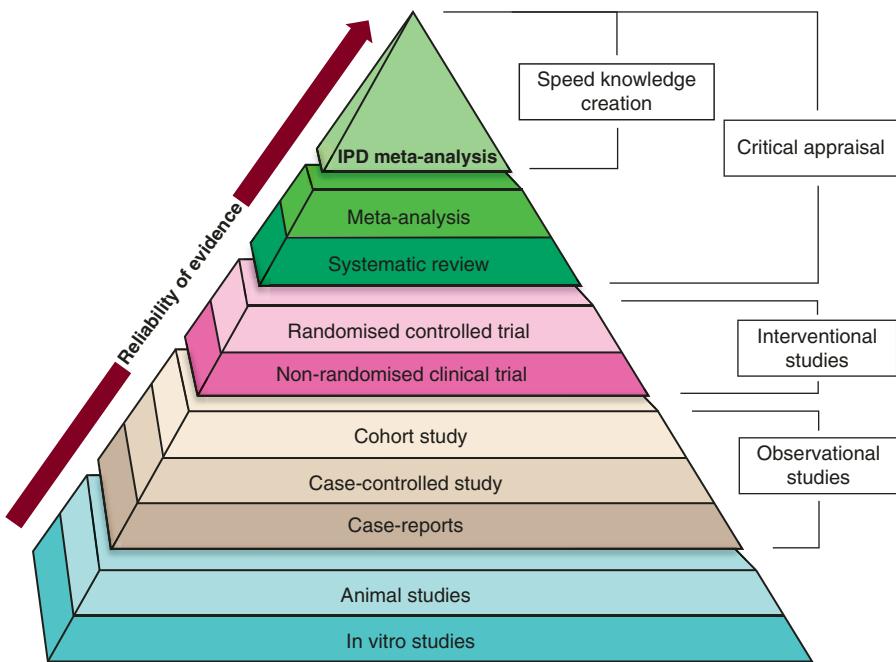


Fig. 21.1 Evidence pyramid – reliability of evidence that can be used for decision-making in health

of the complete protocol, and public disclosure of trial findings in peer-reviewed journals – complemented with public (Internet-based) disclosure of results including aggregate data and IPD-based analyzable data sets – represent a *totality* of evidence and knowledge for a given topic area and are integral to supporting efforts toward evidence-informed decision-making.

Evidence is needed to support many personal and policy decisions in health and in research. Randomized clinical trials, systematic reviews, and increasingly IPD-based meta-analyses are considered gold standards for evidence creation, illustrated by their positions at the top of the pyramid of evidence (Fig. 21.1). Actually, there has been quite an evolution from the acceptance of a systematic review (i.e., reanalyzing the *aggregate or summarized data*, usually obtained from publications) as a gold standard to the growing notion that the gold standard should require the meta-analysis using *the raw data*. This position of clinical trials on the evidence pyramid implies that the reliability of results generated by clinical trials is indeed very important. As the evidence gained from clinical trials might be directly implemented into clinical decision-making, it follows that the quality of these results should be continually scrutinized. Unfortunately, the reliability of trial-based evidence is questionable due to publication and outcome reporting bias of trials, as well as the lack of data sharing – which means that others cannot replicate or verify results. Consequently, incomplete evidence can lead to biased clinical decisions, with often harmful consequences, and can damage public trust in research and medical interventions. Following medical deontology, doctors' prescription habits are supposed to be judiciary, which requires complete and total knowledge of the benefits and

potential harms of prescribed medications. This is difficult at best and impossible if the information about the given diagnostic tools, medications, or devices is not available or is incomplete and thus biased [9, 10].

The full transparency of clinical research is a powerful strategy to diminish publication bias, increase accountability, avoid unnecessary duplication of research, avoid waste, advance research more efficiently [2], provide more reliable evidence for diagnostic and therapeutic prescriptions, speed knowledge creation, and regain public trust [10]. Transparency of clinical trials, at a minimum, means sharing information about design, conduct, and results. The information itself must be explicitly documented, but then an access location or medium for distribution must be provided. Until recently, the public disclosure of clinical trial data was realized by posting them in well-defined, freely accessible clinical trial registries and results databases. Since the first version of this chapter in 2012 [11], a lot has changed. Open-access research data repositories have been developed, and the analyzable data sets (i.e., IPDs and adjacent documentation needed to make data analyzable) can be made publicly available by publishing them in such repositories.

Considering that trials take place internationally and that the knowledge gained by them may be used by anyone anywhere in the world, their quality is also constantly and internationally scrutinized. Thus, the related standards should be internationally defined and relevant. While there are standards for trial registration and registries, the standards for results disclosure and, most importantly, standards for preparing clinical trial data for public sharing (including the definition of the requirements for repositories that host them) have yet to be developed.

Trial Registration

Development of Trial Registration

Although the need for trial registration (i.e., publishing protocol information) has been discussed for several decades, only at the beginning of this millennium did trial registration garner widespread attention from many stakeholders representing varied perspectives. The practical development of trial registration began around 2000 with two critical boosts in 2004 and in 2006. The 2004 New York State Attorney General vs. Glaxo case [12, 13] inspired the International Council of Medical Journal Editors (ICMJE) [14] and Ottawa statement [15] as well as the recommendations of the Mexico Ministerial Summit organized by the World Health Organization (WHO) [16]. These led to the development of international standards for trial registration by the WHO, which were launched in 2006 and changed the landscape of trial registration worldwide [17]. As we learned by the IMPACT Observatory scoping review [18], a number of circumstances had coincided by the year 2000 (earlier than initially thought) which enabled the development of data sharing, beginning with trial registration. These include:

- Internet-enabled storage and retrieval of large data sets
- The definition of data, metadata, and evidence-based (now increasingly called evidence-informed) medicine

- The use of evidence gained by systematic reviews and initial IPD-based meta-analysis in decision-making
- The appreciation of the impact of trial registration on knowledge creation, sharing, and Knowledge translation-KT
- The existence and experience of two major registries: the International Standard Randomized Clinical Trials Number (ISRCTN) <http://www.isrctn.com/>, based in the UK, and [ClinicalTrials.gov](https://clinicaltrials.gov), based in the USA
- Growing awareness of the need to enhance transparency
- The willingness of the international research community to embark on this undertaking
- The awareness of the harmful consequences of decision-making in the context of partial evidence
- The powerful arguments from oncology, pediatrics, rare diseases, AIDS, pregnancy, perinatal medicine, and media reporting trial-related scandals
- The need to stop wasting precious resources in unnecessary duplication of research

The initial international trial registration standards that were launched by WHO in 2006 provided essential contribution toward achieving the evidence-informed decision-making. These standards clearly identify existing registries and trials that need to be registered, define the minimum data set, designate the timing of registration, assign unique numbers to trials, and set international standards to facilitate the development of new national or regional registries as well as the comparability of data across registries. It is important to note that as of 2018, there are no international standards for results disclosure or public sharing of analyzable data. However, these are likely to be developed in the near future and will create numerous opportunities for informatics and information technology (IT) experts to leverage and apply to new applications. Additionally, further evolution of trial registration and its standards has been taking place, again leading to new applications and resources that will undoubtedly impact the development of new research and our subsequent understanding of health, disease, and effective therapies.

The goal of research transparency includes having protocol documents electronically available. For example, the protocol documents should be posted on the registry website, and all trial-related data from them ideally can be cross-referenced to results and findings. However, in reality, a trial protocol can be very complex and lengthy, which can make finding the needed information difficult. To overcome this, an international group defined the set of Standard Protocol Items for RandomIzed Trials (SPIRIT), developed SPIRIT guidelines, and made them publicly available [19–21].

SPIRIT is expected to increase the clarity of clinical research protocols and ensure that the collection of necessary items is indeed specified in the protocol, thus contributing to the overall quality of the protocol and presumably the study and results it generates. The use of SPIRIT guidelines in development of protocols might also facilitate public disclosure, especially in combination with the growing use of electronic data management [22]. It is important to note that even if full protocols are publicly

available, the existing minimum data set of the WHO international standards will still be important as the summary of a protocol. Trial registration standards will have to be revisited frequently as methodology evolves, demands for transparency increase, and with ongoing evaluation and analysis. Trial registries will most certainly expand to include results or cross-references to results databases.

Trial Registries

A *clinical trial registry* is an open-access, Internet-based repository of defined protocol information. Many different kinds of clinical trial registries exist in the public and private domains, such as international-, country-, and region-specific registries, as well as corporate (sponsor-driven) registries. The presence of multiple registries might be seen as a natural consequence of increased pressure and interest and as a positive development; however, a proliferation of registries could potentially lead to information overload and confusion for patients, clinicians, policy-makers, and research sponsors. For example, an inexperienced user may not know which clinical trial registries to trust. It might be expected that this situation will gradually correct itself as the evidence and best practice accumulate. Certainly, the proliferation of trial registries underscores the critical need for international standards that would define required features of registries as well as the content and supporting information that they must provide. Fortunately, such standards exist.

Standards, Policies, and Principles

Because clinical trials are conducted throughout the world, trial registration standards have to be defined on the international level. WHO developed international standards for trial registration, which were endorsed by the ICMJE, most medical journal editors, the Ottawa group, some public funders, organizations, and countries. It is important to note that individual countries often implement international standards by adopting and extending them with additional fields to host more information in their particular registries.

WHO international standards have helped shape many, if not all, trial registries and have been contributing to the quality and the completeness of data for registered trials. Also, it is expected that they will play a major role in further evolution of trial registration. They are sometimes referred to as WHO/ICMJE standards (or even cited only as ICMJE requirements, because the journal editors endorsed the WHO international standards in their instructions to authors and in related FAQs). These international standards define the scope (i.e., *all* clinical trials need to be registered), the registries that meet the well-defined criteria, the timing (i.e., prospective nature of the registration prior to the recruitment of the first trial participant), the content (a minimum data set that needs to be provided to the registry, initially referred to as a 20-item minimum data set), and the assignment of the unique identifier (ID). These international standards also define the criteria that the registry has to meet, which

includes level (nationwide or regional), ownership and governance (public or private nonprofit), trial acceptance, open access, and structure. In particular, structurally, the registry must have at least enough fields to host minimum data set that initially contained the following 20 items:

1. Unique trial number and the name of registry
2. Trial registration date
3. Secondary ID
4. Funding source(s)
5. Primary sponsors
6. Secondary sponsors
7. Responsible contact person
8. Research contact person
9. Public title
10. Scientific title
11. Countries of recruitment
12. Health condition or problem studied
13. Interventions (name, dose, duration of the intervention studied, and comparator)
14. Inclusion/exclusion criteria
15. Study type (randomized or not, how many arms, who is blinded)
16. Anticipated start date (and later on the actual start date)
17. Target sample size
18. Recruitment status (not yet recruiting, recruiting, temporarily stopped recruiting, or closed for recruitment)
19. Primary outcome(s) (name, prespecified time point of measurement)
20. Key secondary outcomes

Since 2012 few additional items were added to the list, each with precise definition and description, thus forming the version 1.3.1 of the WHO data set [23]. These new items are:

21. Ethics review
22. Completion date
23. Summary results
24. IPD sharing statement

In order to foster the implementation of standards, to facilitate creation of new registries, to identify the best practice, and to help develop trial registration policies, WHO formed a freely accessible search portal in 2007, followed in 2008 by the formation of a network of registries and of the Working Group on Best Practice for Clinical Trial Registries. The WHO International Clinical Trials Registry Platform (ICTRP) is a unique global portal to the trials in registries that meet criteria as data providers (i.e., WHO primary registries and [ClinicalTrials.gov](#)), but the platform does not provide access to the full extent of registries' data. Instead, the predefined 24-item data set provided by the registries is displayed (in English). The unique

identifier displayed is meant to be used in any communication about a trial, including in the ethics committees/boards' communications, consent forms, reports, publications, amendments, and press releases. This enables users and computer applications to collect trial data from many sources, allowing users to view the full picture of a given trial, from start to finish.

WHO ICRTP is also supporting a development of policies and regulations and posts them on its website. Many organizations are developing policies on clinical trial registration. While some countries recommend the trial registration (Canada, Australia) or make it a compulsory prerequisite in drug marketing authorization process (approving new drug for the market) such as the USA and the EU, so far only few countries have also developed regulations making trial registration compulsory. Some of these countries (e.g., India) also have registries, while Argentina, Israel, and Switzerland have regulations but do not yet have a registry.

Characteristics and Design Features of Trial Registries

The distinction between patient and trial registries might be confusing as they both capture certain disease-related information and often use Internet-based depositaries. However, these two types of registries are quite different. Patient registries (Chap. 13) contain records and data on *individuals*, whereas trial registries focus on the descriptive aspects of a *research study* at various stages of its implementation and often provide a link to *study results*. While trial registries can be accessed via the WHO ICTRP global search portal, at present there is no single global search portal that can be used to identify or access patient registries.

Clinical trial registries contain predefined information about ongoing and completed clinical trials, regardless of the disease or condition addressed. Patient registries contain the disease-specific information of individual patients. In a clinical trial registry, each entry represents one trial and contains selected information from protocol documents of the trial. Clinical trials are prospective interventional studies, and they may recruit either healthy volunteers or patients with various diseases. Each trial may include any number from a few to thousands of participants. In a patient registry, each entry is an individual patient with the same disease or a condition of the same group, often chronic diseases (e.g., cancer, psychosis, and rare disease patient registries).

The most important difference between trial and patient registries is the purpose. The main goal of trial registries is to provide various stakeholders with information about ongoing and completed trials, in order to enhance transparency and accountability as well as to reduce the publication bias, increase the quality of published results, prevent harmful health consequences, and most importantly, provide knowledge that will ultimately enhance patient care. Patient registries, on the other hand, are developed in order to answer epidemiological questions such as incidence and prevalence and better understand the natural course of disease including morbidity or mortality.

Some trial registries also aim to inform potential trial participants about open or upcoming trials in order to enhance recruitment. Besides being tools for

transparency, registries can also function as learning tools, and one could argue that registries might help improve the quality of the protocol and, as a result, the quality of the trials as they are completed. For example, while entering data in predefined fields, the researcher might realize that he or she is lacking some information (i.e., elements he or she forgot to define and include in the protocol) and will address the missing element(s) by editing and enhancing the protocol.

The first version of the protocol is the initial protocol that has been approved by the local ethics committee and submitted to the trial registry. Updates for trial registries are expected and consist of providing information about the protocol in various stages of the trial: prior to recruitment, during the implementation (recruitment, interventions, follow-up), and upon completion. During trial implementation, changes of protocol, called *amendments*, often take place for various reasons. Amendments to a protocol are instantiated as new protocol versions, which are dated and numbered sequentially as version 2, 3, 4, etc. Annual updates of registry data enable posting of such amendments after approval by the ethics committees. The ability to manage multiple versions of protocol documents is an important feature for a trial registry. The basic rule for the registry is to preserve all of the descriptive data of a protocol that is ever received. Once registered, trials are never removed from the registry, but rather a *status* field indicates the stage of a trial (e.g., prior to recruitment, recruiting, do not recruit any more, completed). Earlier versions of protocol-related data are kept, are not overwritten, and should still be easily accessible by trial registry users.

WHO endorses trial registries that meet international standards and calls these *primary registries*. Registries that do not meet all the criteria of international standards are considered *partner registries*, and they provide data to the WHO search portal via one or more primary registries. The need for international access and utilization of registries implies the need for a common language. While some of these registries initially collect data in the language of the country or region, they provide data to the WHO portal in English because the WHO ICTRP currently accepts and displays protocol data in English only.

It is important to note that registries that adhere to international standards tend to add additional data fields to meet their registry-specific, often country-specific, needs. Regardless of these additional fields, the essential 24 items should always be included and well-defined. Although they are bound by the international standards, the presentation of a registry's website (i.e., the web-based access and query interface) is not the same across primary registries. Some registries collect and display protocol descriptive data beyond the basic predefined 24-item fields. Those registries that collect more data typically have more extensive and detailed data for each trial record and are potentially more useful for consumers. Some registries have free-text entry fields with instructions about which data need to be provided in the fields targeted to those registering their trials, while other registries employ self-explanatory and structured fields, such as drop-down lists [24].

