Health Informatics

Rachel L. Richesson
James E. Andrews *Editors*

Clinical Research Informatics

Springer

Health Informatics

Founding Series Editors
Kathryn J. Hannah • Marion J. Ball

Rachel L. Richesson • James E. Andrews
Editors

Clinical Research Informatics

Springer

Editors
Rachel L. Richesson, Ph.D., MPH
Division of Informatics
Duke University School of Nursing
Durham
NC
USA

James E. Andrews, Ph.D.
Director and Associate Professor
College of Arts and Sciences
School of Information
University of South Florida
Tampa
FL
USA

ISBN 978-1-84882-447-8 e-ISBN 978-1-84882-448-5
DOI 10.1007/978-1-84882-448-5
Springer London Dordrecht Heidelberg New York

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Control Number: 2012931153

Product liability: The publisher can give no guarantee for information about drug dosage and application thereof contained in this book. In every individual case the respective user must check its accuracy by consulting other pharmaceutical literature.

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Contents

Part I Contexts of Clinical Research Informatics

Part II Data Management and Systems in Clinical Research

Part III Knowledge Representation and Discovery

Part IV The Future of Clinical Research, Health, and Clinical Research Informatics

Contributors

James E. Andrews, Ph.D. School of Information, College of Arts and Sciences, University of South Florida, Tampa, FL, USA

Antonella Bacchieri, M.S. Department of Biostatistics and Data Management, Research and Development, Sigma-tau Industrie Farmaceutiche Riunite S.p.A., Pomezia, Rome, Italy

Olivier Bodenreider, M.D., Ph.D. Lister Hill National Center for Biomedical Communications, US National Library of Medicine, National Institutes of Health, Bethesda, MD, USA

Cynthia A. Brandt, M.D., MPH Yale Center for Medical Informatics, Emergency Medicine, Yale University, New Haven, CT, USA

Department of Anesthesiology, VA Connecticut Healthcare System, New Haven, CT, USA

Christopher G. Chute, M.D., DrPH Division of Biomedical Statistics and Informatics, Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA

Mollie R. Cummins, Ph.D., APRN College of Nursing, University of Utah, Salt Lake City, UT, USA

Giovanni Della Cioppa, M.D. Global Clinical Research and Development, Vaccines, Novartis Vaccines and Diagnostics, Siena, Italy

Peter J. Embi, M.D., M.S. Department of Biomedical Informatics and Internal Medicine, The Ohio State University Medical Center, Columbus, OH, USA

Kin Wah Fung, M.D., M.S., M.A. Lister Hill National Center for Biomedical Communications, US National Library of Medicine, National Institutes of Health, Bethesda, MD, USA

W. Ed Hammond, Ph.D. Duke Center for Health Informatics, Duke Translation Medicine Institute, Durham, NC, USA

Applied Informatics Research, Duke Health Technology Solutions

Biomedical Informatics Core, Duke Translational Medicine Institute

Department of Community and Family Medicine, School of Medicine

Department of Biomedical Engineering, Pratt School of Engineering

Fuqua School of Business, Duke University, Durham, NC, USA

Elizabeth J. Horn, Ph.D., MBI Genetic Alliance Registry & BioBank, Genetic Alliance, Washington, DC, USA

J. David Johnson, Ph.D., M.A., B.S. Department of Communication, University of Kentucky, Lexington, KY, USA

Karmela Krleža-Jeri , M.D., M.Sc., D.Sc. Knowledge Translation, Canadian Institutes of Health Research, Ottawa, ON, Canada

Department of Epidemiology and Community Health, Medical Faculty, University of Ottawa, Ottawa, ON, Canada

Rebecca Daniels Kush, Ph.D., B.S. Executive Operations, CDISC, Austin, TX, USA

Feifan Liu, Ph.D. Department of Electric Engineering and Computer Science, University of Wisconsin Milwaukee, Milwaukee, WI, USA

Luis N. Marenco, M.D. Yale Center for Medical Informatics, Yale University, New Haven, CT, USA

Stephane M. Meystre, M.D., Ph.D. Department of Biomedical Informatics, University of Utah, Salt Lake City, UT, USA

Joyce A. Mitchell, Ph.D. Department of Biomedical Informatics, University of Utah, Salt Lake City, UT, USA

Robert O. Morgan, Ph.D. Department of Management, Policy and Community Health, University of Texas School of Public Health, Houston, TX, USA

Prakash M. Nadkarni, M.D. Center for Medical Informatics, Yale University School of Medicine, New Haven, CT, USA

Meredith Nahm, Ph.D. Informatics, Duke Translational Medicine Institute, Duke University, Durham, NC, USA

Scott P. Narus, Ph.D. Department of Biomedical Informatics, University of Utah, Salt Lake City, UT, USA

Joyce C. Niland, Ph.D. Division of Information Sciences, Beckman Research Institute, City of Hope, Duarte, CA, USA

Philip R.O. Payne, Ph.D. Department of Biomedical Informatics, The Ohio State University, Columbus, OH, USA

Rachel L. Richesson, Ph.D., MPH Department of Informatics, Duke University, School of Nursing, Durham, NC, USA

Kavita R. Sail, Ph.D. Department of Health Care Informatics, McKesson/US Oncology Specialty Care Unit, The Woodlands, TX, USA

Ida Sim, M.D., Ph.D. Department of Medicine, University of California San Francisco, San Francisco, CA, USA

Sharon F. Terry, M.A. Genetic Alliance, Registry and Biobank, Genetic Alliance, Washington, DC, USA

A.C. (Kees) van Grootheest, M.D., Ph.D. Mathematics and Natural Sciences, Netherlands Pharmacovigilance Centre, s-Hertogenbosch, the Netherlands

Department of Pharmacy, University of Groningen, s-Hertogenbosch, the Netherlands

Kendra Vehik, Ph.D., MPH Department of Pediatrics, Pediatrics Epidemiology Center, University of South Florida, Tampa, FL, USA

Chunhua Weng, Ph.D. Department of Biomedical Informatics, Columbia University, New York, NY, USA

Hong Yu, Ph.D. Department of Electric Engineering and Computer Science, University of Wisconsin Milwaukee, Milwaukee, WI, USA

Part I
Contexts of Clinical Research Informatics

Chapter 1
Introduction to Clinical Research Informatics

Rachel L. Richesson and James E. Andrews

Abstract This chapter provides essential definitions and overviews important constructs and methods within the emerging 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

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. 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 disease. Clinical research enables new understanding and practices for prevention, diagnosis, or treatment of a disease or its symptoms. 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

R.L. Richesson, Ph.D., MPH (✉)
Division of Informatics, Duke University School of Nursing,
311 Trent Drive, Durham, NC 27710, USA
e-mail: rachel.richesson@dm.duke.edu

J.E. Andrews, Ph.D.
School of Information, College of Arts and Sciences, University of South Florida,
4202 E. Fowler Ave, CIS 1040, Tampa, FL 33620-8100, USA
e-mail: jimandrews@usf.edu

R.L. Richesson, J.E. Andrews (eds.), *Clinical Research Informatics*, Health Informatics,
DOI 10.1007/978-1-84882-448-5_1, © Springer-Verlag London Limited 2012

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), then, is the application of informatics principles and techniques to support the spectrum of activities and business processes that instantiate clinical research. Informatics, as somewhat crudely 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 newly articulated yet 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 offer 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. As a relatively new field, but driven by maturing professional and research communities, 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. 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 [7]. The use of clinical databases for research inquiry was first established in the late 1960s, and by the next decade – more than 40 years ago – 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 [8]. 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 [9]. 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* [10–12]. 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 [13], 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, increasing ability to collect data rapidly from patients or external data sources, and the scope and scale of today's research enterprises have lead 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. Existing informatics tools and data management systems have been adopted and tailored to address the unique issues in clinical research. Because the objectives and workflows of clinical research are unique, or at least highly specific, new tools and solutions have also emerged in the form of CRI. 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 outcomes 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. It is not surprising, however, that the opportunities and needs in CRI are so great and have emerged so rapidly that practitioners and researchers have not had time to amalgamate ideas or define themselves as a professional and scientific community. This collection of works is meant to provide a beginning toward helping 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 during its early evolution. 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 [14]. Rather, this collection is targeted for those with at least a basic understanding of the field and who would like to apply informatics principles to clinical research problems and processes. Many of these theories and principles presented in this collection 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. What we have found, however, is that there are clear information problems at the core of clinical research that have dominated CRI. Moreover, from these efforts discernible trends are emerging, and research/practice foci 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 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). 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 in 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, laboratory data, etc.) or topics. However, to restate, it was impractical for this book to attempt to cover every aspect of the field. The most notable omission is a single chapter on regulatory science, but the field is relatively new as an informatics-related focus, and moving rapidly to accommodate explosive changes in research data sources, genetic data, new technologies, evolving patient roles, and new models and ethical issues for research in international settings, especially in developing nations. The topic is, however, touched upon in other chapters, and we hope readers will see the relevance and importance of regulations and ethics to CRI and also appreciate the gradual emergence of regulatory science as a scientific practice area in its own right.

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. Still, 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). It has only been relatively recently that the AMIA CRI Working Group has become a promising research and policy leadership group. Regardless, as noted at a recent CRI-focused professional conference [15], there is a lack of original research in the form of classic randomized interventional research of informatics applications in the clinical research domain. This issue becomes manifest in the book through the chapters and their organization. Some chapters tend to focus on best practices and are instructional in nature, and some are theoretical (usually drawing from the parent or contributing discipline); some are very concrete and easy to define and digest; others are more abstract in nature and therefore require readers to extrapolate as to direct relevance to their own areas of interest.

Organization of the Book

As an attempt to cluster chapters under unifying themes, we chose to organize them at a high level using four broad sections: (1) the context and foundations of clinical research informatics; (2) data management and systems in clinical research; (3) knowledge representation and discovery; and (4) the future of clinical research, health, and clinical research informatics.

Section 1: Contexts 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 an historical perspective from Chris 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 written 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.

After this historical perspective and fundamentals of clinical research design and conduct, this introduction 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, David Johnson and Jim Andrews 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 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 through advocacy groups or other social networks. 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. 7 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, Scott Narus, and Joyce Mitchell 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 Management and Systems in Clinical Research

Six 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, Luis Marenco, and Cynthia Brandt. Their chapter also gives broad overview of system selection and evaluation issues. Their 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 Ida Sim and Joyce Niland 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. Sim and Niland 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 Nahm addresses the idea that central to clinical research is data collection, quality, and management. She focuses on various types of data collected (e.g., clinical observations, diagnoses) and the methods and tools for collecting these. Special attention is given to Case Report Forms (CRFs), the primary mechanism for data collection in clinical research, including discussions regarding the nature, development, and organization of the questions that comprise CRFs. The chapter provides both a theoretical framework for data quality in clinical research and also will serve as practical guidance. Moreover, Nahm draws on the themes of workflows presented by Payne in Chap. 2, and advocates 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 and Kavita Sail 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. In addition, Liz Horn and Sharon Terry discuss the informatics issues involved in cell line/tissue banking, as well as an overview of the complexities and restrictions dealing with storage and repeated analysis of human tissues. Regarding the latter, the authors characterize the broad types of research questions (current and near future) that biobanks support and highlight issues of quality and standardization, particularly regarding the information (annotations) related to those samples. Technologies and languages for indexing samples and merging different data sets are key issues. This chapter mentions virtually all informatics issues that affect the collection, standardization, and interpretation of data for research purposes, as well as highlights international challenges related to the comparability of genetic and pathology data for multisite research efforts. These data are particularly relevant and synergize the ideas regarding consumer involvement that were presented by Jim Andrews and David Johnson in the first section. As patients become more active in research – as participants, sponsors, and consumers – the patient-reported data and patient biological specimens represent domains that are dominated as patient concerns.