The WHO formed the Working Group on Best Practice for Clinical Trial Registries in 2008 in order to identify best practices, improve systems for entering new trial protocol records, and support the development of new registries [25]. The working group includes primary and some partner registries. Since the first edition

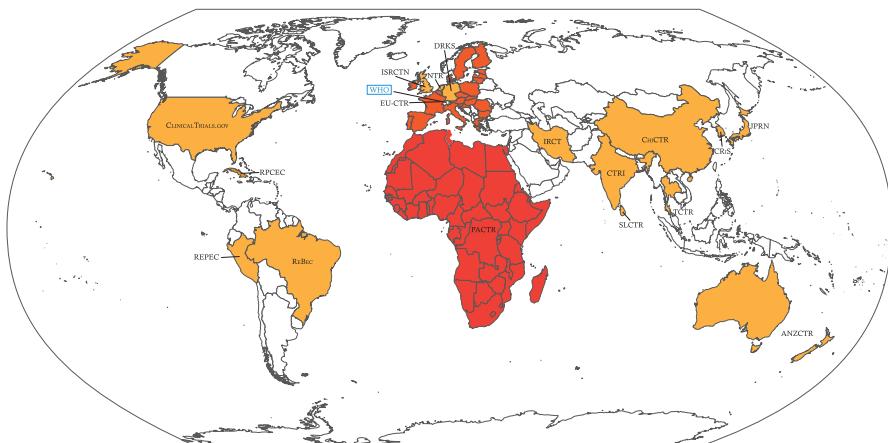


Fig. 21.2 Network of registries providing data to WHO search portal – ICRTP. This map provides the worldwide distribution of registries that directly provided data to WHO as of July 2018. *ANZCTR* Australian New Zealand Clinical Trials Registry, *ReBec* Brazilian Clinical Trial Registry, *ChiCTR* Chinese Clinical Trial Registry, *CRiS* Clinical Research Information Service, Republic of Korea, [ClinicalTrials.gov](#) (USA), *CTRI* Clinical Trials Registry, India, *EU-CTR* EU Clinical Trials Register, *RPCEC* Cuban Public Registry of Clinical Trials, *DRKS* German Clinical Trials Register, *IRCT* Iranian Registry of Clinical Trials, [ISRCTN.org](#) (UK), *JPRN* Japan Primary Registries Network, *NTR* The Netherlands National Trial Register, *PACTR* Pan African Clinical Trial Registry, *REPEC* Peruvian Clinical Trial Registry, *SLCTR* Sri Lanka Clinical Trials Registry, *TCTR* Thai Clinical Trials Registry, WHO Search Portal, Geneva. Note: The source of information: WHO ICRTP [17]. Since 2012 three registries, EU-CRT, TCTR, and REPEC joined the WHO primary registry network that directly provide data to WHO

of this book in 2012, 3 additional primary registries were developed, and as of June 2018, there were 17 registries that directly provide data to the WHO portal, specifically 16 WHO primary registries and the [ClinicalTrials.gov](#) registry which is not a part of primary registry network but provides data to the search portal. As can be seen from the geographic distribution shown in Fig. 21.2, the network includes at least one registry per continent.

Clinical trial registries can cross-reference a registered trial to its website if one exists; many large trials establish their own websites. Also, registries provide links and cross-references to publications in peer-reviewed journals, and some also cross-reference to trial results databases and research data repositories. It is expected that the number of these links will increase as results databases and repositories continue to be developed.

Timing

A responsible registrant, usually a specially delegated individual from the trial team or sponsoring organization, provides protocol-related data to the trial registry. Because all research protocols must be reviewed and approved by the ethics

committee or board of the local institution in order to conduct the study, the descriptive protocol data set is usually submitted to the trial registry after institutional ethics approval. Otherwise, registration in the trial registry is considered conditional until the ethics approval is obtained.

Although international standards require registration prior to recruitment of trial participants, this is still not fully implemented [24, 26]. Such prospective registration is important as it not only guarantees that all trials are registered but also that the initial protocol is made publicly available. For various reasons, the protocol might be changed early on, and/or a trial might be stopped within the first few weeks. Information about early protocol changes or stopped trials is lost unless trials are prospectively registered. Full data sharing is essential for the advancement of science and helps to avoid repeating such trials. Registries record the date of initial registration and date all subsequent updates. Additionally, the assignment and subsequent use of a unique ID for each trial upon registration enables any stakeholder to easily find what interests them.

Some countries hesitate to simply “import” the international standards or policies out of fear that these might change and put the country (regulator, or funding agency) in an odd position. One can debate the justification of such positions, but they are a reality. Implicit application of international standards occurs more often, with or without referencing them. Such is the case with the Declaration of Helsinki (DoH) [27], which obliges physicians via their national medical associations and is thus implicitly implemented. The DoH gradually addressed clinical trial registration and results disclosure, and the latest, 2013, Declaration explicitly calls for the registration and results disclosure of trials [27–29].

Quality of Registries

The quality of various trial registries can be judged by the extent to which they meet the predefined goal of achieving high transparency of trials. Considering that meeting international standards is a prerequisite to qualify as a WHO primary registry, the quality and utility of trial registries mainly depend upon the quality and accuracy of data and the timing of reporting [17]. To realize research transparency, clinical trials need to be registered prior to the recruitment of trial participants; this principle has not yet been fully achieved [26, 30, 31].

Registries constantly work on ensuring and improving the quality of data. The aim is to have correct data that are meaningful and precise. Accuracy of data requires regular updates in case of any changes and keeping track of previous versions. Registries impose some logical structure onto submitted data, but the quality is largely in the hands of data providers (i.e., principal investigators or sponsors). Many researchers and some registries perform analysis and evaluation of registry data [24, 31, 32]. IT experts might contribute by developing new, system-based solutions for quality control of entered trial data. Quality of data is a particularly sensitive issue as trial registries are based upon self-reporting by researchers, their teams, or sponsors. Following international standards and national requirements are

prerequisites for attaining an acceptable level of data quality. (Note that the practical and theoretical aspects of data quality are described in Chap. 11.)

The numerous and ongoing analyses and evaluations of implementation of standards and the quality of registries will enable revisions and updates, thereby improving trial registries at large. Furthermore, trial registries should reflect the reality of clinical trials methodology, which is constantly developing. Understandably, this presents a continuing challenge to those involved with the IT aspects of the data collection.

Registries that meet international standards might accept trials from any number of countries with data in the country's native language; therefore, it is essential to ensure the high quality of the translation of terms from any other language to English. Criteria that define quality also include transfer-related issues such as coding and the use of standard terms, such as those developed by the Clinical Data Interchange Standards Consortium (CDISC) [33]. For this reason, definitions of English terms used across registries created in different countries also require standardization, and there have been efforts to this end, notably those on the standard data interchange format developed by CDISC. Standardization of terms is an important issue, and solutions must balance the resources required for researchers and trial registry administrators to implement standard coding against the potential benefits for information retrieval, interoperability, and knowledge discovery. The ability of protocol data to be managed and exchanged electronically, including difficulties with computerized representation due to various coding standards for several elements such as eligibility criteria, is described in Chap. 10.

One of concerns for trial registries is the issue of duplicate registration. Duplicate registration of trials, especially of multicenter and multi-country trials, has been observed from the very beginning and was discussed by the WHO Scientific Advisory Group (SAG) while developing the standards. The concern is that duplicate registration in WHO primary registries/registries acknowledged by the ICMJE might lead to counting one trial as two, or even as several trials, and might skew conclusions of systematic reviews. Therefore, these registries perform intra-registry deduplication process, while the WHO search portal established mechanisms of overall deduplication called *bridging*. In that process, most registries have created a field for an identification number (ID) that a particular trial was given by another registry. They usually also have the field for the ID from the source, which is assigned by the funder and/or sponsor. Parallel registration in a hospital, sponsor-based, or WHO partner registry does not count as duplicate registration; only the registration in more than one primary registry of the WHO/registries recognized by the ICMJE qualifies as duplication. This is because those other registries have to provide their data to one primary registry or [ClinicalTrials.gov](#) to meet criteria of international standards and then data are provided to the WHO search portal.

It is important to note that clinical trials are sometimes justifiably registered in more than one primary registry. For example, international trials might be registered in more than one primary registry if regulators in different jurisdictions require registration in specific registries. In these cases, researchers need to cross-reference IDs assigned from one registry to another. For this reason, the creation of a field in the registry to host the ID(s) received by other registries is important. Also, it is

important that researchers provide the same trial title and the same version of protocol information in case of duplicate registration. The latter is particularly important in case of delayed registration in one of the registries and/or of initial data entry from a protocol that was already amended. Primary registries usually date the e-data entry, but it would be very useful to also number and date the protocol versions.

In 2009, as a part of implementing international standards, WHO established the universal trial number (UTN) [17], and registries developed a field to host it. This number is also meant to help control duplicate registrations. While designing a registry, it is thus necessary to anticipate the field to host the UTN. Likewise, nonprimary registries as well as eventual trial websites should create fields for UTN and IDs assigned by primary registries.

Evolution and Spin-Off

Mandates for registries determine their scope, substance, and consequent design. Although relatively new, trial registries are experiencing constant and rapid evolution, and the learning curve is steep for registrants, registry staff, registry users, and of course, IT professionals. The major impetus for the progress of trial registries followed the development of the WHO international standards in 2006 that expanded their scope from randomized controlled trials (RCTs) to all trials, regardless of the scope and type, and from a few items that indicated the existence of a trial to a summary of the protocol. At the same time, registries expanded fields and started to accept trials from other countries. Initially, registration included only RCTs that aimed at developing new drugs and collected only basic information. Of course, there is still significant potential for improvement. For example, many trials are still registered retrospectively or with a delay, but this is expected to get better with time [30, 34, 35].

Further evolution of the international trial registration standards is expected to respond to the evolution of trial methodology. For example, phases 0, I, and II might need different fields, while some fields designed for RCTs no longer apply. This has to be kept in mind while designing a registry.

Some registries, such as ClinicalTrials.gov, primarily originated from a mandate to enable potential trial participants to find a particular RCT and to enroll in it. Overall the main purpose of registries has shifted from a recruitment tool to a transparency tool while still focusing on benefits to trial participants. While registries still facilitate patients and clinicians searching by various criteria for ongoing studies, they are also becoming a source of data on various completed trials.

The trigger for trial registration was the lack of transparency and the subsequent and disastrous health consequences shown by the New York State Attorney General vs. Glaxo trial [12, 13]. This case mobilized stakeholders and elicited consequent action from various interest groups, i.e., journals, research communities, consumer advocates, regulators, etc. Nowadays, trial registries aim to inform research and clinical decisions as well as to control publication bias in response to scientific and ethical requirements of research. As a result of the international dialogue among various stakeholders, most registries now aim to meet the needs of all involved in order to elevate research to another level.

Apparently, the compliance with international standards is weak and selective when registration is voluntary, but it is gradually becoming compulsory in many jurisdictions. Still, even when regulated, compulsory registration does not necessarily meet all the requirements of the WHO international standards. For example, in the USA, registration in ClinicalTrials.gov is required by law [36]. Investigators must comply or risk a penalty; however, the law does not require registration of all trials, and it allows a delay of 21 days for registration of trials that are covered by the Food and Drug Administration Amendments Act (FDAAA) of 2007.

The experience gained so far is expected to inspire the registration of other types of studies or the development of other research-type registries. Such “spin-off” is already taking place and includes registration of observational studies in trial registries. Another example of a spin-off is the international initiative to develop a registry of systematic reviews of clinical trials and corresponding standards. The registry PROSPERO, international prospective register of systematic reviews [37], was launched in February 2011. It is expected that such registries will function based on similar principles as trial registries. For example, PROSPERO is prospectively registering a systematic review (i.e., its design and conduct, protocol, or equivalent) and is displaying a link to eventual publication of the completed review. All the information is provided by the researcher and publicly displayed on PROSPERO’s website. The registration and the usage are free of charge and freely accessible. Individual studies are the unit (record) of entry in such registries, and a mechanism for cross-referencing of study entries across various registries will be established. For example, systematic review registries might establish a cross-reference to trial registries. Such spin-off would require development of standards and creation of specific fields. Registries might provide fields to capture results or link to various levels of reporting trial results and findings, such as links to publications, capturing aggregate results data in results fields, and linking to a database with microlevel data and registry of systematic reviews.

In addition to the WHO international trial registration standards, some countries develop their own specific standards, which may meet and expand or somewhat differ from the existing standards. For example, FDAAA differs by exempting the so-called phase I and some device trials from compulsory registration. Consequently, ClinicalTrials.gov offers fields for such trials, but their registration is voluntary. There are also initiatives to develop regional registries and software that will facilitate development of individual country registries in a given region such as in the Americas [30].

Creation and Management of a Trial Registry: The User Perspective

Design of Trial Registries

As mentioned earlier, every primary trial registry now contains fields for a 24-item minimum data set as defined by the international standards and usually a few additional ones. These include the fields for the ID assigned by any other registry, the unique trial registration number (UTRN) assigned by WHO, trial website URL,

publications, etc. The required items are often expanded in several fields. For example, there may be special fields to indicate whether healthy volunteers are being recruited or to specify which participants are blinded. In parallel with registration of a minimum data set, arguments have been built for publishing the full protocol, and some journals have already started doing so. It will be particularly useful to have publicly available electronic versions of structured protocols, following SPIRIT guidelines. However, even if and when that happens, the data provided in trial registries will be useful as a summary of the protocol. These two major tools of protocol transparency (trial registry and publicly available SPIRIT-based protocol) each attract different users but undoubtedly will provide a foundation for a number of navigation and analytic tools directed toward researchers, consumers, and policy-makers.

International Standards

International standards were the major impetus for the development of trial registries. Among other advantages, standards ensure the trustworthiness of data and comparability among registries. It is important that data provided is precise and meaningful, which depends on the precision of instructions for registration and also on the fields [24]. These instructions, inspired by the WHO standards, might be developed by regulators in combination with the registry and/or journal editors as for example the Australian Clinical Trial Toolkit [38]. Registries usually have levels of compulsory completion of fields that cannot be skipped. Furthermore, they might indicate which fields or items are required by the WHO standards and/or by the appropriate national regulator. It is important to note that at this time, there are no standards for registration of observational studies, so currently registries use the trial fields and allow other descriptive data to be added.

Data Fields

The design of fields for trial registries is extremely important. Possibilities include free-text, drop-down, or predefined entries. It is advisable to define which data is needed and develop a drop-down list whenever possible. Such a drop-down list should include all known possibilities and the category “other” with text field to elaborate. Considering the rapidly developing field of clinical trials, it is necessary to anticipate additional items in a drop-down list.

Well-defined fields are prerequisite to obtain high-quality protocol data in trial registries. For example, if a registry field is free text and the data entry prompt reads *type of trial*, the answer will likely be simply “randomized controlled trial” or “randomized clinical trial” or even just the acronym “RCT.” However, the registry might prespecify in a drop-down list whether the trial is controlled or uncontrolled and whether it is an RCT and whether its design is parallel, crossover, etc.

Although phases I–IV are still in use as descriptive terms, they will probably be replaced with more specific descriptions of studies in the future. Elaboration of those numbered phases is already taking place: the phase 0 has been added, and existing phases are subdivided into a, b, and c (e.g., phase II a, b, etc.). In some cases, two phases are streamlined into one study (e.g., I/II or II/III).

Other examples of terminology issues arise within the *Study Design* field, which might include allocation concealment (nonrandomized or randomized) control, endpoint classification, intervention model, masking or blinding, and who is blinded. Thus, in the case of RCTs, the trial registry data will not simply classify a study as an RCT but will also indicate if it is a parallel or crossover trial, which participants are blinded, whether the trial is one center or multicenter, and if the latter plans to recruit in one or several countries.