Finally, and also related to patients, is Rachel Richesson and Kendra Vehik's discussion of the use of patient data registries for observational research. Their discussion includes the variety of registries, their proliferation and overlap, and standards and best practice issues. Technical issues such as access, data updates, and data quality are fully described, as are sociopolitical issues related to data ownership, policy, and international regulations. The registry issues represent rapidly growing area of research activity that is being driven in large part by patient advocacy organizations, but also includes registry sponsors and developers from pharmaceutical industries, commercial providers, government agencies, and academic medical centers.

Section 3: Knowledge Representation and 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 that is gaining 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.

Section 4: The Future

In this final section of the text, we chose to include a representation of different 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.

Data sharing is tremendous challenge with perhaps the greatest potential for impact in all areas of clinical research. In the U.S., a domestic national health information infrastructure is being defined by the "collect once, use many" paradigm, which has broad support but a lack of consensus for how (or if) it can be accomplished on a massive scale. Rebecca Kush covers various scenarios for data sharing, including who needs to share data and why. More importantly, she describes the history and future strategy of cooperation between major standards development organizations in health care and clinical research. In her leadership role in the Clinical Data Interchange Standards Consortium (CDISC), she provides an illuminating perspective of the momentous but arduous alliance of the health care and research professional community silos. Further, she defines interoperability, along with examples that demonstrate the need for interoperability in satisfying regulations and in supporting global science. She also describes current technical, regulatory, and logistical challenges related to the reuse of electronic health record for clinical research.

Ed Hammond and Rachel Richesson cover the topic of standards – a central topic and persistent challenge for informatics efforts. Their focus is on the standards development process, identification of relevant standards, and selection and implementation issues. Like Dr. Kush's previous chapter, they address the collaboration and harmonization between clinical care data standards (and professional communities) and research data standards. Specifically, an argument for research standards (in the form of common data elements) that are complementary to health-care standards is introduced.

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. Kees van Grootheest and Rachel Richesson provide an 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 safety monitoring, new methods for reporting of 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-Jeri 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-care system and of the relationship between clinical research, evidence-based 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. Milano: 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. Available from http://www.oenb.at/de/img/executive_summary--nih_directors_panel_on_clinical_research_report_12_97_tcm14–48582.pdf. Cited 15 May 2011.
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. Forrest WH, Bellville JW. The Use of computers in clinical trials. Br J Anaesth. 1967;39:311.
8. Pryor DB, Stead WW, Hammond WE, et al. Features of TMR for a successful clinical and research database. Proc Annu Symp Comput Appl Med Care (SCAMC). 1982;Nov 2:79–84.
9. Blois MS. Medical records and clinical databases: what is the difference? MD Comput. 1984;1(3):24–8.
10. Carlson RW, Tu SW, Lane NM, et al. Computer-based screening of patients with HIV/AIDS for clinical-trial eligibility. Online J Curr Clin Trials. 1995;Mar 28:Doc No 179. Citation is located at: http://www.ncbi.nlm.nih.gov/pubmed/7719564
11. Mansour EG. Barriers to clinical trials. Part III: knowledge and attitudes of health care providers. Cancer. 1994;74(9 Suppl):2672–5.
12. Breitfeld PP, Weisburd M, Overhage JM, 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.
13. International Conference on Harmonsiation. Guideline for Good Clinical Practice E6(R1), current step 4 version, 10 June 1996. Source located:http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E6_R1/Step4/E6_R1__Guideline.pdf
14. Shortliffe EH, CImino J, editors. Biomedical informatics: computer applications in health care and biomedicine. Health Informatics. Hannah KJ, editor. New York: Springer Science + Business Media, LLC; 2006.
15. Embi PJ. AMIA 2011 summit in CRI. CRI year-in-review. In: AMIA summit in Clinical Research Informatics, San Francisco; 2011.

Chapter 2
From Notations to Data: The Digital Transformation of Clinical Research

Christopher G. Chute

Abstract 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 computer 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. Similarly, in the past decade, informatics, including clinical research informatics as a recognized subfield, has come into its own.

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

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

C.G. Chute, M.D., DrPH
Division of Biomedical Statistics and Informatics,
Department of Health Sciences Research,
Mayo Clinic, 200 First Street, SW, Rochester, MN 55905, USA
e-mail: chute@mayo.edu

R.L. Richesson, J.E. Andrews (eds.), *Clinical Research Informatics*, Health Informatics,
DOI 10.1007/978-1-84882-448-5_2, © Springer-Verlag London Limited 2012

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 computer 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 decade, 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 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, and a host of continuous data properties. 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 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 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 to Analog Processors

The advent of digital signal processing, first manifest in analog to digital converters, has fundamentally transformed clinical research. In effect, it is the marrying of quantitative data to computing capability. Digital to analog converters (DACs) 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 generation of digital "snapshots."

DACs were first practical during the Second World War, when they were experimented with to carry telephonic signals over long distances without degradation. The telephony industry brought this capability into the civilian world, and commercial DACs 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 DACs. While case reports exist of physiologic researchers using DACs in the 1950s, this did not become common practice until the cost and performance characteristics of this technology became practical in the early 1970s.

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 DACs are 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 keypunch machinery.

While it is obviously true that not all biomedical data or clinical informatics arose from transducer-driven DAC 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 has 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 trillion-fold increase in computing power over the last 50 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.

Network Capacity

Early computing devices were reliant on locally connected devices for input and output. The most primitive interface devices were plugboard or toggle switches that required human configuration; the baud rates of such devices are perhaps unimaginably slow. Today, 100-Gb network backbones are not uncommon, giving yet 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^{14} increase in data storage capacity.

Taken together, these advances total an approximate 10^{50} increase in computational power (albeit we are cheating somewhat adding exponents, which is really multiplying in nonlogarithmic space) over the past 50 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.