Data Quality

In order to ensure the quality of data entered, instructions in the form of guidelines or learning modules are needed. Registries are developing such instructions to help researchers achieve better quality of data submitted. For example, the Australian New Zealand Clinical Trial Registry developed “data item definition and explanation” [39]. International standards, the two countries’ regulations, funders, and registries’ policies all inform the content of this tool. Initial analysis of data entry in existing acceptable registries showed that a substantial amount of meaningless information was entered in open-ended text fields [40], but it has also shown improvement in this area over time [31, 41]. Finding the balance between general versus specific information is important. For example, indicating that the trial is blinded or double-blinded is much less informative than specifying who is blinded.

Many registrants will do only what is required, which is often determined by regulations, policies of funders, or simply recommended by WHO international standards and ICMJE instructions. The following is one potential look at levels of required data fields.

First-Level Fields First-level fields are required by the regulator. For example, ClinicalTrials.gov has fields that cannot be skipped because the FDAAA requires them; ISRCTN also has fields that cannot be skipped, which are aligned with the WHO international standards. While designing a registry, one should keep in mind the possibility of expansion and provide a few fields for such unexpected information.

Second-Level Fields Second-level fields are not made compulsory by some registries but are required by others. For example, because public funders or journal editors may require additional information beyond the international standards, there is an expectation that the relevant information will be provided by registrants; however, registries themselves cannot necessarily make these fields compulsory on their end, and consequently, some registries might not have these fields. Because adding fields to registries can sometimes be difficult, posting such additionally required information elsewhere in the registry is allowed. It may be placed along with or below other information or in the *Other or Additional information* field. For this reason, it is necessary to anticipate creation of such fields. For example, Canadian Institutes of Health Research (CIHR) requires the explicit reporting and public visibility of the ethics approval and confirmation of the systematic review justifying the trial.

Third-Level Fields Third-level fields are optional and contain information that might be suggested by the registry, research groups, or offered by the researcher as important for a given trial. Such third-level data are usually entered in the *Additional information* field. This variation in fields means that, although there are international standards, there are differences among registries, specifically in the number of fields and their elaboration. The current stage of trial registries might be considered the initial learning stage, and the analysis and evaluation of current practices will point to better policies and practices for the future.

Maintenance of Trial Registries

The researcher or sponsor of a trial provides annual updates of the trial record, and all of these updates should be displayed in the registry. These updates aim at capturing all amendments (i.e., changes of the protocol, the stage of trial implementation, eventual early stopping, etc.). It is important that these updates have dedicated fields and do not overwrite previous information. Such an approach enables the identification of changes and tracks the flow of the trial implementation. The registry can be designed so that a reminder is sent automatically to registrants so that they can obtain the annual update. As mentioned earlier, registries develop special mechanisms of deduplication within the registry and with other registries.

Results Databases

Traditionally the main vehicle to disseminate trial results and findings in a trustworthy way has been via publication in a peer-reviewed journal. Due to publication and outcome reporting bias and the availability of the Internet, there is a growing international discussion about Internet-based databases of summary results. Public disclosure of results in such databases will complement publication in peer-reviewed journals, and it is an integral part of the transparency tool set.

Theoretically results databases are complex, and they might include aggregate data, metadata, and analyzable data sets. Clinical trial databases in public domain are being developed by trial registries. Currently three registries developed them: ClinicalTrials.gov, European clinical trial registry, and the Japanese UMIN. Similarly, to trial registries, results databases are expected to build hyperlinks, the most important ones being between the given trial in the registry and related publications or systematic reviews and meta-analysis. As of 2018, results databases and repositories are far less developed than trial registries. As identified by the international meeting of the Public Reporting Of Clinical Trials Outcomes and Results (PROCTOR) group in 2008 [42], and discussed later on by us [10] especially in the IMPACT Observatory [43], and by others [44], there are numerous issues to be resolved in order to get the results data, especially microlevel data sets, publicly disclosed.

Standards

There are no international standards for public disclosure of trial results, and there are no standards for preparing and use of the analyzable data sets, based on cleaned, anonymized individual participant data (IPD) and adjacent needed documentation (metadata, dictionary, etc.). However, there is much discussion on how these should be designed, and some initiatives have been contributing to accumulation of experience [28, 42, 45]. In 2010, the journal *Trials* started posting them on the Internet as the series “Sharing clinical research data,” edited by Andrew Vickers. The topic of results disclosure actually includes a spectrum of information from aggregate (summary) data to fully analyzable, i.e., IPD-based data sets. In 2017, following several years of consensus building process that involved participants from various areas and backgrounds, the ECRIN leg of the CORBEL project developed a set of recommendations regarding clinical trial data sharing [44]. Of note, clinical trial registries generally only enable the public disclosure of summary data and findings of clinical trials many of which are also published in peer-reviewed journals, while the IPD-based analyzable data sets are published in repositories.

Some of the outstanding challenges and disclosure issues regarding summary results and analyzable data are comparable to those of trial registries. These include the need to develop international standards, quality and completeness of data, timing of reporting, and standardization of terms. Other issues are more specific to the practical details of public disclosure of analyzable data sets. Those include the cleaning of data, quality of data, accountability, defining which adjacent documentation is needed, who is the guarantor of truth, privacy issues/anonymization, intellectual property rights, and issues related to anonymization efforts [46].

Many of these issues suggest a need to develop levels of detail related to levels of access. In the era of electronic data management, some of these steps, such as cleaning of raw data, are becoming less of an issue as they take place simultaneously with the data collection. Much can be learned from other areas especially from the experience of genome data sharing, for which many have shown that data sharing has boosted the development of the field [47, 48].

A lot has changed since the first version of this chapter published in 2012 [11], when these data were either protected in the hands of regulators or might have been shared with systematic reviewers only upon request and only under certain conditions. Meanwhile many constituencies engaged in making data available, especially in order to facilitate systematic reviews that include of IPD data sets (meta-analyses). For example, journal editors are increasingly encouraging data sharing upon publication of trial findings in their respective journals [49].

Data sharing is becoming more and more appealing to all stakeholders [50–53]. Earlier hesitation has been gradually lightening, and we are witnessing increased transparency and a consecutive change of the research paradigm. Although many issues have yet to be resolved, this area is constantly and rapidly evolving, and by the

time this book is printed, there will likely be more progress. However, several dilemmas and issues are still present and will require research and resolution. These include the lack of standards on how to prepare data sets for public sharing, heterogeneity of repositories, and finding the balance of privacy versus transparency [43]. All of these elements create specific challenges, require interdisciplinary work, and present an opportunity for clinical research informatics and information technology experts.

Repositories

Repositories, i.e., research data repositories, are electronic databases hosting research raw data and facilitating their reuse. They are the newest research transparency tool complementing trial registries and results databases.

As mentioned earlier when talking about data sharing from clinical trials, we are talking about the cleaned anonymized individual participant data (IPD) sets and adjacent documentation forming the analyzable data.

Repositories can be classified by the scientific area they cover or the level (university, region, country, international) at which they are organized. Re3data [54] (described below) classifies them into disciplinary, institutional, and other. Some of repositories hosting clinical trial data are based at universities and accept data only from researchers from a given university or consortium, such as Edinburgh DataShare or DRUM (Data Repository for the University of Minnesota). Figshare, on the other hand, accepts data from anywhere. Dryad accepts data if the research is published. Most general open-access repositories in public domain host data from any research. Their number is growing, and as of June 2018, there were 2109 repositories registered in re3data. However, only a small portion of them host clinical trial data. In our ongoing study we identified about a dozen general open access repositories in public domain that also host clinical trial data and analyzed their basic features [43, 55–57]. However, besides general research data repositories, there are also disease-specific repositories and research data repositories organized by funders, such as several repositories run by the NIH institutes.

With the exception of the Japanese register UMIN [58] that hosts clinical trial data of trials that are already registered in it, there is currently no domain repository in public domain, i.e., repository devoted to hosting exclusively clinical trial data.

It is important to note that the data management should begin at data collection, and public funders are increasingly demanding that the data management plan be developed up front. This leads to the understanding that the data preservation and storage of academic trials starts at the academia, that the institution – academia conducting a trial should anticipate data sharing and act accordingly – preferably develop a database and then might send data to established repositories. Indeed, several universities have been doing this. One of the first was the Edinburgh University that established Edinburgh DataShare repository which also hosts clinical trial data. It started with a JISK project led by

Edinburgh University in partnership with two other UK universities (Oxford and Southampton). While it initially hosted data from the international stroke trial, it is now hosting data from other studies conducted at the Edinburgh University [59]. The key role in setting and running of this repository has been played by research librarians. Actually, management and storage of research data have become a field of interest of research librarians, and they are increasingly engaged in this field.

Some repositories hosting clinical trial data might limit the uploading of data to members of a given university or consortium, but all of them enable open access to data for secondary use. There is a limited control of data quality at entry and no curatorship of data already in the repository. Basically, repositories rely on the clinician trialist – data provider to clean, anonymize, and organize data for publication.

Several specific projects and software have been influencing development of this field. One of them is Dataverse, which is an open-source web application to share, preserve, cite, explore, and analyze research data [60]. A Dataverse repository is the software installation, which then hosts multiple virtual archives called Dataverses. Each Dataverse contains data sets, and each data set contains descriptive metadata and data files (including documentation and code that accompany the data). As an organizing method, Dataverses may also contain other Dataverses. There are 33 Dataverse repositories (installations) around the word, and one of them, Harvard Dataverse, also hosts CT data [61].

It is important to point to the Research Data Alliance (RDA) which aims at building the social and technical infrastructure to enable open sharing of data. It functions through interest and working groups that elaborate specific topics and provide recommendations for the community [62].

Few related tools to data sharing by repositories include persistent identifiers/PID, DataCite, re3data, and the CoreTrustSeal of certification organization [63].

Re3data is a registry of research data repositories from various academic disciplines. In 2014 it merged with another similar tool, Databib, and it is now managed by DataCite. Re3data registers repositories from various disciplines and describes basic features of each of them. “It presents repositories for the permanent storage and access of data sets to researchers, funding bodies, publishers and scholarly institutions. re3data.org promotes a culture of sharing, increased access and better visibility of research data. The registry went live in autumn 2012 and is funded by the German Research Foundation (DFG)” [54, 64].

Citability and findability of published data are very important. Among other benefits, they stimulate public data sharing. Citability and to certain extend findability are achieved by assigning the *persistent identifier (PI or PID)* to published data sets [43]. PID is a long-lasting reference to a document, file, web page, or other object. The term “persistent identifier” is usually used in the context of digital objects that are accessible over the Internet. Once plugged in the web browser, it will link to related data sets which enables citation of given

data sets [65]. Persistent identifiers help the research community locate, identify, and cite research data with confidence.

DataCite is a leading global nonprofit organization that provides persistent identifiers (DOIs) for research data [66]. DataCite assigns DOI persistent identifier to each repository registered in re3data. Repositories in turn assign persistent identifier to hosted data sets, i.e., data sets published in them. In our ongoing scanning of general repositories within the IMPACT Observatory we noticed that most of the open access general repositories in public domain that host clinical trial data assign DOI, or some other PID [57].

The research community realized the importance of ensuring the quality of repositories, and in 2017, the *CoreTrustSeal* certification organization was established, developed by *the ICSU World Data System (WDS)* and *the Data Seal of Approval (DSA)* under the umbrella of RDA. The CoreTrustSeal has a set of criteria that a given repository has to meet [63]. The re3data indicates for each indexed repository whether it is certified or whether it supports repository standards.

The User Perspective

Some of repositories that host clinical trial data are open for hosting of data from certain groups of researchers, usually those linked to a given university, or area, but all of them allow open access to data they host. The lack of standards and heterogeneity of repositories makes the analysis of hosted data across several repositories very difficult if not impossible, without contacting the original data provider. It can be expected that the interest and the need for reanalysis will trigger development of needed standards. Such standards should be developed by the research community, not by repository. Ideally, internationally renowned organizations, such as WHO, will lead standard development and include key stakeholders in the consensus building process, as was the case with development of the trial registration standards.

Summary and Future

The future of clinical research and informatics is closely interwoven, and it can be expected that these evolving fields will mutually inform and influence each other. Clinical trial transparency and especially sharing of analyzable data sets are lagging behind most other research areas. There are barriers to overcome, some of which are specific for clinical trials, and they will probably continue presenting exciting challenges for researchers, information technology (IT) experts, and in fact all interested to further existing tools and figure out the sustainable strategies for public disclosure of trial information – from protocol via results to data, including the stewardship and reuse of such data in knowledge creation which will in turn speed development of new and more powerful diagnostics and therapeutics.

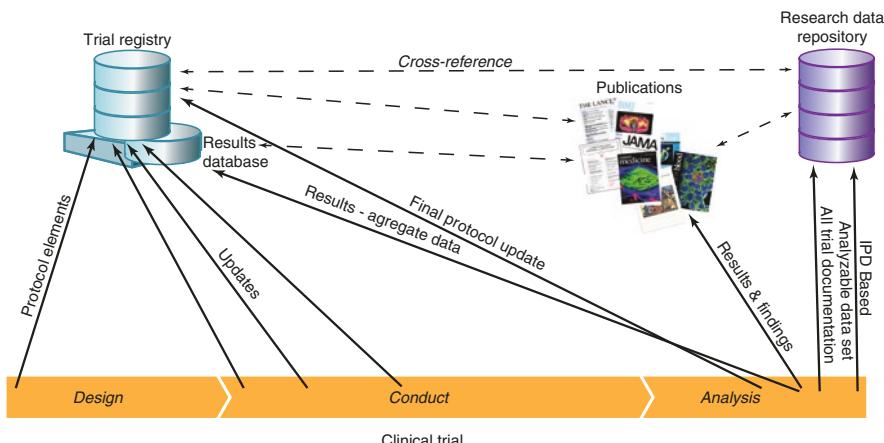


Fig. 21.3 Anticipated flow of data from clinical trial to public domain. Please note that while all parts of the data flow have evolved since 2012, the major change of this flow of data took place by the establishment of open-access research data repositories in public domain

It is anticipated that data flow from trials to the public domain and the linking and cross-referencing of related data will create a more efficient system of information sharing and knowledge creation (Fig. 21.3). Although it has not yet been completely accomplished, there is a clear tendency to move in that direction, which will ensure a high level of transparency, getting closer to open data and open science.

Furthermore, it is expected that existing systematic reviews will be updated with the meta-analysis of IPD-based analyzable data to inform various levels of decision-making with the updated evidence. Finally, in an ongoing effort to increase transparency of research and to build on the experience of trial registries, other types of studies are being registered in trial registries, and other types of research registries are being developed. However, although there are no standards and guidelines for the preparation of clinical trial data for public release and although repositories are heterogeneous, the existence of open-access repositories is a big step forward toward opening of clinical trial data.

Trial registries host defined protocol items, and they are in constant evolution, from the elaboration of fields to the establishment of hyperlinks. It can be expected that the analysis and evaluation of the existing primary registries' experience will inform the best practice and potential expansion of the data included, like adding fields to host more data than required by the initial 20-item international standards. This has already taking place, and, for example, WHO recently revised standards (version 1.3.1.) include four more protocol items: ethics review, completion date, summary results, and IPD sharing plan [23].

Furthermore, there is a strong push for publication of the full protocol, either in the registry or elsewhere. It will certainly be particularly useful to have publicly available electronic versions of structured protocols, following SPIRIT guidelines. If this were to happen, the protocol data set that is available in registries will

continue to provide valuable summaries of protocols with links to other trial related information including the full protocol, publications, trial website, systematic review, meta-analysis, results databases and research data repositories and thus continue to play an important role in achieving trial transparency.