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, 10-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 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 multinodal neural networks, 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. Absent 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 big science given its requirements for large cyclotrons, 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 nonscalable. 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.

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, manifest 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.

Chapter 3
The Clinical Research Environment

Philip R.O. Payne

Abstract Clinical research is an information and resource intensive endeavor, incorporating a broad variety of stakeholders spanning a spectrum from patients to providers to policymakers. Increasingly, the modern clinical research environment incorporates a number of informatics methods and technologies, informed by socio-technical and information-theoretic frameworks. In this chapter, we introduce the major facets that serve to define the clinical research setting, including the design of clinical studies, clinical research workflow, and information management needs incumbent to 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.

Keywords Clinical research workflow • Clinical research funding • Clinical research information management needs • Design of informatics platforms • Common clinical research settings • Large-scale research consortia • Information management requirements

In this chapter, we describe the clinical research environment, including an overview of common activities and processes, as well as the roles played by various actors involved throughout the lifecycle of clinical studies, including interventional and observational study designs. This discussion summarizes 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.

P.R.O. Payne, Ph.D
Department of Biomedical Informatics, The Ohio State University,
3190 Graves Hall, 333 West 10th Avenue, Columbus, OH 43210, USA
e-mail: philip.payne@osumc.edu

R.L. Richesson, J.E. Andrews (eds.), *Clinical Research Informatics*, Health Informatics,
DOI 10.1007/978-1-84882-448-5_3, © Springer-Verlag London Limited 2012

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 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 in the clinical research environment.

Clinical Research Processes, Actors, and Goals

In the following section, we introduce broadly applicable processes, actors, and goals that serve to characterize the modern physical and sociotechnical clinical research environment. Taken as a whole, these components represent a complex, information-intensive enterprise that incorporates a broad variety of professionals and participants and what are nominally concurrent and tightly interrelated goals or objectives. Given such a challenging environment, the role of informatics in addressing potential barriers to the efficient, effective, high-quality, and timely conduct of clinical research programs is an area of intensive and ongoing interest in the biomedical informatics community [1, 2].

Common Clinical Research Processes

At a high level, the processes that comprise clinical research can be divided into eight general classes, as summarized below. Of note, we will place particular emphasis in this section on describing such processes relative to the conduct of interventional clinical studies. 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. An example of such workflow components, relative to the context of an interventional clinical trial, is illustrated in Fig. 3.1. Key processes include the following.

Identifying Potential Study Participants

This process usually involves either (1) the preencounter and/or point-of-care review of an individual's personal characteristics and medical history 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, or (2) the identification of a cohort of potential study participants from whom data can be

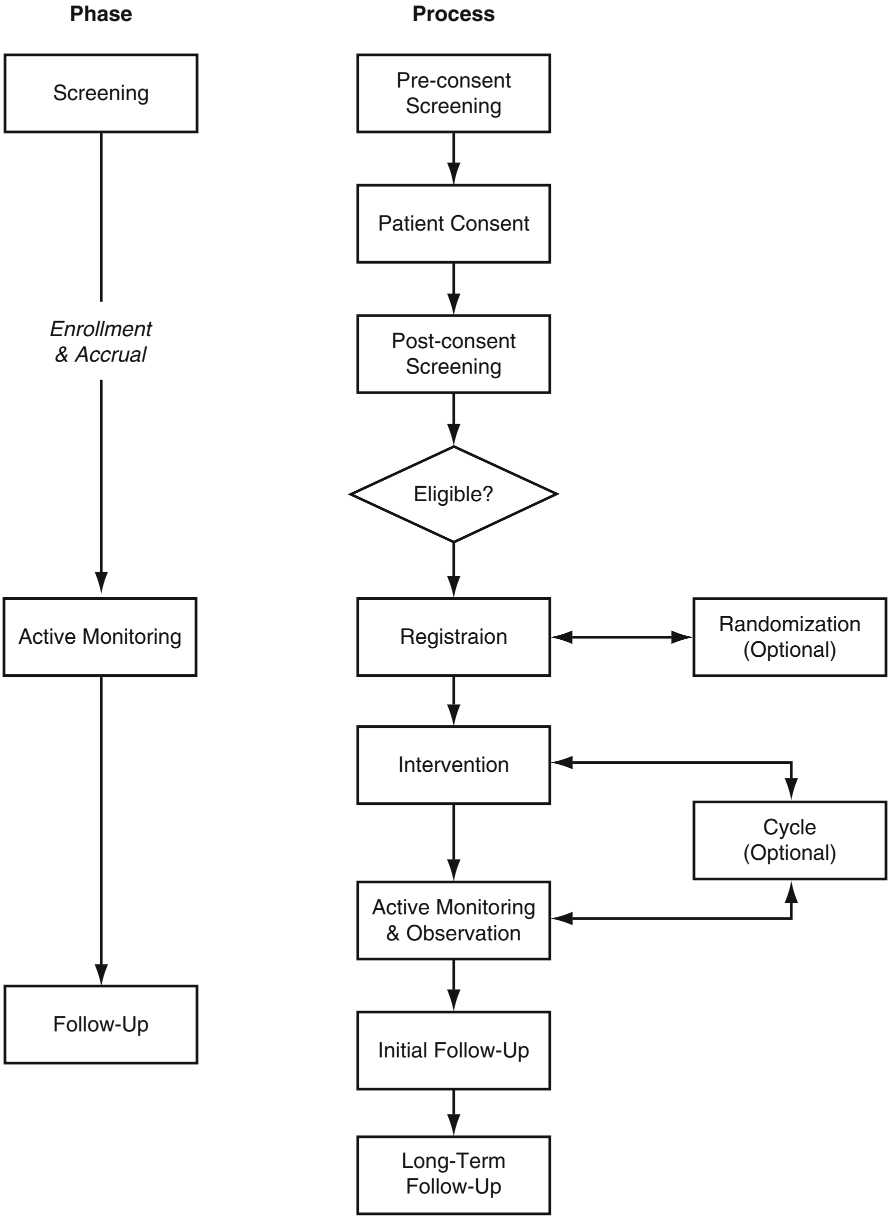


Fig. 3.1 Interventional clinical trial phases and associated execution-oriented processes

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 workflows. In many cases, potential participants are identified partially and then referred for further screening via interview or other similar mechanisms. Due to prevailing confidentiality and privacy laws and regulations, if the individual performing such eligibility screening is not directly involved in the medical 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 medical 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 additional interviews and/or testing 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.) 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.