Results databases are in their early stage of development, and they currently lack international standards. They are being formed by trial registries and aim at providing summary/aggregate results data of registered trials in predefined tables. Out of 17 general open-access registries in public domain that are linked to the WHO, only 3 developed summary clinical trial results databases: ClinicalTrials.gov, EU CRT (European Clinical Trial Register, <https://www.clinicaltrialsregister.eu/>) and Japanese registry, UMIN. As mentioned earlier, UMIN also displays IPDs. These databases differ. Each of them follows the rules of their respective countries, and at the same time, they are meeting the WHO and ICMJE request to register and share summary results. Apparently, the need to synchronize has been understood, and it seems that ClinicalTrials.gov and EMA/European Clinical Trial Registry are working on developing comparable data fields which might inform future development of international standards of data sharing.

Open-access research data repositories in public domain are certainly the most important tool for data opening and can play a major role in enabling public availability of research data. However, they are heterogenous, and there are still no international standards to govern the public disclosure of analyzable data sets which include cleaned, anonymized IPDs (i.e., usually numeric or encoded) and documentation sufficient to make the data reusable.

Development of such standards will require participation of all interested constituencies in thorough planning, analysis of quality control, resources, as well as dealing with specific issues, such as privacy, i.e., anonymization methods and practices. It is important to note that although there are no standards and guidelines for the preparation of clinical trial data for public release and although repositories are heterogenous, the existence of open-access repositories and a possibility to publish data in them are a big step forward toward opening of clinical trial data.

The progress achieved as well as the interest and expectations this data opening process has created so far is encouraging, but still a lot needs to be done. As mentioned earlier, there are numerous initiatives contributing to increasing the transparency of clinical trials and opening of its data beyond described in this chapter. There are also initiatives and projects addressing the needed standards development as mentioned CORBEL project [44]. It can be expected that this process will be observed and supported in various ways by key players at various levels, including regulators, public funders, clinicians, academia, pharmacists, journal editors, industry, patients, consumers, consumer advocates, and general public. Thus, researchers and IT experts will not be alone in this process as the clinical trials and their contribution to creation of the evidence needed for decisions in health are of paramount interests to numerous stakeholders.

The dynamics of the process are so immense and complex that they merit assessment of actions, initiatives, and practice of various players and their interactions. It is equally important to assess the impact of these dynamics on opening of

analyzable data for reuse, on the consequent transformation of clinical trial research all adjacent issues. An observatory or natural experiment is the methodology of choice to collect, assess, and disseminate such data and thus inform the process and indicate trends. The IMPACT Observatory aims to do just that and become a tool, a hub, informing the process of opening of trial data [43].

Acknowledgments The author would like to thank Nevena Jeric at Apropo Media for graphic design.

Disclaimer The views expressed here are the author's and do not represent the views of the MedILS or any other organization.

References

1. Vickers AJ. Sharing raw data from clinical trials: what progress since we first asked “Whose data set is it anyway?” *Trials* [Internet]. 2016;17:227. Available from: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4855346/>.
2. Krleža-Jerić K. Clinical trial registration: the differing views of industry, the WHO, and the Ottawa Group. *PLoS Med*. 2005;2:1093–7.
3. Clinical Study Data Request [Internet]. [cited 2016 Sep 1]. Available from: <https://www.clinicalstudycdarequest.com/>.
4. YODA Project [Internet]. [cited 2016 Jul 19]. Available from: <http://yoda.yale.edu/>.
5. Project Data Sphere [Internet]. [cited 2016 Aug 1]. Available from: <https://www.projectdatasphere.org/>.
6. Vivli-A global clinical trial data sharing platform: proposal, definition and scope background and objectives; 2016 June.
7. The Vivli Platform is live – Vivli [Internet]. [cited 2018 Aug 23]. Available from: <https://vivli.org/news/the-vivli-platform-is-live/>.
8. European Medicines Agency – clinical data publication – documents from advisory groups on clinical-trial data [Internet]. [cited 2018 Aug 21]. Available from: http://www.ema.europa.eu/ema/index.jsp?curl=pages/special_topics/document_listing/document_listing_000368.jsp&m_id=WC0b01ac05809f3f12#section1.
9. Krleža-Jerić K. Sharing of data from clinical trials and research integrity. In: Steneck N, Anderson M, Kleinert S, Mayer T, editors. *Integrity in the Global Research Arena; Proceedings of the World Conference on Research Integrity*. 3rd ed. Montreal, Quebec, Singapore: World Scientific Publishing; 2013. p. 91.
10. Krleža-Jerić K. Sharing of clinical trial data and research integrity. *Period Biol*. 2014;116(4):337–9.
11. Krleža-Jerić K. Clinical trials registries and results databases. In: Ritchesson RL, Andrews JE, editors. *Clinical research informatics*. London: Springer; 2012. p. 389–408.
12. Bass A. Side effects; a prosecutor, whistleblower, and a bestselling antidepressant on trial. 1st ed. Chapel Hill: Algonquin Books; 2008. 260 p.
13. Gibson L. GlaxoSmithKline to publish clinical trials after US lawsuit. *BMJ*. 2004;328(7455):1513.
14. DeAngelis CD, Drazen JM, Frizelle FA, Haug C, Hoey J, Horton R, et al. Clinical trial registration: a statement from the international committee of medical journal editors. *JAMA* [Internet]. 2004 [cited 2016 Jul 13];292(11):1363–4. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/15355936>.
15. Krleža-Jerić K, Chan A-W, Dickersin K, Sim I, Grimshaw J, Gluud C. Principles for international registration of protocol information and results from human trials of health related interventions: Ottawa statement (part 1). *BMJ* [Internet]. 2005 [cited 2016 Jul 14];330(7497):956–8. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/15845980>.

16. The Mexico Statement on Health Research [Internet]. Mexico City; Available from: http://www.who.int/rpc/summit/agenda/Mexico_Statement-English.pdf.
17. International Clinical Trials Registry Platform (ICTRP) [Internet]. [cited 2016 Jul 12]. Available from: <http://www.who.int/ictrp/en>.
18. Mahmić-Kaknjo M, Šimić J, Krleža-Jerić K. Setting the impact (improve access to clinical trial data) observatory baseline. Biochem Med. 2018;28(1):7–15. 010201. <https://doi.org/10.11613/BM.2018.010201>.
19. Chan AW, Tetzlaff JM, Altman DG, Laupacis A, Gøtzsche PC, Krleža-Jerić K, et al. SPIRIT 2013 statement: defining standard protocol items for clinical trials. Ann Intern Med. 2013;158(3):200–7.
20. Chan A-W, Tetzlaff JM, Gøtzsche PC, Altman DG, Mann H, Berlin JA, et al. SPIRIT 2013 explanation and elaboration: guidance for protocols of clinical trials. BMJ [Internet]. 2013;346:e7586. Available from: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3541470&tool=pmcentrez&rendertype=abstract>.
21. The SPIRIT Statement [Internet]. [cited 2018 Aug 17]. Available from: <http://www.spirit-statement.org/spirit-statement/>.
22. El Emam K, Jonker E, Sampson M, Krleža-Jerić K, Neisa A. The use of electronic data capture tools in clinical trials: web-survey of 259 Canadian trials. J Med Internet Res [Internet]. 2009;11(1):e8. Available from: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2762772&tool=pmcentrez&rendertype=abstract>.
23. WHO Trial registration Data set version 1.3.1. [Internet]. [cited 2018 Aug 19]. Available from: <http://www.who.int/ictrp/network/trds/en/>.
24. Reveiz L, Chan A-W, Krleža-Jerić K, Granados CE, Pinart M, Etxeandia I, et al. Reporting of methodologic information on trial registries for quality assessment: a study of trial records retrieved from the WHO search portal. PLoS One [Internet]. 2010;5(8):e12484. Available from: <http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0012484>.
25. WHO/The WHO Registry Network [Internet]. WHO. World Health Organization; 2016 [cited 2018 Aug 17]. Available from: <http://www.who.int/ictrp/network/en/>.
26. Reveiz L, Krleža-Jerić K, Chan A-W, de Aguiar S. Do trialists endorse clinical trial registration? Survey of a Pubmed sample. Trials [Internet]. 2007;8:30. Available from: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2147029&tool=pmcentrez&rendertype=abstract>.
27. WMA Declaration of Helsinki – ethical principles for medical research involving human subjects – WMA – The World Medical Association [Internet]. 2013 [cited 2018 Aug 20]. Available from: <https://www.wma.net/policies-post/wma-declaration-of-helsinki-ethical-principles-for-medical-research-involving-human-subjects/>.
28. Krleža-Jerić K, Lemmens T. 7th revision of the Declaration of Helsinki: good news for the transparency of clinical trials. Croat Med J [Internet]. 2009 [cited 2016 Jul 14];50(2):105–10. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/19399942>.
29. Goodyear MDE, Krleža-Jeric K, Lemmens T. The Declaration of Helsinki. BMJ [Internet]. 2007 [cited 2016 Jul 14];335(7621):624–5. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/17901471>.
30. Krleža-Jerić K, Lemmens T, Reveiz L, Cuervo LG, Bero LA. Prospective registration and results disclosure of clinical trials in the Americas: a roadmap toward transparency. Rev Panam Salud Publica [Internet]. 2011 [cited 2016 Jun 16];30(1):87–96. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/22159656>.
31. Zarin DA, Tse T, Williams RJ, Rajakannan T. The status of trial registration eleven years after the ICMJE policy. N Engl J Med [Internet]. 2017;376(4):383–91. Available from: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5813248/>.
32. Rising K, Bacchetti P, Bero L. Reporting bias in drug trials submitted to the food and drug administration: review of publication and presentation. Ioannidis J, editor. PLoS Med [Internet]. 2008;5(11):e217. Available from: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2586350/>.
33. CDISC [Internet]. [cited 2018 Aug 20]. Available from: <https://www.cdisc.org/>.

34. Harriman SL, Patel J. When are clinical trials registered? An analysis of prospective versus retrospective registration. *Trials* [Internet]. 2016;17:187. Available from: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4832501/>.
35. Viergever RF, Karam G, Reis A, Ghersi D. The quality of registration of clinical trials: still a problem. Scherer RW, editor. *PLoS One* [Internet]. 2014;9(1):e84727. Available from: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3888400/>.
36. Food and Drug Administration Amendments Act (FDAAA) of 2007 [Internet]. Office of the Commissioner; [cited 2018 Aug 22]. Available from: <https://www.fda.gov/regulatoryinformation/lawsenforcedbyfda/significantamendmentstothefdact/foodanddrugadministration-amendmentsactof2007/default.htm>.
37. Prospero-International prospective register of systematic reviews [Internet]. [cited 2018 Aug 21]. Available from: <https://www.crd.york.ac.uk/prospero/>.
38. Australia Clinical Trials ToolkitAustralian Clinical Trials [Internet]. [cited 2018 Aug 22]. Available from: <https://www.australianclinicaltrials.gov.au/clinical-trials-toolkit#overlay-context=home>.
39. Australia New Zealand Clinical Trials Registry. Data item definition/explanation [Internet]. [cited 2018 Aug 14]. Available from: <http://www.anzctr.org.au/docs/ANZCTR%20Data%20field%20explanation.pdf>.
40. Zarin DA, Tse T, Ide NC. Trial registration at ClinicalTrials.gov between May and October 2005. *N Engl J Med* [Internet]. 2005;353(26):2779–87. Available from: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1568386/>.
41. Askie LM, Hunter KE, Berber S, Langford A, Tan-Koay AG, Vu T, Sausa R, Seidler AL, Ko H SR. The clinical trials landscape in Australia 2006–2015 [Internet]. Sydney: Australian New Zealand Clinical Trials Registry; 2017 [cited 2018 Aug 19]. 83 p. Available from: <http://www.anzctr.org.au/docs/ClinicalTrialsInAustralia2006-2015.pdf#page=1&zoom=auto,557,766>.
42. Krleža-Jerić K. International dialogue on the public reporting of clinical trial outcome and results – PROCTOR meeting. *Croat Med J*. 2008;49:267–8.
43. Krleža-Jerić K, Gabelica M, Banzi R, Martinić MK, Pulido B, Mahmić-Kaknjo M, et al. IMPACT Observatory: tracking the evolution of clinical trial data sharing and research integrity. *Biochem Medica* [Internet]. 2016;26(3):308–17. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/27812300%0A>, <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5082220>.
44. Ohmann C, Banzi R, Canham S, Battaglia S, Matei M, Ariyo C, et al. Sharing and reuse of individual participant data from clinical trials: principles and recommendations. *BMJ Open* [Internet]. 2017;7(12):e018647. Available from: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5736032/>.
45. Bian Z-X, Wu T-X. Legislation for trial registration and data transparency. *Trials* [Internet]. 2010;11(1):64. Available from: <https://doi.org/10.1186/1745-6215-11-64>.
46. El Emam K, Rodgers S, Malin B. Anonymising and sharing individual patient data. *BMJ* [Internet]. 2015;350:h1139. Available from: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4707567/>.
47. Collins F. Has the revolution arrived? *Nature* [Internet]. 2010;464(7289):674–5. Available from: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5101928/>.
48. Collins FS, Green ED, Guttmacher AE, Guyer MS. A vision for the future of genomics research. *Nature* [Internet]. 2003;422:835. Available from: <https://doi.org/10.1038/nature01626>.
49. Taichman DB, Sahni P, Pinborg A, Peiperl L, Laine C, James A, et al. Data sharing statements for clinical trials: a requirement of the International Committee of Medical Journal Editors. *PLoS Med* [Internet]. 2017;14(6):e1002315. Available from: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5459581/>.
50. Gøtzsche PC. Why we need easy access to all data from all clinical trials and how to accomplish it. *Trials* [Internet]. 2011 [cited 2018 Aug 20];12(1):249. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/22112900>.
51. Zarin DA, Tse T. Sharing Individual Participant Data (IPD) within the Context of the Trial Reporting System (TRS). *PLoS Med* [Internet]. 2016;13(1):e1001946. Available from: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4718525/>.

-
52. Rockhold F, Nisen P, Freeman A. Data sharing at a crossroads. *N Engl J Med* [Internet]. 2016 [cited 2018 Aug 13];375(12):1115–7. Available from: <http://www.nejm.org/doi/10.1056/NEJMp1608086>.
 53. Eichler H-G, Abadie E, Breckenridge A, Leufkens H, Rasi G, Doshi P, et al. Open clinical trial data for all? A view from regulators. *PLoS Med* [Internet]. 2012 [cited 2016 Jul 14];9(4):e1001202. Available from: <http://dx.plos.org/10.1371/journal.pmed.1001202>.
 54. Re3Data; Registry of Research Data Repositories [Internet]. [cited 2018 Aug 12]. Available from: www.Re3data.org.
 55. Krleža-Jeric K, Hrynaszkiewicz I. Environmental Scan of Repositories of Clinical Research Data: How Far Have We Got With Public Disclosure of Trial Data? [Internet]. figshare; 2018. Available from: https://figshare.com/articles/Environmental_Scan_of_Repositories_of_Clinical_Research_Data_How_Far_Have_We_Got_With_Public_Disclosure_of_Trial_Data/_5755386.
 56. Krleža-Jeric K, Gabelica M, Mahmic-Kaknjo M, Malicki M, Utrobicic A, Simic J, et al. Setting of an Observatory of clinical trial transition regarding data sharing; IMPACT Observatory. Poster, Cochrane Colloquium Vienna, 2015. Available from: https://figshare.com/articles/Setting_of_an_Observatory_of_clinical_trial_transition Regarding_data_sharing_IMPACT_Observatory/5753226.
 57. Gabelica M, Martinic MK, Luksic D, Krleža-Jeric K. Clinical trial transparency and data repositories; an environmental scan of the IMPACT (Improving Access to Clinical Trial Data) Observatory. Poster, 8th Croatian Cochrane Symposium, Split. 2016. <https://doi.org/10.6084/m9.figshare.7390559.v1>. Available from: https://figshare.com/articles/Clinical_trial_transparency_and_data_repositories_an_environmental_scan_of_the_IMPACT_Improving_Access_to_Clinical_Trial_Data_Observatory/7390559.
 58. UMIN-ICDR Individual Case data repository [Internet]. [cited 2018 Aug 12]. Available from: <http://www.umin.ac.jp/icdr/index.html>.
 59. Edinburgh DataShare [Internet]. [cited 2018 Aug 19]. Available from: <https://datashare.is.ed.ac.uk/>.
 60. The Dataverse Project [Internet]. [cited 2018 Aug 19]. Available from: <https://dataverse.org/>.
 61. Harvard Dataverse [Internet]. [cited 2018 Aug 20]. Available from: <https://dataverse.harvard.edu/>.
 62. Research Data Alliance RDA [Internet]. [cited 2018 Aug 17]. Available from: <https://www.rd-alliance.org/about-rda/who-rda.htmlNo.Title>.
 63. CoreTrustSeal [Internet]. [cited 2018 Jun 28]. Available from: <https://www.coretrustseal.org/about>.
 64. Pampel H, Vierkant P, Scholze F, Bertelmann R, Kindling M, Klump J, et al. Making research data repositories visible: the re3data.org registry. *PLoS One*. 2013;8(11):e78080.
 65. Persistent Identifier [Internet]. [cited 2018 Aug 19]. Available from: https://en.wikipedia.org/wiki/Persistent_identifier.
 66. DataCite [Internet]. [cited 2018 Aug 12]. Available from: <https://www.datacite.org/index.html>.



Future Directions in Clinical Research Informatics

22

Peter J. Embi

Abstract

Given the rapid advances in biomedical science, the growth of the human population, and the escalating costs of health care, the need to accelerate the pace of biomedical discoveries and their translation into health-care practice will continue to grow. Indeed, the need for more efficient and effective support of clinical research to enable the development, evaluation, and implementation of cost-effective therapies is more important now than ever before. Furthermore, the fundamentally information-intensive nature of such clinical research endeavors and the growth in both health technology adoption and health-related data available for interventions and analytics beg for the solutions offered by CRI. As a result, the demand for informatics professionals who focus on the increasingly important field of clinical and translational research will increase. Despite the progress made to date, new models, tools, and approaches will be needed to fully leverage and mine these digital assets and improve CRI practice, and this innovation will continue to drive the field forward in the coming years.

Keywords

Clinical research informatics · Biomedical informatics · Translation research · Electronic health records · Future trends · US policy initiatives · Health IT infrastructure · Data analytics · Learning health systems · Evidence-generating medicine

P. J. Embi, MD, MS (✉)

Regenstrief Institute, Inc, and Indiana University School of Medicine, Indianapolis, IN, USA
e-mail: pembi@regenstrief.org

As evidenced by the production of the new edition of this book and reflected in its chapters, clinical research informatics (CRI) has clearly become established as a distinct and important biomedical informatics subdiscipline [1]. Given that clinical research is a complex, information- and resource-intensive endeavor, one comprised of a multitude of actors, workflows, processes, and information resources, this is not a surprise. As described throughout the text, the myriad stakeholders in CRI, and their roles in the health care, research, and informatics enterprises, are continually evolving, fueled by technological, scientific, and socioeconomic changes. The changing roles in health care and biomedical research bring new challenges for research conduct and coordination but also bring potential for new research efficiencies, more rapid translation of results to practice, and enhanced patient benefits as a result of increased transparency, more meaningful participation, and increased safety.

As Fig. 22.1 depicts, the pathway from biological discovery to public health impact (the phases of translational research) clearly is served by informatics applications and professionals working in the different subdomains of biomedical informatics. Given that all of these endeavors rely on data, information, and knowledge for their success, informatics approaches, theories, and resources have and will continue to be essential to driving advances from discovery to global health. Indeed, informatics issues are at the heart of realizing many of the goals for the research enterprise.

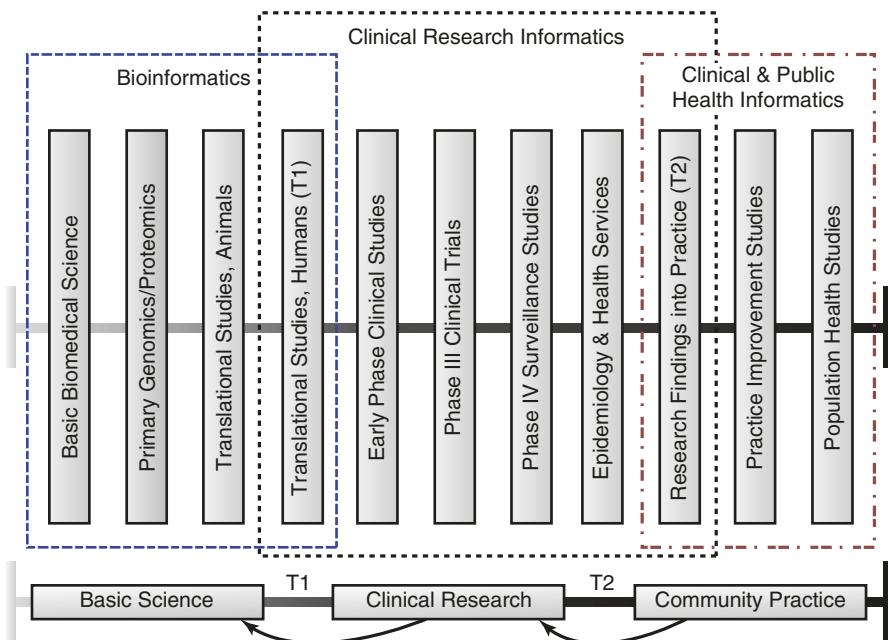


Fig. 22.1 Clinical and translational science spectrum research and informatics. This figure illustrates examples of research across the translational science spectrum and the relationships between CRI and the other subdomains of translational bioinformatics, clinical informatics, and public health informatics as applied to those efforts. (From Embi and Payne [1], with permission)

Initiatives, Policy, and Regulatory Trends in CRI

It should therefore come as no great surprise that recent years have seen the emergence of several national and international research initiatives, as well as policy and regulatory efforts focused on accelerating and improving clinical research capacity and capabilities. Indeed, a range of initiatives funded by US health and human service agencies are helping to advance the field. These include initiatives by the US National Institutes of Health (NIH), including important efforts related to the NIH Clinical and Translational Science Award (CTSA) [2, 3] programs, the establishment of visible and well-funded data science initiatives at NLM, and increased funding as a result of the twenty-first-century Cures Act toward the Cancer Moonshot and the evolution of the All of Us Research Program for advancing precision and personalized medicine.

In recent years, the CTSA program in particular has had fostered significant growth in both the practice and science of CRI and fostering professional development of CRI, given one of its major emphases the advancement of CRI, and the closely related domains of translational research informatics, translational bioinformatics, and biomedical data science efforts. Recent examples that are likely to play larger roles in the coming years, involved CRI activities that foster informatics innovations to support pragmatic and multi-site clinical research as well as recruitment innovations [4]. Other NIH activities advancing efforts related to “big data” and “data science” also have direct relevance to CRI [5, 6]. The growth of data science illustrated by the maturation of the Big Data to Knowledge (BD2K) awards the first phase designed to stimulate data-driven discovery via innovative methods, software, and training and more recently a second phase of awards designed to make the aforementioned products of research usable, discoverable, and broadly disseminated, embracing approaches that make biomedical data findable, accessible, interoperable, and reusable or “FAIR.” Additionally, other CRI-related efforts led by institutes like the National Cancer Institute (NCI) [7–10] and National Library of Medicine [11, 12] will continue to advance work in the field. Beyond NIH, funders like the Agency for Healthcare Research and Quality (AHRQ) and the Patient-Centered Outcomes Research Institute (PCORI) are also driving advances in research data methods and techniques for CRI-related efforts, including comparative effectiveness and health services research [13–15].

In addition to such initiatives focused on advancing the science and practice of CRI, investments by institutions and by the government through the US Department of Health and Human Services (DHHS), the US Office of the National Coordinator for Health Information Technology (ONC), and the US Centers for Medicare and Medicaid Services (CMMS) have incentivized the adoption and “meaningful use” of electronic health records (EHRs). The Medicare Access and CHIP Reauthorization Act of 2015 (MACRA) emphasizes the use of patient registries for quality measurement and reporting. The resultant widespread health IT infrastructure now in place, while initially focused primarily on improving patient care, is starting to enable interoperable infrastructure that is allowing for data reuse across research networks [16–18]. While initially separate efforts, recent efforts to translate between prevailing data models and adopt common interchange standards, as well as updates to

antiquated regulatory structures should enable increased interactions and enable more robust reuse of data and information from clinical care for public health and research improvements. A driving goal, to create and enable the learning health system, is now within reach, and early examples are coming online and more are likely to follow [19].

Creating Learning Health Systems Data and Knowledge Management, Evidence Generation, and Quality Improvement.

Just as biomedical informatics approaches and resources are essential to realizing the potential of such systems for enhancing clinical care, so too are CRI methods, theories, and tools critical to realizing the vision of a learning health system that enables systematic evidence generation and application via clinical practice [20]. Indeed, fully leveraging our health care and research investments to advance human health will require even more emphasis on making sense of the ever-increasing amounts of data generated through health care and research endeavors. It is work in the field of CRI that will enable and improve such research activities, from the translation of basic science discoveries to clinical trials to the leveraging of health-care data for population level science and health services research that enables its impact on care.

Importantly, these advances will continue to require increased effort not just to the development and management of technologies and platforms but also to the foundational science of CRI in an increasingly electronic world [21]. By facilitating all of the information-dense aspects of clinical research, population management, and quality improvement, CRI methods and resources will enable the conduct of increasingly pragmatic and rigorous research programs to generate new and impactful knowledge [22]. In fact, the now ubiquitous presence of EHRs will allow the systematic collection of essential data that will drive quality improvement research, outcomes research, clinical trials, comparative effectiveness research, and population level studies to a degree not heretofore feasible [23]. In addition to the technological and informatics underpinnings already mentioned, realizing this promise will require increased attention and efforts by experts focused on advancing the domain of CRI.

As depicted (Fig. 22.2), an informatics-enabled learning health system will enable the virtuous cycle of evidence generation and application, leveraging both real-world experiences and data, and applying increasingly computable knowledge artifacts in order to drive evidence-directed care and population management. Such a system will enable (a) the study of linkages from molecules to populations, (b) the development of tools and methods to enable evidence generation from real-world practice experience, (c) build bridges between health systems and research enterprises, and (d) enable the implementation and study of solutions to systematically improve health-care delivery.

Indeed, as the preceding chapters have also demonstrate, advances in CRI have already begun to enable significant improvements in the quality and efficiency of clinical research [24–26]. These have occurred through improvements in processes at the individual investigator level, through approaches and resources developed and implemented at the institutional level, and through mechanisms that have enabled and facilitated the endeavors' multicenter research consortia to drive team science. As research becomes increasingly global, initiatives like those mentioned above

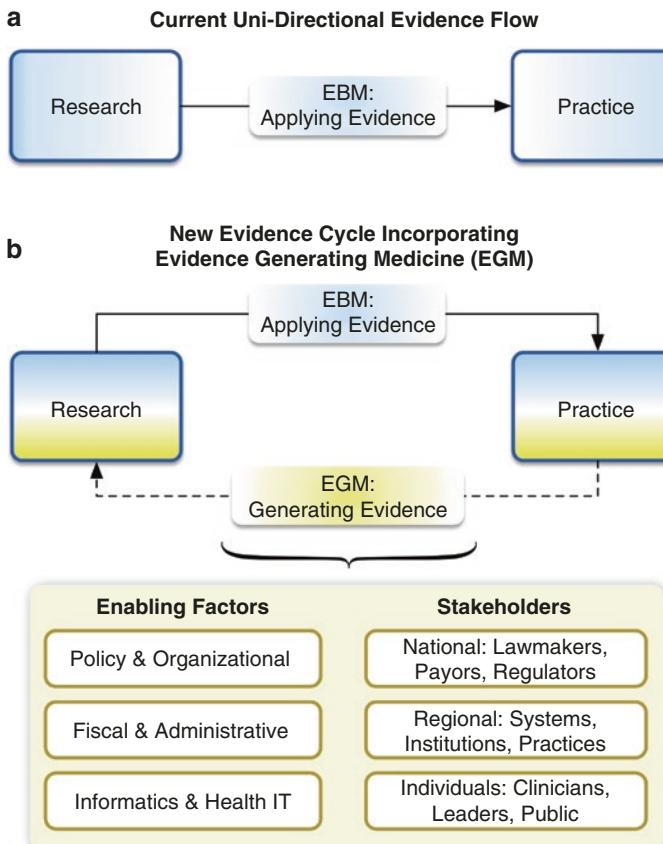


Fig. 22.2 Enabling a virtuous cycle of EBM and EGM is critical to realizing a learning health system, and there remain numerous enabling factors and key stakeholders that must be addressed and aligned to overcoming current challenges. (From Embi and Payne [20], reproduced with permission)

provide opportunities for collaboration and cooperation among CRI professionals across geographical, institutional, and virtual borders to identify common problems, solutions, and education and training needs. Increasingly, investigators and professionals engaged in these groups are explicitly self-identifying as CRI experts or practitioners, further evidence for the establishment of CRI as an important, respected, and distinct informatics subdiscipline.

Multidisciplinary Collaboration

CRI professionals come to the field from many disciplines and professional communities. In addition to the collaborations and professional development fostered by such initiatives as the CTSA mentioned above, there is also a growing role for

professional associations that can provide a professional home for those working in the maturing discipline. The American Medical Informatics Association (AMIA) is the most well-recognized such organization. Working groups focused on CRI within organizations like AMIA continue to see considerable growth in interest and attendance over the past decade. There has also been the emergence of operational professionals often referred to as chief research information officers (CROIs) who are akin to CMIOs but focused on the research IT portfolios of academic health centers [27].

The past several years have also seen a growth in scientific conferences dedicated to CRI and the closely related informatics subdiscipline of translational bioinformatics (TBI). The main meeting hosted by AMIA has seen growing attendance and productivity among the informatics and clinical/translational research communities. In addition, journals like AMIA's *JAMIA*, *Applied Clinical Informatics*, and *JAMIA Open*, as well as other leading journals in the field, have also seen growth in CRI-focused publications. The importance of CRI has led to editorial board members with CRI expertise, and even journal space special issues are dedicated to important topics in CRI [28]. Given its growth, it is likely that journals specifically focused on this domain will emerge in the years to come. In addition, other important informatics groups and journal, such as International Medical Informatics Association (IMIA), and non-informatics associations and journals (e.g., DIA, The Society for Clinical Trials, Clinical Research Forum, and many other professional medical societies) also increasingly provide coverage and opportunities for professional collaboration among those working to advance CRI. Efforts like these continue foster the maturity and growth so critical to advancing the field.

Challenges and Opportunities

Despite these many advances, significant challenges and opportunities remain to be addressed if this relatively young discipline is to evolve and realize its full potential to accelerate and improve clinical and translational science. Indeed, as reported in 2009 by Embi and Payne, the challenges and opportunities facing CRI are myriad. In that manuscript, these were placed into 13 distinct categories that spanned multiple stakeholder groups (Fig. 22.3) [1].

This conceptualization of CRI activities includes those related to education and original (informatics) research, research support services and activities, and policy leadership. The stakeholders for all of these span the individual, institutional, and national levels and include those with clinical research as well as informatics perspectives and priorities. These broad groups of stakeholders and the wide range of diverse CRI activities should all be considered as the field evolves and as research agendas, educational and training efforts, and professional resources are developed.