Scheduling and Tracking Study-Related Participant Events

Once a participant has been identified, screened, and enrolled in a study, they are usually scheduled for a series of encounters as defined by the study protocol calendar. 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 schedule or calendar. 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.

Ensuring the Quality of Study Data

Throughout a given study, research 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., medical 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 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 subjects related reporting and the monitoring of such compliance is 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.

According to a recent study conducted by Khan and colleagues, the five most common tasks performed by research staff during clinical studies are: (1) completing case report forms, (2) seeking study information, (3) completing EDC instruments, (4) seeking general information (e.g., medical or other supporting information related to a study protocol), and (5) identifying potential clinical trial participants. In the same study, it was determined that the preceding tasks are most commonly performed using the following five types of tools or approaches: (1) paper-based forms and information sources, (2) verbal communications, (3) computer-based information systems, (4) manual processes (e.g., reviewing or organizing information sources), and (5) telephones [3].

Common Clinical Research Actors

The clinical research environment can include a broad variety of actors fulfilling multiple roles. Such actors 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 [4–8].

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 only 3% 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 [9–12].

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 [13, 14].

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 [15]. During the conduct of clinical studies, AHCs or equivalent entities may take on any number or combination of the following responsibilities:

- Obtaining local regulatory 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 studies 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 [7].

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 a studies 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/or 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 [4, 6, 7, 16].

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/associates/assistants, 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 [4, 6, 7, 16].

Clinical Research Organizations

Clinical 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 the 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 [17].

Sponsoring Organization

Sponsoring organizations are primarily responsible for the origination (except in the case of investigator-initiated clinical trials, as discussed earlier) and funding of clinical research programs. 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 lifecycle:

- 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, 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 a 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 [4, 6].

Federal Regulatory Agencies

Federal regulators are primarily responsible for overseeing the safety and legality 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 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 lifecycle, 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). Given the ever-increasing adoption of information technology (IT) 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 [1, 18], 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.

Other Clinical Research Actors

Additional actors who play roles in the clinical research setting include the following [19–21]:

- 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 Sect. 3, 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 share 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) 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 [3, 4, 22–25].

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 of 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

[a] The gold standard for such methodological approaches is the randomized controlled trial (*RCT*)

A Framework for Information Management Requirements in Clinical Research

In order to better understand the relationships between the information needs of clinical researchers and available informatics tools or platforms, it is helpful to conceptualize the conduct of clinical research programs as a multiple-stage sequential model [1]. At each stage in this model, a combination of general-purpose, clinical, and research-specific IT 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 [26–30].
- Electronic medical records (EMRs or alternatively, Electronic Health Records or EHRs) can be utilized to collect clinical data on research participants in a structured form that can reduce redundant data entry [31–36].
- 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 [37–39]. (Also see Chap. 15)
- Decision-support systems can be used to alert providers at the point of care that an individual may be eligible for a clinical trial [39–41].
- 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 [15, 33, 42].

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 [43, 44].
- Protocol authoring tools can allow geographically distributed authors to collaborate on complex protocol documents [45–49].
- Participant screening tools can assist in the identification and registration of research participants [39, 41, 50].
- Research-specific web portals provide researchers with a single point of access to research-specific documents and information [51–53].
- 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 [33, 54–56].
- Research-specific decision-support systems provide protocol-specific guidelines and alerts to researchers, for example, tracking the status of participants to ensure protocol compliance [33, 49].

Fundamentally, the ability to use IT in support of clinical research relies on the ability to collect, store, and analyze data in a computationally tractable format. Electronic Data Capture (EDC) is a broad label for tools that enable the capture of protocol-specific data elements in a structured manner, and is a vigorous and active area of commercial and open source software development. In a report published by Forrester Research as early as 2005, it was projected that the number of global clinical trials utilizing EDC between 2001 and 2006 would see a 12-fold increase over a 6-year period [57]. A more recent report published by CenterWatch in 2007 demonstrated that 99% of surveyed sites ($n = 103$) were using EDC technologies for at least one of their active clinical studies, with 73%, 36%, and 2% of sites using EDC for at least 25%, 50%, or 100% of their active clinical studies, respectively [58–61].

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, including the four broad categories of such issues as summarized below [3, 62, 63].

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, as well 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, serves 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 [64–67]. 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) [68–70]. 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.

Trends in Clinical Research Funding

In the preceding sections of this chapter, we have outlined the basic theories and methods that facilitate the design and conduct of clinical research programs, as well as the significant actors and workflow characteristics that define the domain and current state of clinical research practice. Throughout these discussions, we have referenced emergent issues surrounding the funding and resources available to support such activities. In this section, we will explore this topic in more detail, touching on the emergence of notable large-scale research programs and consortia in the United States, as well the increasing situation of clinical research programs in community or international settings, instead of US-based academic health centers.