One of the keys to enabling a learning health system is the ability to enable systematic evidence generation through practice. A key challenge today remains the now artificial but persistent paradigm that dictates clinical care and research

		Stakeholder(s)			
		Individual Researchers & IT/Informatics Professionals	Organizational Institutions & Organizations	National/International Funders, Regulators, Agencies	
Scope	CRI Academics & Advancement	Educational Needs Scope of CRI CRI Innovation & Investigation	X X X	X X X	
	Practice of CRI	Research Planning & Conduct Data Access, Integration & Analysis Recruitment Workflow Standards	X X X X X	X X X X	
	Society & Leadership	Socio-organizational Leadership & Coordination Fiscal & Administrative Regulatory & Policy Issues	X X X X	X X X X	X X X X
	Lessons Not Learned		X	X	X

Fig. 22.3 Major challenges and opportunities facing CRI. This figure provides an overview of identified challenges and opportunities facing CRI, organized into higher-level groupings by scope, and applied across the groups of stakeholders to which they apply. (From Embi and Payne [1], with permission)

activities as distinct activities that are related only in the application of research evidence to practice, via evidence-based medicine [20]. Instead, CRI activities are increasingly demonstrating and creating environments that recognized a virtuous cycle of evidence generation and application, where “Evidence Generating Medicine” (EGM) paradigm is realized. As defined, EGM involves, “the systematic incorporation of research and quality improvement considerations into the organization and practice of healthcare to advance biomedical science and thereby improve the health of individuals and populations” [20]. An EGM-enabled environment recognizes and supports the fact that (a) clinical care activities are not entirely distinct from research activities, (b) EGM must be enabled during practice to advance both research and care, (c) EGM activities are in fact ongoing, (d) advancing EGM is key to the desired EBM lifecycle, and (e) multiple enabling factors and stakeholders are essential to making this reality (Fig. 22.4) [20].

Another major challenge to be overcome in order to realize the promise of CRI is the need to address the severe shortage of professionals currently working to advance in the CRI domain. As with many biomedical informatics subdisciplines, training in CRI is and will remain interdisciplinary by nature, requiring the study of

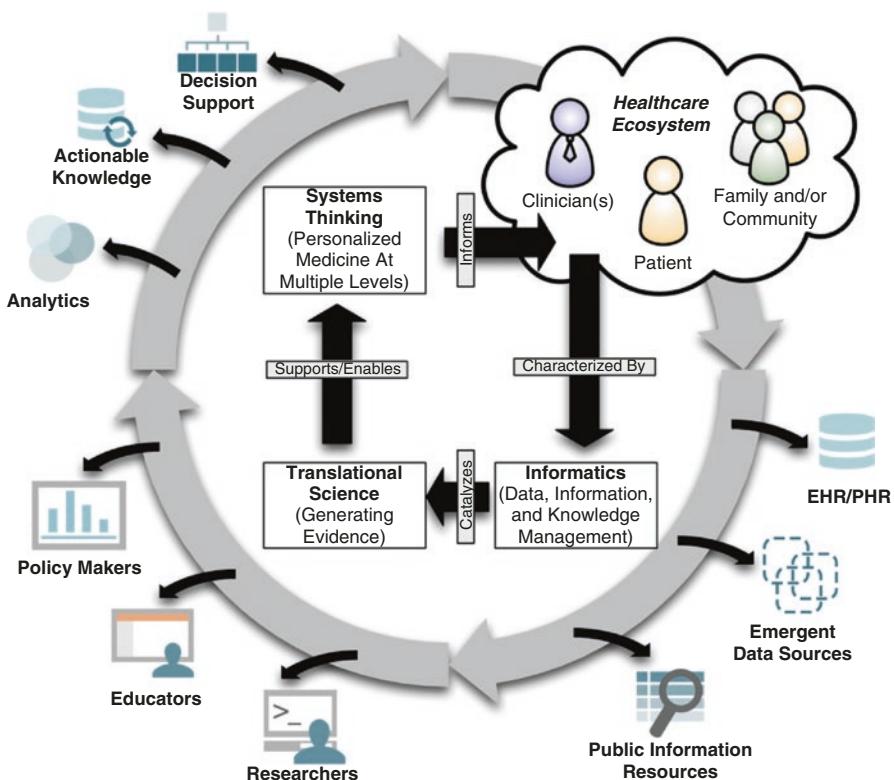


Fig. 22.4 Creating an informatics-enabled evidence-generating medicine (EGM) system: the virtuous cycle of evidence generation and application that fuels a learning health system. (From: Payne and Embi [29], reproduced with permission)

topics ranging from research methods and biostatistics, to regulatory and ethical issues in CRI to the fundamental informatics and IT topics essential to data management in biomedical science. As the content of this very book illustrates, the training needed to adequately equip trainees and professionals to address the complex and interdisciplinary nature of CRI demands the growth of programs focused specifically in this area.

Furthermore, while there is certainly a clear need for more technicians convergent in both clinical research and biomedical informatics to work in the CRI space, there remains a great need for scientific experts working to innovate and advance the methods and theories of the CRI domain. In recent years, the National Library of Medicine, which has long supported training and infrastructure development in health and biomedical informatics, recognized this need by clearly calling out clinical research informatics as a domain of interest for the fellowship training programs it supports. While most welcome and important, the availability of such training and education remains extremely limited. Significantly, more capacity in training and education programs focused on CRI will be needed to establish and grow the cadre

of professionals focused in this critical area if the goals set forth for the biomedical science and health-care enterprise are to be realized. This will require increased attention by sponsors and educational institutions.

In addition to training the professionals who will focus primarily in CRI to advance the domain, there is a major need to also educate current informaticians, clinical research investigators and staff, and institutional leaders concerning the theory and practice of CRI. Programs like AMIA's 10×10 initiative and tutorials at professional meetings offer examples like a course focused in CRI that help to meet such a need [30]. Such offerings help to ensure that those called upon to satisfy the CRI needs of our research enterprise are able to provide appropriate support for utilization of CRI-related methods or tools, including the allocation of appropriate resources to accomplish organizational aims.

As the workforce of CRI professionals grows, the field can be expected to mature further. While so much of the current effort of CRI is quite appropriately focused on the proverbial “low-hanging fruit” of overcoming the significant day-to-day IT challenges that plague our traditionally low-tech research enterprise, significant advances will ultimately come about through a recognition that biomedical informatics approaches are crucial centerpieces in the clinical research enterprise. Indeed, just as the relationship between clinical care and clinical research is increasingly being blurred as we move toward the realizing of a “learning health system,” so too are there corollaries to be drawn between the current formative state of CRI and the experiences learned during the early decades of work in clinical informatics. Those working to lead advances in CRI would do well to heed the lessons learned from the clinical informatics experiences of years past. Future years can be expected to see CRI not only instrument, facilitate, and improve current clinical research processes, but advances can be expected to fundamentally change the pace, direction, and effectiveness of the clinical research enterprise and discovery. Toward that end, groups are already working to develop maturity models and deployment indices that can be used to measure and compare CRI infrastructures as to their level of maturity and ability to support the research enterprise [31]. Such measures of CRI maturity will only grow and become more useful to inform progress in the years to come. Guided by such measures, we should expect to see CRI efforts continue to improve, with consequent improvements to scientific discovery, healthcare quality, and real-world evidence generation as learning health systems continue to evolve and mature.

Conclusion

In conclusion, the future is bright for the domain of CRI. Given the rapid advances in biomedical discoveries, the growth of the human population, and the escalating costs of health care, there is an ever-increasing need for clinical research that will enable the testing and implementation of cost-effective therapies at the exclusion of those that are not. The fundamentally information-intensive nature of such clinical research endeavors begs for the solutions offered by CRI. As a result, the demand for informatics professionals who focus on the increasingly important field of clinical and

translational research will only grow. New models, tools, and approaches must continue to be developed to achieve this, and the resultant innovations are what will continue to drive the field forward in the coming years. It remains an exciting time to be working in this critically important area of informatics study and practice.

References

1. Embi PJ, Payne PR. Clinical research informatics: challenges, opportunities and definition for an emerging domain. *J Am Med Inform Assoc.* 2009;16(3):316–27.
2. Zerhouni EA. Translational and clinical science – time for a new vision. *N Engl J Med.* 2005;353(15):1621–3.
3. Zerhouni EA. Clinical research at a crossroads: the NIH roadmap. *J Investig Med.* 2006;54(4):171–3.
4. NCATS. CTSA Trial Innovation Network. <https://ncats.nih.gov/ctsa/projects/network>.
5. Bourne PE, Bonazzi V, Dunn M, Green ED, Guyer M, Komatsoulis G, Larkin J, Russell B. The NIH big data to knowledge (BD2K) initiative. *J Am Med Inform Assoc.* 2015;22(6):1114. <https://doi.org/10.1093/jamia/ocv136>. No abstract available.
6. NIH Strategic Plan: <https://www.nih.gov/sites/default/files/about-nih/strategic-plan-fy2016-2020-508.pdf>.
7. Oster S, Langella S, Hastings S, et al. caGrid 1.0: an enterprise grid infrastructure for biomedical research. *J Am Med Inform Assoc.* 2008;15(2):138–49.
8. Saltz J, Oster S, Hastings S, et al. caGrid: design and implementation of the core architecture of the cancer biomedical informatics grid. *Bioinformatics.* 2006;22(15):1910–6.
9. Niland JC, Townsend RM, Annechiarico R, Johnson K, Beck JR, Manion FJ, Hutchinson F, Robbins RJ, Chute CG, Vogel LH, Saltz JH, Watson MA, Casavant TL, Soong SJ, Bondy J, Fenstermacher DA, Becich MJ, Casagrande JT, Tuck DP. The cancer biomedical informatics grid (caBIG): infrastructure and applications for a worldwide research community. *Fortschr Med.* 2007;12(Pt 1):330–4. PMID: 17911733.
10. Kakazu KK, Cheung LW, Lynne W. The cancer biomedical informatics grid (caBIG): pioneering an expansive network of information and tools for collaborative cancer research. *Hawaii Med J.* 2004;63(9):273–5.
11. Citation to clinicaltrials.gov final rule: <https://prsinfo.clinicaltrials.gov>.
12. Citation to common rule change: <https://www.hhs.gov/ohrp/regulations-and-policy/regulations/finalized-revisions-common-rule/index.html>.
13. Holte E, Segal C, Lopez MH, Rein A, Johnson BH. The electronic data methods (EDM) forum for comparative effectiveness research (CER). *Med Care.* 2012;50(Suppl):S7–10. <https://doi.org/10.1097/MLR.0b013e318257a66b>.
14. Fleurence RL, Curtis LH, Califf RM, Platt R, Selby JV, Brown JS. Launching PCORnet, a national patient-centered clinical research network. *J Am Med Inform Assoc.* 2014;21(4):578–82. <https://doi.org/10.1136/amiainjnl-2014-002747>. Epub 2014 May 12.
15. PCORNet PPRN Consortium, Daugherty SE, Wahba S, Fleurence R. Patient-powered research networks: building capacity for conducting patient-centered clinical outcomes research. *J Am Med Inform Assoc.* 2014;21(4):583–6. <https://doi.org/10.1136/amiainjnl-2014-002758>. Epub 2014 May 12.
16. Califf RM. The patient-centered outcomes research network: a national infrastructure for comparative effectiveness research. *N C Med J.* 2014;75(3):204–10. <https://www.ncbi.nlm.nih.gov/pubmed/24830497>.
17. Hripcak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, Suchard MA, Park RW, Wong IC, Rijnbeek PR, van der Lei J, Pratt N, Norén GN, Li YC, Stang PE, Madigan D, Ryan PB. Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. *Stud Health Technol Inform.* 2015;216:574–8. PMID:26262116.

18. Klann JG, Abend A, Raghavan VA, Mandl KD, Murphy SN. Data interchange using i2b2. *J Am Med Inform Assoc.* 2016;23(5):909–15. <https://doi.org/10.1093/jamia/ocv188>. Epub 2016 Feb 5. PMID: 26911824.
19. Friedman CP, Wong AK, Blumenthal D. Achieving a nationwide learning health system. *Sci Transl Med.* 2010;2(57):57cm29.
20. Embi PJ, Payne PR. Evidence generating medicine: redefining the research-practice relationship to complete the evidence cycle. *Med Care.* 2013;51(8 Suppl 3):S87–91. <https://doi.org/10.1097/MLR.0b013e31829b1d66>. PMID: 23793052.
21. Payne PR, Embi PJ, Niland J. Foundational biomedical informatics research in the clinical and translational science era: a call to action. *J Am Med Inform Assoc.* 2010;17(6):615–6.
22. Richesson RL, Green BB, Laws R, Puro J, Kahn MG, Bauck A, Smerek M, Van Eaton EG, Zozus M, Hammond WE, Stephens KA, Simon GE. Pragmatic (trial) informatics: a perspective from the NIH health care systems research collaboratory. *J Am Med Inform Assoc.* 2017;24(5):996–1001. <https://doi.org/10.1093/jamia/ocx016>.
23. Embi PJ, Kaufman SE, Payne PRO. Biomedical informatics and outcomes research. *Circulation.* 2009;120:2393–9., Originally published December 7, 2009. <https://doi.org/10.1161/CIRCULATIONAHA.108.795526>.
24. Payne PR, Johnson SB, Starren JB, Tilson HH, Dowdy D. Breaking the translational barriers: the value of integrating biomedical informatics and translational research. *J Investigig Med.* 2005;53(4):192–200.
25. Sung NS, Crowley WF Jr, Genel M, et al. Central challenges facing the national clinical research enterprise. *JAMA.* 2003;289(10):1278–87.
26. Chung TK, Kukafka R, Johnson SB. Reengineering clinical research with informatics. *J Investigig Med.* 2006;54(6):327–33.
27. Sanchez-Pinto LN¹, Mosa ASM, Fultz-Hollis K, Tachinardi U, Barnett WK, Embi PJ. The emerging role of the chief research informatics officer in academic health centers. *Appl Clin Inform.* 2017;8(3):845–53. <https://doi.org/10.4338/ACI-2017-04-RA-0062>.
28. Embi PJ, Payne PR. Advancing methodologies in clinical research informatics (CRI): foundational work for a maturing field. *J Biomed Inform.* 2014;52:1–3. <https://doi.org/10.1016/j.jbi.2014.10.007>. No abstract available.
29. Payne PRO, Embi PJ, editors. *Translational informatics: realizing the promise of knowledge-driven healthcare.* London: Springer; 2014.
30. The Ohio State University-AMIA 10x10 program in Clinical Research Informatics. <http://www.amia.org/education/academic-and-training-programs/10x10-ohio-state-university>. Accessed 14 Jul 2011.
31. Knosp BM, Barnett W, Embi PJ, Anderson N. Maturity models for research IT and Informatics reports from the field. In: Proceedings of the AMIA summit on clinical research informatics; 2017. p. 18–20. <https://knowledge.amia.org/amia-64484-cri2017-1.3520710/t001-1.3521784/t001-1.3521785/a011-1.3521792/ap011-1.3521793#pdf-container>.