Large-Scale Research Consortia

We will use both the NCI-sponsored Cancer Biomedical Informatics Grid (caBIG) and NCRR-sponsored Clinical and Translational Science Award (CTSA) programs as exemplary cases of the evolution of the clinical research policy and funding, with an emphasis on their import relative to the conduct of biomedical informatics research and development:

- The caBIG program was launched in 2004 with the goal of developing an infrastructure capable of enabling multisite research and data sharing spanning National Cancer Institute's (NCI) funded centers in order to facilitate large-scale efforts intended to facilitate distributed clinical research and to inform personalized healthcare delivery [71]. Funding for the program had been reduced, but active development and support continues for platforms such as caGrid and caTissue Suite. The caBIG program developed a suite of data sharing and analysis platforms capable of supporting such goals. Further, as has been noted by Dr. Kenneth Buetow, discussing the interplay between caBIG and the data management requirements inherent to the translational sciences: "The tools, standards, and infrastructure developed for the caBIG program can provide a comprehensive solution to many of these data management issues" [71]. A number of characteristics of the caBIG program should be noted, including the (1) pursuit of an architectural model developed and overseen by a project-specific cadre of informaticians, (2) an explicit focus on modeling and representing current and desirable research workflows, and (3) an explicit attempt to make the model and subsequent tools applicable to disease research outside of oncology.
- The CTSA program, which began in 2006, seeks to establish a network of academic health centers (AHC), each with a scholarly home for the clinical and translational sciences. As has been stated by former National Institutes of Health (NIH) Director, Dr. Elias A. Zerhouni: "The development of this consortium represents the first systematic change in our approach to clinical research in 50 years. Working together, these sites will serve as discovery engines that will improve medical care by applying new scientific advances to real-world practice" [72]. A critical component of the CTSA program is the creation of informatics infrastructure and services for use by the research community at recipient organizations. Such efforts at most CTSA sites have focused on areas such as: training, consultative services, database design/hosting and data warehousing, data sharing infrastructure, and the execution of complex data analyses. The CTSA program was explicitly designed to integrate different parts of the research

process (from basic science to study conceptualization and design, to clinical research, to application of findings). There are also explicit provisions for education of clinical researchers, enhanced communication and collaboration across all actors in the research process, and the specification of a new breed of scientists called *translational scientists*.

Both the caBIG and the CTSA programs can be broadly categorized as affecting the availability of resources to support clinical research and clinical research informatics as follows:

- Resources and funding availability are focused on a small number of centers of excellence and associated with the need to satisfy high-level requirements defined by funding programs and policies put in place by governmental agencies and legislators.
- An emphasis is being placed on the design and use of modular components capable of enabling networks or networks, intended to provide for increased speed and economies of scale in the clinical research setting.
- Informaticians are primarily focused on providing information management services, with formative clinical research informatics research and development occurring as a secondary or indirect objective of such requirements and funding mechanisms.

Research in Community Practice and International/Global Settings

Beyond the preceding movement toward large-scale, government funding CRI research, and development efforts, there is an orthogonal movement toward the conduct of clinical research in settings beyond the traditional AHC, including community practice and international environments. Such a movement has significant implications for the CRI domain, as it represents a shifting target for the living laboratory in which such efforts are situated, as summarized below:

- Community Practice Settings: Increasingly, both private- and public-sector study sponsors are engaging community-based practice sites to conduct clinical studies. Such engagement is primarily motivated by the high financial and administrative overhead imposed by AHCs as well as the need to access increasingly broad audiences of potential participants for large-scale or otherwise difficult to conduct clinical studies. A primary vehicle of providing such community engagement is the Practice-Based Research Network (PBRN) or equivalent organizations, in which a centralized administrative body facilitates overall study operations at a distributed network of small- to medium-scale community practices. An example of this is a variety of Oncology Networks (e.g., CCOP) funded by the National Cancer Institute. With this shift in setting, the engagement of community practices in the design and evaluation of CRI platforms is becoming increasing critical. It is likely that this trend will continue for the foreseeable future [22–25, 73]. Advances in information technology in past few decades that now facilitate distributed communication and collaboration (e.g., teleconference, e-mail of documents and images, facsimile transmission of data) have created an environment that can better support distributed clinical research.

- International Settings: In addition to the movement of clinical research programs into community practice settings, there is a simultaneous effort underway to conduct many large-scale or early-phase clinical studies in international settings. One of the most common themes in this regard is the movement of clinical trials to what are known as the BRIC countries, namely, Brazil, India, and China. Recent reports have demonstrated increased participant recruitment and study completion rates, as well as lower costs and comparable data quality to that seen in US AHC-based trials from studies conducted in BRIC settings. This trend has the potential to significantly shift the focus of cutting-edge clinical research away from the USA, with its complex legal, regulatory, and funding environments. Such a shift would require a significant change in focus for the CRI community within the USA, Canada, the European Union, and Japan, none of which have traditionally conducted research or development activities in BRIC countries. Furthermore, the movement of clinical research originating in the USA to international settings introduced a broad variety of legal and ethical challenges, largely revolving around the harmonization of local, regional, national, and international norms and standards for research practices. These challenges represent yet another dimension of the challenges facing CRI in light of this shifting direction in clinical research conduct [74–77].