Index

A

- Academic health centers (AHCs), 34–38
Acceliant’s ePRO platform, 260
Advancing Clinical Genomic Trials on Cancer Master Ontology (ACGT-MO), 330
AdaBoost model, 369
Adaptive randomization methods, 70
Adverse Drug Reaction Classification System (ADReCS), 442
Adverse event (AE), 32, 405, 438, 439
Agency for Healthcare Research and Quality (AHRQ), 271, 272
American Health Information Management Association (AHIMA), 294, 421
American Medical Informatics Association (AMIA), 5, 95, 486
Analog signal processing, 18–19
Analog to digital converters (ADC), 19
Analyzable data, reuse of, 474
Anatomical therapeutic chemical (ATC), 323
ANN, *see* Artificial neural network
Anonymization, 382
Application programming interface (API), 326, 328
Argonaut, 419
Artificial neural networks (ANNs), 346, 347, 363, 366
AskHERMES, 369
Audiovisual computer-assisted self-interviewing (A-CASI) systems, 259, 260

B

- Bayesian network, 112, 346, 349, 363, 364, 369
Big data, discovery and application, 446
Bioinformatics, 442

- Biomedical ontologies, 315, 434, 440, 482
Biomedical Research Integrated Domain Group (BRIDG), 203, 204, 318, 319, 382, 416
BioPortal Annotator, 327–329
BIRNLex, 331
BRIDG, *see* Biomedical Research Integrated Domain Group
Business associate agreement (BAA), 299

C

- Canadian Institutes of Health Research (CIHR), 469
Cancer biomedical informatics grid (caBIG) program, 158–159, 321, 353
Cancer Data Standards Repository (caDSR), 406, 429
Cancer Genome Project, 161
Capability Maturity Model, 241
CarePlan, 413
Case report forms (CRFs), 31, 194
CDISC, *see* Clinical Data Interchange Standards Consortium
Centers for Medicaid and Medicare Services (CMS), 271, 410
Chemical entities of biological interest (ChEBI), 319
Chief Clinical Information Officer (CCIO), 305
Chief Health Information Officer (CHIO), 305
Chief Information Officer (CIO), 305
Chief Medical Information Officer (CMIO), 305
China Food and Drug Administration (CFDA), 388
Classification and regression trees (CART), 348

- Clinical and translational science awards (CTSA) program, 325
- Clinical/Contract Research Organizations (CROs), 35, 36
- Clinical Data Acquisition Standards
Harmonization (CDASH) standards, 385, 391, 406
- Clinical Data Interchange Standards
Consortium (CDISC), 13, 202, 381, 385, 404, 465
- CDASH standards, 406
- Operational Data Model, 407
- Retrieve Form Data Capture, 396
- Clinical data, reusing, 380
- Clinical decision support (CDS) logic, 410
- Clinical decision-making, 456
- Clinical Information Modeling Initiative (CIMI), 415
- Clinical research
academic health centers, 34–35
- administrative managers/coordinators, 38
- big science emergence
modern astronomy, 23
- particle physics, 23
- social transformation, 24
- socially interdependent process, 24
- biomedical data, 20
- budgeting and fiscal reconciliation, 32
- clinical study, screening and enrolling participants in, 30
- complexity
computing capacity, 20
- information processing, 20
- complex technical and communications processes, 41
- computational power, 21
- contexts and attempts, 5, 6
- CROs, 35–36
- data and information systems, 10, 11
- data and information management requirements in, 39–40
- data-driven discovery, 12–14
- data quality, 31
- design patterns, 38, 39
- DSMBs, 38
- emerging policy trends, 106, 107
- evidence generating medicine, 43–44
- federal regulatory agencies, 37
- foundations of, 9, 10
- fundamental theorem, 4
- healthcare and clinical research
information systems vendors, 37–38
- history, 18
- human subjects protection reporting and monitoring, 32
- information exchange, 41–42
- cognitive complexity, 42
- innovation, 42
- interruptions, 41
- interventional clinical trial phases and associated execution-oriented processes, 28–29
- knowledge representation, 12, 13
- learning healthcare systems, 43–44
- local storage, 21
- network capacity, 21
- paper-based information management practices, 41
- patients and advocacy organizations, 33–34
- potential study participants, 30
- precision/personalized medicine, 43
- programs, 38
- recruitment
computational solutions, 112, 113
- computer-based medical records systems, 113, 114
- data repositories, 115
- EHR systems, 114, 115
- sociotechnical challenges, 116, 117
- workflows, 110, 111
- regulatory and sponsor reporting and administrative tracking, 31
- RWE generation, 44–45
- scope, 7–9
- sponsoring organizations, 36
- stakeholders, 33
- standards
comparable information, 25
- consistent information, 25
- constructs, 25
- interoperable systems, 25
- study encounters and associated data collection tasks, 31
- study-related events, scheduling and tracking, 30–33
- tasks and barriers, 32–33
- telephonic signals, 19
- workflow, 40
- Clinical research funding, 28
- Clinical research information systems (CRISs)
clinical research subjects, 177
- concepts, 173
- current inefficiencies, 194, 195
- EHR-related systems, 173
- electronic data capture

- certain experimental designs, 180
certain research designs, 180
cross-field validation, 179
data library, 180, 181
dynamic lists, 180
skip logic, 180
specific privileges, 180
validation, 179
essential functions, 173
events, 177
implementation
 representing experimental designs, 174
 supports multiple studies, 173, 174
Initiatives, Policy and Regulatory Trends, 483, 484
pragmatic clinical trials
 analysis, 188
 comparison therapy choice
 flexibility, 188
 follow-up intensity, 188
 outcome, 188
 patient selection criteria, 188
 personnel, 188
 practitioner adherence, 188
 subject compliance, 188
 therapeutic flexibility, 188
quality control, 181
real-time self-reporting, 178
scope, 174, 175
standards, 187
structured data, 178, 179
study protocol, 193, 194
study stages
 patient-monitoring and safety, 185, 186
 planning and protocol authoring,
 182–184
 protocol management, 184, 185
 recruitment and eligibility
 determination, 184
time windows, 176, 177
validation and certification, 186, 187
vendor models, 172
workflow, 176
Clinical trial management systems, 405, 406
Clinical trial registry, 459, 461, 463
Clinical Trials Transformation Initiative (CTTI), 271, 419
ClinicalASR, 369
Coalition For Accelerating Standards and Therapies (CFAST), 395
Coasian transactions, 435–437
Code of Federal Regulations (CFR), 92
Cognitive complexity, 41, 42
Common clinical registry framework (CCRF)
 model, 279, 280
Common data model (CDM), 283, 320, 440
Common Protocol Template (CPT), 395
Common Rule, 92–94, 96, 98, 294, 299, 300
 common rule revisions, 94–96
 data sharing policies, 105, 106
 food and drugs regulation and guidance, 96
foundational federal legislation
 Food, Drug and Cosmetic Act of 1938,
 89, 90
 Public Health Services Act of 1944, 91
HIPAA privacy rule and research, 97
regulatory science
 digital health, 102, 103
 real-world evidence, 100–102
21st Century Cures Act, 103
Complete crossover design, 76
Comprehensive Health Enhancement Support System (CHESS), 132
Computable study protocol
 accurate data capture, 199, 200
 complete study plan, 197
 computability and standardization, 201
 decision support, 198, 199
 facilitating timely, 199, 200
 interpretation and application of results,
 200, 201
 statistical analysis and reporting, 200
 study data and artifacts, 201
Computational approach, 445
 signal detection, 440
Computer adaptive testing (CAT) system, 265
Computer-based medical records systems, 113
Computerized physician order entry (CPOE) systems, 40
Computer-mediated support groups (CMSP), 132
Computing validity, 255
Concurrent validity, 255
Conditional random fields (CRFs), 363, 367, 370
Consolidated clinical document architecture (C-CDA), 275
Construct validity, 255
Content standards, 275–277
Content validity, 254–255
Continuity of Care Document (CCD), 418
Convergent validity, 255
Coordinated Research Infrastructure Building Enduring Life-science Services (CORBEL), 390, 395
Core Protocol version 1, 104

CoreTrustSeal certification organization, 473, 474
 Covered entity, 299
CRF, *see* Case report forms
CRIS, *see* Clinical research information systems
 Criterion validity, 255
 Critical Path Institute (C-Path), 395, 418
 Cronbach's alpha, 253, 254
 Current procedural terminology (CPT), 321, 324

D

Data content ontology, 320, 321
 Data exchange standards, 275
 Data governance, 239–241
 clinical data management, 217, 218
 coherent data governance program, 293
 conceptual model, 292
 data and information governance program, 307
 data lifecycle, 297, 298
 data manifold, 295
 data protection to research ethics, 299–302
 data-information-knowledge, 296
 decision matrix, 308
 definition, 292, 293
 electronic patient data, 293
 implementation, 306–309
 information governance, 295, 302–304
 life cycle of data, 297–298
 master data, 309
 organization and roles, 305, 306
 outcomes, 294
 policies, 158
 research, 293–294
 structures and processes, 294
 value of data, 296, 297
 Data integrity, 388
 Data mining, 342, 343, 345, 346
 tools, 39
 Data model, 439
 Data safety and monitoring boards (DSMBs), 38
 Data Seal of Approval (DSA), 474
 Data sharing, 380, 382, 471
 collaborations, initiatives and tools, 394, 396
 projects and networks, 386
 reuse, benefits, 383
 Data standards, adoption and implementation of, 391, 392

Data storage
 analytic sophistication, 22, 23
 data density, 22
 design complexity, 22
 DataCite, 474
 Decision tree (DT), 348, 363, 365
 Decision-support systems, 40
 Declaration of Helsinki (DoH), 464
 Deep belief networks (DBNs), 368
 Deep learning approach, 367, 368, 370
 De-identification, 299, 382
 Delayed hepatotoxicity, 444, 445
 Deming Wheel, 389
 Depression, 132, 134, 447
 DermLex, 328
 Description logics (DLs), 315
 Digital Health, 100, 102–103
 Digital Imaging and Communications in Medicine (DICOM), 411
 Digital signal processing (DSP), 19, 20
 Digital to analog converters (DACs), 19
 Digitization of healthcare data, 435
 Digitize action collaborative, 416
 Directed acyclic graph (DAG), 364
 Discovery science, 28
 Discriminant validity, 255
 Dose-titration design, 77
 Drug therapy, 446
 DrugBank, 323, 332

E

Economical translation of data, 441
 eDiaries, 384
 EHR4CR project, *see* Electronic Health Records for Clinical Research
 EHR, *see* Electronic health record
 Electronic case report form (eCRF), 382
 Electronic clinical research study, 392
 Electronic data capture (EDC), 31, 283
 Electronic data management, 471
 Electronic health data, 381
 Electronic health record (EHR), 39, 114, 171, 220, 332, 358, 442, 443
 clinical natural language, 361
 converting clinical data to research variables challenges, 372, 373
 data and data sharing, 389
 data quality issues, 372
 definition, 358
 eligibility screening, 359, 360
 legal computerized medical record, 358
 low-risk clinical studies, 175, 176

- patient-encounter data, 181
patient-oriented research, 358
patient phenotype retrieval, 359, 360
sample, 359
secondary use of data, 360
standards-based structured laboratory test, 358
- Electronic Health Records for Clinical Research (EHR4CR), 386, 396
- Electronic mail, 258, 259
- Electronic medical record (EMR), 202, 207, 208, 319
- Electronic medical records and genomics (eMERGE) network, 332
- Electronic population of data, 386
- Electronic/computable phenotyping, 332
- Eligibility Rule Grammar and Ontology (ERGO) project, 205
- Employer identification number (EIN), 277
- Enterprise data warehouses (EDW), 353
 data marts or registries, 410
- Epidemiological studies, 54
 prospective studies, 55
 retrospective studies, 55
- e-protocol, 191, 193–202
- Equivalence/non-inferiority studies, 65, 253
- eSource
 data interchange, 381, 384, 385
 implementations, 384, 387–389
 methodology assessments, 389
 process, 386
 solution, 386
- E2B data model, 440
- European Clinical Research Infrastructure Network (ECRIN), 395
- European Medicines Agency (EMA), 404
- Evidence generating medicine, 43–44
- Evidence-informed decision-making, 455
- Executed study protocol, 192
- Experimental designs
 antidotes against bias
 blinding, 70
 cluster randomization, 70
 double-blind clinical trial, 71
 simple randomization, 69
 stratified randomization, 70
 basic concepts, 67
 crossover designs, 74, 75
 variants of, 77
 definitions, 67
 innovative approaches, 78, 82, 83
 parallel group designs, 71, 72, 74
 variants of, 77
- single treatment group, 68
- Experimental studies
 between-group studies, 56
 clinical trial, 56–58
 study treatments
 concomitant treatments, 64
 control treatment, 63
 experimental treatment, 63
- superiority vs. non-inferiority, equivalence/non-inferiority studies, 65–67
- treatment effect definition
 end-point to group indicator, 60
 group indicator to signal, 60
 measurement to end-point, 59
 measurements, 58
 within-group studies, 56
- Expert determination, 300
- Exposomics, 149
- External AEs, 32
- F**
- Factorial designs, 77
- Fast Healthcare Interoperability Resources (FHIR), 275, 385, 396, 411, 412
- Fayyad's knowledge discovery, 343
- Federal regulatory agencies, 37
- Federal-wide assurance (FWA), 300, 301
- Food and Drug Administration (FDA), 172–173
- Food and Drug Administration Amendments Act (FDAAA), 271, 467
- Food and Drug Administration's (FDA) sentinel initiative, 397
- Food, Drug and Cosmetic (FD&C) Act, 89, 90
- Function and outcomes research for
 comparative effectiveness in
 total joint replacement (FORCE-TJR), 272
- G**
- GenBank, 24, 149–151
- Gene ontology (GO), 153, 157, 316, 319, 328
- General health data exchange standard, 411
- Genomes project, 155
- Genomics metadata, 412
- Global medical device nomenclature agency, 317
- Glomerular filtration rate (GFR), 372
- Good Clinical Practice (GCP), 6
- Graphical processing units (GPU), 21
- GRU RNN model, 370

H

Healthcare information technology (HIT) platforms, 37–38

Health consumerism clinical trial involvement, 138, 139 dynamic relationship with their own health data, 139 empowered consumers, 139 patient empowerment or management, 127 patient researcher, 133, 134 personalization of medicine, 140 television and radio, 134

Health data, reusing, 380, 382–383

Health Data Standards to Clinical Research, 407

Health Insurance Portability and Accountability Act (HIPAA), 91, 97, 299

Health Level Seven (HL7), 202, 274, 396, 411 clinical document architecture standard, 418

Patient Care workgroup, 279, 280 reference information model (RIM), 320

Health outcome of interest (HOI), 443

Health plan identifier (HPID), 277

Healthcare delivery systems, 404

Healthcare Information and Management Systems Society (HIMSS), 419

Healthcare information systems, 407, 413, 417

Healthcare Information Technology IT Standards Panel (HITSP), 385

Healthcare Standards, 409–411

HedgeScope, 369

Hidden Markov models (HMM), 363, 364, 367

Health Insurance Portability and Accountability Act (HIPAA), 91, 110, 116, 174, 277, 371 data governance, 299 privacy rule, 97–98, 299 research, 97–98

HIPAA, *see* Health Insurance Portability and Accountability Act

Honest broker, 301, 305

Hot deck method, 348

Human genome project, 148

Human subjects protection reporting, 32

I

ICD-10-CM, 275, 276, 324

IMPACT Observatory, 457, 470, 477

Incomplete crossover designs, 76

Individual Participant Data (IPD), 454 direct researcher-to-researcher contact, 454, 455 initiatives and project, 454

IPD-based meta-analyses, 456 pooled and meta-analysis, 454 publicly accessible repositories, 454 research data repositories, 455

Infectious Diseases Data Observatory (IDDO), 396

Information and communication technologies (ICTs), 124

Information architecture, 300

Information-based recruitment workflows, 111

Information governance, 302–304, 307

Innovative Medicine Initiative (IMI), 396

Institutional Review Board (IRB), 93, 115, 301

Integrated and interoperable health information systems, 408

Integrating the Health Enterprise (IHE), 396, 419

Interactive voice response (IVR) systems, 259

Internal AEs, 32

Internal consistency reliability, 254

Internal Review Board (IRB), 305

International Classification for Nursing Practice (ICNP), 317

International classification of diseases (ICD), 317, 321, 323, 324

International Classification of Primary Care, 327

International Conference on Harmonization (ICH), 196, 216

International Medical Informatics Association (IMIA), 486

International Patient Summary Implementation Guide, 420

International research community, 458

International Society for Pharmacoeconomics Outcomes Research (ISPOR), 263

International Standard Randomized Clinical Trials Number (ISRCTN), 132, 458

International standards landscape, 416

International trial registration standards, 458

Interoperability, 37, 105, 159, 185, 203, 274, 280, 353, 381, 385, 387, 396–397, 411, 415, 419

CCRF model, 279, 280

clinical models and data elements, 276, 277

clinical phenotype, 278
coding systems and controlled terminologies, 276
content standards, 275
data exchange standards, 275
outcome measures, 279
UDI, 277, 278
Interrater reliability, 254
Intraclass correlations (ICC), 253
Item anchors, 261
Item response theory (IRT), 265

J

Janus clinical trials repository, 330
Joint Initiative Council (JIC), 417

K

Kappa coefficients, 253, 254
KDD, *see* Knowledge discovery in databases
KEGG, 332
k nearest neighbor (*k*-NN), 366
classification method, 348
Knowledge discovery
artificial neural networks, 346, 347
association rule, 349
Bayesian networks, 349
data mining, 345
data selection, 343, 344
decision trees, 348
enterprise data warehouses, 353
Fayyad's knowledge discovery, 343
infrastructure for, 353, 354
interpretation and evaluation,
349, 350
k-nearest neighbor classification
method, 348
knowledge discovery in databases
process, 342, 343
limitations, 352
mitigation of bias, 351, 352
preprocessing, 344
PRISMS, 354
rare instances, 351
ROC curve, 350
support vector machine methods, 348
training/testing curves, 347
transformation, 345
Utah PRISMS Center, 354
Knowledge discovery in databases (KDD),
342, 344, 350–353

L

Latin square design, 76, 77
Learning based approach, 369
Learning Health Community (LHC),
392, 396
Learning health systems (LHS), 273, 303,
408, 486
core values, 392, 393
Linked Data, 332
Logical observation identifiers, names and
codes (LOINC), 276, 317,
320–322, 333
Logistic regression, 363

M

Machine learning approach, 422, 441
algorithmic computation, 438
combined with statistical techniques, 443
discriminative model
ANN model, 366
CRFs, 367
decision trees, 365
K-NN method, 366
logistic regression, 365
SVM model, 366
generative model
Bayesian network, 364
HMM, 364
MRF model, 364, 365
Naive Bayes classifier, 363
high quality curated datasets, 444
unsupervised clustering, 367
Mapping process, 414
Markov random field (MRF) model, 363–365
Maximum entropy Markov models
(MEMMs), 367
MedDRA, *see* Medical Dictionary for
Regulatory Activities
Medical and industrial research, 441
Medical device epidemiology network
(MDEpiNet), 277
Medical Dictionary for Regulatory Activities
(MedDRA), 327, 429, 440, 442
Medical language extraction and encoding
(MedLEE) system, 117, 368
Medidata, 260
Merit-based incentive payment system
(MIPS), 322
Metabolomics, 149, 154
Microbiomics, 149
Mobile devices, 137, 260, 389
Mobile Health (mHealth), 137, 384

- Mobile patient-reported outcomes (MPRO), 260
- Molecular biology
diagnostic methods, 161, 162
functional analysis data, 152–154
human variation, 154, 155
integration platforms, 156–160
mechanisms of disease, 160, 161
molecular data to support clinical research, 160
molecular epidemiological data, 162
omics data clinical application, 155, 156
sequence analysis data, 149, 151
structure analysis data, 152
therapeutic applications studies, 161, 162
- Multilayer artificial neural network, 346
- Multistage designs, 77
- N**
- Naive Bayes classifier, 363
- Named Entity Recognition (NER) tasks, 442
- National Academy of Medicine (NAM), 219
- National cancer informatics program (NCIP), 353
- National Cancer Institute (NCI), 113, 353
- National Cancer Institute Thesaurus (NCIT), 315, 317, 321
- National Center of Biomedical Ontology, 317
- National Council for Pharmacy Drug Program (NCPDP), 323, 411
- National Institute for Standards and Technology (NIST), 417
- National Institutes of Health (NIH), 214, 264
- National Library of Medicine (NLM), 91, 327, 334
Value Set Authority Center, 410
- National patient-centered clinical research network (PCORnet), 332
- National provider identifier (NPI), 277
- Natural language processing (NLP), 372
deep learning approach, 367, 368, 370
learning based approach, 369
machine learning approach (*see* Machine learning approach)
rule based approach, 368
sublanguage approach
definition, 361
semantics and discourse level, 362
syntax level, 362
vocabulary level, 362
- NCIT, *see* National Cancer Institute Thesaurus
- Nearest neighbor, 363
- NegScope, 369
- NIH Clinical and Translational Science Award (CTSA), 483
- NLM Value Set Authority Center, 410
- Non-informatics associations and journals, 486
- O**
- Observational Health Data Sciences and Informatics (OHDSI), 320, 397, 414
- Observational Medical Outcomes Pilot (OMOP), 440
CDM, 414
- Observational studies, 6, 22, 38, 54–56, 204, 270, 280, 342, 447, 467
- Occam's razor, 225
- OCRe, *see* Ontology of clinical research
- OneMind, 397
- Online analytic processing (OLAP), 344
- Ontologies
biomedical, 315
biomedical investigations, 315, 318, 319
BioPortal applications, 328, 329
BRIDG model, 319, 320
clinical module, 318
clinical research, 317
data integration, 330–332
electronic/computable phenotyping, 332–334
workflow management, 329, 330
- and computational methods, 441
- CPT, 324
- data contents, 320
- definition, 313
- development
building blocks, 315
DL, 315
foundry, 316, 317
harmonization effort, 317
OBO syntax, 316
OBO-Edit, 316
ontological distinctions, 314
OWL, 315
OWL syntax, 316
RDF/S, 316
SKOS, 316
web of data, 315
- ICD, 323, 324
- LOINC, 322, 323
- NCIT, 321
- post hoc mapping and alignment, 334
- research module, 318
- RxNorm, 323
- Semantic Network, 326
- SNOMED CT, 321, 322

- study design module, 318
traditional data warehouses, 325
UMLS, 325
UMLS and BioPortal, 334
UMLS applications, 327
UMLS Metathesaurus, 326
UTS, 326
- Ontology for biomedical investigations (OBI), 318
Ontology of clinical research (OCRe), 315, 318
Ontology Web Language (OWL), 201, 204, 315, 316
Ontology-based Trial Management Application (ObTiMA), 330
Open access Research Data Repositories in public domain, 476
Open biomedical ontologies (OBO), 315–317
 OBO Foundry, 316
 OBO-Edit, 316
Open science and data, 454
OpenPVSignal, 447
Operational Data Model (ODM), 391, 407
Orphanet, 317
- P**
- Paper-based information management practices, 41
Participant screening tools, 40
Patient advocacy organizations, 33–34
Patient Centered Outcomes Research Institute, 397
Patient engagement models, 135, 136
Patient registries
 biomarker discovery, 271
 biomedical and health services research, 286, 287
 CMS centralized repository of registries, 271
 comparative effectiveness research, 271
 definitions, 270
 EHR systems, 274
 envision registries, 271
 error and bias types, 281
 FORCE-TJR registry, 272
 inclusion criteria types, 270
 informatics approaches
 clinical system, 283
 cost and commitment, 282
 critical functions of, 285, 286
 efficiency calculus, 283
 federated form, 284
 FHIR API, 283
minimal mapping work, 283
NQRN clinical registry maturational framework model, 285
PCORnet, 283
structured reporting, 284
tidy data, 284
- interoperability and data standards
 CCRF model, 279, 280
 clinical models and data elements, 276, 277
 clinical phenotype, 278
 coding systems and controlled terminologies, 276
 content standards, 275
 data exchange standards, 275
 outcome measures, 279
 UDI, 277, 278
- LHS, 273
limitations of, 280, 281
NIH inventory, 271
patient safety, 271
post-market surveillance, 271
qualified registries, 271
risk mitigation and evaluation systems, 271
RoPR, 271
support clinical trials planning and recruitment, 271
- Patient-reported outcomes (PRO)
 characteristics of, 250–252
 collections of instruments, 263
 CONSORT reporting guidelines, 252
 definition, 250
 electronic mail, 258, 259
 item and scale development, 260–262
 IVR systems, 259
 mailed surveys, 257, 258
 measurement issues
 demographic characteristics, 252
 design issues, 253
 measurement properties, 252
 reliability, 253, 254
 responsiveness, 256
 validity, 254–256
 modification of existing, 263
 online repositories, 264
 personal/face-to-face administration, 256
 PROMIS item banks, 264, 265
 response options types, 262
 screen text devices, 259, 260
 telephone administration, 257
 web surveys, 258
- Patient-Reported outcome measurement information system (PROMIS), 264
item banks, 264, 265

- PCORnet common data model (CDM), 306, 414
- Pediatric research using integrated sensor monitoring systems (PRISMS), 354
- Personal/face-to-face administration, 256
- Personal identifying information (PII), 301
- Pharmaceuticals and Medical Device Agency (PMDA), 388
- Pharmacovigilance
- definition, 434
 - development of, 434
 - drug safety, 434
 - in clinical research, 433
 - regulatory and legal requirements, 434
- Physician Data Query (PDQ), 113
- Planning, quality improvement, 389–391
- Pluripotent storage model, 285
- Postgenomic era
- genomics data, 149–151
 - molecular biology, 149
- Postmarketing phase, AEs, 443
- Pragmatic clinical trials, 173, 218
- analysis, 188
 - comparison therapy choice flexibility, 188
 - follow-up intensity, 188
 - outcome, 188
 - patient selection criteria, 188
 - personnel, 188
 - Practitioner adherence, 188
 - subject compliance, 188
 - therapeutic flexibility, 188
- Precision medicine, 43, 446
- Precision Pharmacovigilance, 446
- ProcedureRequests, 413
- Process Analysis and Design, 392
- Prospective studies, 55
- Protected health information (PHI), 299, 371
- Protective factor, 54
- Protocol
- care and research, 208
 - computable study protocol
 - accurate data capture, 199, 200
 - complete study plan, 197
 - computability and standardization, 201
 - decision support, 198, 199
 - facilitating timely, 199, 200
 - interpretation and application of results, 200, 201
 - statistical analysis and reporting, 200
 - study data and artifacts, 201
- Core Protocol version 1, 104
- CPT, 395
- eligibility criteria representation standards, 205
- EMR data, 207, 208
- e-protocol, 193, 197
- executed study protocol, 192
- study design improvement, 206, 207
- Protocol authoring tools, 40
- Protocol representation standards, Health Level 7, 202
- Pseudonymization, 382
- Public health reporting, 408
- Public Health Services Act (PHSA), 89, 91
- Public protection and disaster relief (PPDR), 331
- Public Reporting Of Clinical Trials Outcomes and Results (PROCTOR) group, 470
- Publication bias, 457, 461
- PubMed, 39, 332, 427
- PV/drug safety, 435–437
- Q**
- Quality of life (QoL), 250
- Quasi-registries, 282
- Quality reporting registry, 270
- R**
- Random error, 53, 67
- Randomization, 55
- Re3data, 473
- Real world data (RWD), 44, 389, 394
- Real world evidence (RWE), 44–45, 383
- Receiver operating curve (ROC)
- analysis, 350
- Recurrent neural networks (RNNs), 368, 370
- REDCap, 172
- Reengineering, 6, 42, 393, 404
- Reference Information Model (RIM), 202, 411
- Registry of patient registries (RoPR), 271
- Regulated Clinical Research Information Management (RCRIM), 416
- Relation ontology, 315
- Reliability, 253–254
- Representational state transfer (RESTful)
- APIs, 275
 - web services, 327, 328
- Research Data Alliance (RDA), 473
- Research data repositories, 463, 472, 473
- Research design, biomedical studies. *see* Study design
- Research methods, 10, 137, 220, 442, 488
- Research transparency, 458, 464, 472
- Research-specific decision-support systems, 40

- Resource description framework schema (RDF/S), 316
- Resources For Health (RFH) project, 411
- Responsiveness, 256
- Retrieve Form Data Capture (RFD), 396
- Retrieve Protocol for Execution (RPE) standards, 396
- Retrospective studies, 55
- Risk factor, 54
- Rule based approach, 368
- RxNorm, 276, 277, 321, 323, 325, 333
- S**
- Safe Harbor method, 300
- Scientific Advisory Group (SAG), 465
- SDO Charter Organization (SCO), 417
- Secondary use of data, 382
- Semantic interoperability, 202, 325, 381
- Semantic web, 114, 316, 332, 353
- Shared Health and Research Electronic Library (SHARE), 397
- Sharing data, 380
- Simple knowledge organization system (SKOS), 316
- Simultaneous treatment design, 77
- Skip logic, 180
- SNOMED clinical terms (SNOMED CT), 276, 277, 315, 317, 321, 322, 325, 327, 333, 443
- Software-as-a-Medical Device (SaMD), 102
- Software-inside-a-Medical Device (SIMD), 102
- SPECIALIST Lexicon, 326, 327
- Split-half reliability, 254
- Sponsoring organizations, 36
- Stability, 253
- Stand-alone desktop systems, 260
- Standard operating procedures (SOPs), 238
- Standard Protocol Items for RandomIzed Trials (SPIRIT), 203, 458, 459
- Standards Coordinating Organization (SCO), 417
- STARBRITE project, 385
- Storage Standard for Medical Information Exchange (SS-MIX), 385
- Strategic Health IT advanced research projects (SHARP), 332, 360
- Structured Product Labeling, 416
- Study Data Tabulation Model (SDTM), 406
- Study design
- classification, 54
 - clinical research, 407
 - definitions, 53
- distinctive characteristics, 54
- epidemiological studies, 54
- experimental studies, 55, 56
- measurement-related variability, 53
- minimal intervention studies, 56–58
- phase I, 50
- phase II, 50
- phase III, 51
- phase IV, 51
- phenotypic variability, 53
- temporal variability, 53
- See also* Experimental designs
- Study protocol
- clinical research informatics, 193, 194
 - clinical research study, 192, 193
 - computable study protocol
 - accurate data capture, 199, 200
 - complete study plan, 197
 - computability and standardization, 201
 - decision support, 198, 199
 - facilitating timely, 199, 200
 - interpretation and application of results, 200, 201
 - statistical analysis and reporting, 200
 - study data and artifacts, 201
 - executed study protocol, 192
 - Sublanguage theory, 361, 362
- Substitutable Medical Applications and Reusable Technologies (SMART), 397
- Support vector machines (SVMs), 348, 363, 366
- learning algorithm, 369
- Surrogate end-points, 61
- Symbolic Text Processor (SymText), 369
- Systematic error, 53
- Systems biology approach, 442
- T**
- Telephone administration, 257
- Test-retest reliability, 254
- T-helper system, 113
- Tidy data, 284
- Touch-screen systems, 260
- Traceability, 215, 381, 388
- Traditional health communicators
- information environment, 127, 128
 - interpersonal communication in social networks, 128, 129
 - self-help and advocacy, 131, 132
 - third parties, 130, 131
 - WWW-based social media applications, 126

- Training/testing curves, 347
TransCelerate, 419
Transcription errors, 392
Transcriptome, 148, 149
Translational bioinformatics (TBI), 482, 483, 486
Translational Research and Patient Safety in Europe project (TransFoRm), 385, 386, 397
Translational science, 392, 482
Transparency and accountability, clinical trials, 455
Trial registration
 amendments, 462
 anticipated flow of data, 475
 characteristics and design features, 461–463
 creation and management
 data fields, 468, 469
 data quality, 469
 design of, 467
 first-level fields, 469
 international standards, 468
 maintenance, 470
 second-level fields, 469
 study design field, 469
 third-level fields, 470
 decision-making, 475
 development, 457
 disastrous health consequences, 466
 disease-related information, 461
 and IBD, 455
 international standards, 459, 460, 471
 international trial registration
 standards, 466
 learning tools, 461
 network of, 463
 primary registries, 462, 466, 475
 protocol-related data, 462, 463
public disclosure, 454, 471
quality of, 464–466
registration of selected protocol elements, 454
standards, 459
WHO IC RTP, 461
WHO international standards, 459
Tuskegee Syphilis Study, 92
21st Century Cures Act, 103, 277
- U**
- U.S. Centers for Medicare & Medicaid Services (CMS), 322
U.S. National Cancer Institute (NCI), 321
UMLS terminology services (UTS), 326, 327
Unified medical language system (UMLS), 199, 321, 325
Unified Modeling Language (UML), 201
Unique device identification (UDI), 277
Universal trial number (UTN), 466
Upper-level ontologies, 315, 318
- V**
- Valuable learning health systems, 380
Value Set Authority Center (VSAC), 278, 334
VitalHealth's QuestLink platform, 260
Voice over Internet protocols (VOIP), 259
- W**
- Web surveys, 258–259
Whole genome sequences (WGS), 22
Working Group on Best Practice for Clinical Trial Registries, 462
World Health Organization (WHO), 124, 323, 324, 425, 457–463, 468, 475, 476