Discussion

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, actors, settings, 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, actors, 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 prevailing trends in funding policy and clinical study conduct. 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 CRI 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. Payne PR, Johnson SB, Starren JB, et al. Breaking the translational barriers: the value of integrating biomedical informatics and translational research. J Invest Med. 2005;53:192–200.
2. Embi PJ, Payne PR. Clinical research informatics: challenges, opportunities and definition for an emerging domain. J Am Med Inform Assoc. 2009;16:316–27.
3. Khan SA, Kukafka R, Bigger JT, Johnson SB. Re-engineering opportunities in clinical research using workflow analysis in community practice settings. AMIA Annu Symp Proc. 2008;2008: 363–7.
4. Finlayson SR. The effect of the current regulatory environment on clinical research: an overview. Surgery. 2009;145:626–30.
5. Holzhauser K, Cooke M, Winch S, Davis C. Developing a research-active clinical environment within the emergency department: a case study. J Prof Nurs. 2008;24:36–41.
6. Gennery B. Academic clinical research in the new regulatory environment. Clin Med. 2005;5: 39–41.
7. Barker KL. Building and supporting effective academic clinical research programs in today's market-driven health care environment: is research valued and how much and by whom? Exp Biol Med. 2003;228:1269–71.
8. Kovach JS. Clinical cancer research in a managed-care environment. Cancer Invest. 1998;16:350–2.
9. Madsen SM, Holm S, Riis P. Attitudes towards clinical research among cancer trial participants and non-participants: an interview study using a Grounded Theory approach. J Med Ethics. 2007;33:234–40.
10. Laurence DR. Participants in research: patients in clinical trials are protected by consumer regulation. BMJ. 2005;331:110.
11. Madsen SM, Mizra MR, Holm S, et al. Attitudes towards clinical research amongst participants and nonparticipants. J Intern Med. 2002;251:156–68.
12. Phipps E, Fleetwood J, Piraino A. Research ethics and clinical trials: fostering collaboration with research participants. Account Res. 1999;6:215–26.
13. Ayme S, Kole A, Groft S. Empowerment of patients: lessons from the rare disease community. Lancet. 2008;371:2048–61.
14. Watson MS, Epstein C, Howell RR, et al. Developing a national collaborative study system for rare genetic diseases. Genet Med. 2008;10:325–9.
15. Sung NS, Crowley WF, Genel M, et al. Central challenges facing the national clinical research enterprise. JAMA. 2003;289:1278–87.
16. Cox K. Setting the context for research: exploring the philosophy and environment of a cancer clinical trials unit. J Adv Nurs. 2000;32:1058–65.
17. Spilker B. Guide to clinical trials, vol. xxv. New York: Raven; 1991. p. 1156.
18. Payne PR, Embi PJ, Sen CK. Translational informatics: enabling high throughput research paradigms. Physiol Genomics. 2009;39:131–40.
19. Williams BA, Bottegal MT, Francis KA, et al. A post hoc analysis of research study staffing: budgetary effects of the Health Insurance Portability and Accountability Act on research staff work hours during a prospective, randomized clinical trial. Anesthesiology. 2007;107: 860–1.
20. Rutledge DN, Ropka M, Greene PE, et al. Barriers to research utilization for oncology staff nurses and nurse managers/clinical nurse specialists. Oncol Nurs Forum. 1998;25:497–506.

21. Medoff-Cooper B, Lamb AH. The clinical specialist-staff nurse research team: a model for clinical research. Clin Nurse Spec. 1989;3:16–9.
22. Goode JV, Mott DA, Chater R. Collaborations to facilitate success of community pharmacy practice-based research networks. J Am Pharm Assoc. 2008;48:153–62.
23. Pruchnicki MC, Rodis JL, Beatty SJ, et al. Practice-based research network as a research training model for community/ambulatory pharmacy residents. J Am Pharm Assoc. 2008;48: 191–202.
24. Westfall JM. Practice-based research is community engagement. J Am Board Fam Med. 2009; 22:423–7.
25. Williams RL, Shelley BM, Sussman AL. The marriage of community-based participatory research and practice-based research networks: can it work? A research involving outpatient settings network (RIOS Net) study. J Am Board Fam Med. 2009;22:428–35.
26. Briggs B. Clinical trials getting a hand. Health Data Manag. 2002;10:56–60, 62.
27. Ebbert JO, Dupras DM, Erwin PJ. Searching the medical literature using PubMed: a tutorial. Mayo Clin Proc. 2003;78:87–91.
28. Eveillard P. Bibliographic databases. Medline via PubMed. Rev Prat. 2000;50(16 Suppl):1–34.
29. Eysenbach G, Tuische J, Diepgen TL. Evaluation of the usefulness of internet searches to identify unpublished clinical trials for systematic reviews. Med Inform Internet Med. 2001;26: 203–18.
30. Eysenbach G, Wyatt J. Using the internet for surveys and health research. J Med Internet Res. 2002;4:E13.
31. Bates DW, Ebell M, Gotlieb E, et al. A proposal for electronic medical records in U.S. primary care. J Am Med Inform Assoc. 2003;10:1–10.
32. Clark O, Clark L, Djulbegovic B. Is clinical research still too haphazard? Lancet. 2001;358: 1648.
33. Marks RG, Conlon M, Ruberg SJ. Paradigm shifts in clinical trials enabled by information technology. Stat Med. 2001;20:2683–96.
34. McDonald CJ. The barriers to electronic medical record systems and how to overcome them. J Am Med Inform Assoc. 1997;4:213–21.
35. McDonald CJ, Overhage JM, Tierney WM, et al. The Regenstrief Medical Record System: a quarter century experience. Int J Med Inform. 1999;54:225–53.
36. Padkin A, Rowan K, Black N. Using high quality clinical databases to complement the results of randomised controlled trials: the case of recombinant human activated protein C. BMJ. 2001;323:923–6.
37. Butler D. Data, data, everywhere. Nature. 2001;414:840–1.
38. Evans D. Database searches for qualitative research. J Med Libr Assoc. 2002;90:290–3.
39. Marks L, Power E. Using technology to address recruitment issues in the clinical trial process. Trends Biotechnol. 2002;20:105–9.
40. Bates DW, Leape LL, Cullen DJ, et al. Effect of computerized physician order entry and a team intervention on prevention of serious medication errors. JAMA. 1998;280:1311–6.
41. Butte AJ, Weinstein DA, Kohane IS. Enrolling patients into clinical trials faster using RealTime Recuiting. Proc AMIA Symp. 2000;2000:111–5.
42. Teich JM, Merchia PR, Schmiz JL, et al. Effects of computerized physician order entry on prescribing practices. Arch Intern Med. 2000;160:2741–7.
43. Holford NH, Kimko HC, Monteleone JP, Peck CC. Simulation of clinical trials. Annu Rev Pharmacol Toxicol. 2000;40:209–34.
44. Kim JH, Kohane IS, Ohno-Machado L. Visualization and evaluation of clusters for exploratory analysis of gene expression data. J Biomed Inform. 2002;35:25–36.
45. Fazi P, Grifoni P, Luzi D, et al. Is workflow technology suitable to represent and manage clinical trials? Stud Health Technol Inform. 2000;77:302–6.
46. Fazi P, Luzi D, Manco M, et al. WITH: a system to write clinical trials using XML and RDBMS. Proc AMIA Symp. 2002;2002:240–4.
47. Goodman KW. Using the web as a research tool. MD Comput. 2000;17:13–4.

48. Rubin DL, Gennari J, Musen MA. Knowledge representation and tool support for critiquing clinical trial protocols. Proc AMIA Symp. 2000;2000:724–8.
49. Tai BC, Seldrup J. A review of software for data management, design and analysis of clinical trials. Ann Acad Med Singapore. 2000;29:576–81.
50. Lutz S, Henkind SJ. Recruiting for clinical trials on the web. Healthplan. 2000;41:36–43.
51. Greaves AW, Payne PR, Rassenti L, Kpps TJ. CRC Tissue Core Management System (TCMS): integration of basic science and clinical data for translational research. In: AMIA 2003 symposium. Washington, D.C.; 2003.
52. Payne PRO, Greaves AW, Kipps TJ. CRC Clinical Trials Management System (CTMS): an integrated information management solution for collaborative clinical research. In: AMIA 2003 symposium. Washington, D.C.; 2003.
53. Westgren M, Kublickas M. To use internet in collaborative studies and registers. Acta Obstet Gynecol Scand. 2000;79:329–30.
54. Kuchenbecker J, Dick HB, Schmitz K, Behrens-Baumann W. Use of internet technologies for data acquisition in large clinical trials. Telemed J E Health. 2001;7:73–6.
55. Merzweiler A, Knaup P, Weber R, et al. Recording clinical data – from a general set of record items to case report forms (CRF) for clinics. Medinfo. 2001;10:653–7.
56. Wubbelt P, Fernandez G, Heymer J. Clinical trial management and remote data entry on the internet based on XML case report forms. Stud Health Technol Inform. 2000;77:333–7.
57. Forrester. Global trial starts using web-enabled EDC capture. Cambridge: Forrester Research, Inc; 2006.
58. El Emam K, Jonker E, Sampson M, et al. The use of electronic data capture tools in clinical trials: web-survey of 259 Canadian trials. J Med Internet Res. 2009;11:e8.
59. Harris PA, Taylor R, Thielke R, et al. Research electronic data capture (REDCap) – a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform. 2009;42:377–81.
60. Nahm ML, Pieper CF, Cunningham MM. Quantifying data quality for clinical trials using electronic data capture. PLoS One. 2008;3:e3049.
61. Welker JA. Implementation of electronic data capture systems: barriers and solutions. Contemp Clin Trials. 2007;28:329–36.
62. Khan H, Khan S, Iqbal A. Knowledge, attitudes and practices around health research: the perspective of physicians-in-training in Pakistan. BMC Med Educ. 2009;9:46.
63. Khan SA, Kukafka R, Payne PR, et al. A day in the life of a clinical research coordinator: observations from community practice settings. Stud Health Technol Inform. 2007;129: 247–51.
64. Horsky J, Zhang J, Patel VL. To err is not entirely human: complex technology and user cognition. J Biomed Inform. 2005;38:264–6.
65. Horsky J, Kuperman GJ, Patel VL. Comprehensive analysis of a medication dosing error related to CPOE. J Am Med Inform Assoc. 2005;12:377–82.
66. Horsky J, Kaufman DR, Patel VL. Computer-based drug ordering: evaluation of interaction with a decision-support system. Stud Health Technol Inform. 2004;107:1063–7.
67. Horsky J, Kaufman DR, Patel VL. The cognitive complexity of a provider order entry interface. AMIA Annu Symp Proc. 2003;2003:294–8.
68. Payne PR, Mendonca EA, Starren JB. Modeling participant-related clinical research events using conceptual knowledge acquisition techniques. AMIA Annu Symp Proc. 2007;2007: 593–7.
69. Embi PJ, Payne PR, Kaufman SE, et al. Identifying challenges and opportunities in clinical research informatics: analysis of a facilitated discussion at the 2006 AMIA annual symposium. AMIA Annu Symp Proc. 2007;2007:221–5.
70. Starren JB, Payne PR, Kaufman DR. Human computer interaction issues in Clinical Trials Management Systems. AMIA Annu Symp Proc. 2006;2006:1109.
71. Buetow K. An infrastructure for interconnecting research institutions. Drug Discov Today. 2009;14:605–10.

72. Zerhouni EA. Translational research: moving discovery to practice. Nature. 2007;81:126–8.
73. Short A, Jackson W, Nugus P. Expanding clinical research capacity through a community of practice (CoPER). Nurse Educ Pract. 2009;10:52–6.
74. Fidler DP. "Geographical morality" revisited: international relations, international law, and the controversy over placebo controlled HIV clinical trials in developing countries. Harv Int Law J. 2001;42:299–354.
75. Pike R, Keech AC, Simes RJ. Clinical trials research in the new millennium: the international clinical trials symposium, Sydney, 21–23 October 2002. Med J Aust. 2003;178:316–7.
76. Johnatty RN. Clinical trials in developing countries: discussions at the '9th international symposium on long term clinical trials', London, UK, 19–20 June 2000. Curr Control Trials Cardiovasc Med. 2000;1:55–8.
77. Beran RG. The ethics of clinical research and the conduct of clinical drug trials: international comparisons and codes of conduct. Med Law. 2000;19:501–21.

Chapter 4
Methodological Foundations of Clinical Research

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, binding, and *a priori* definition of the statistical analysis is also discussed. Finally, we look at how adaptive clinical trials can shorten the time and reduce the cost of classical research programs. Such adaptation strategies are relatively new in clinical research and allow for modification of the sample size, adjusting study duration, and other changes.

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

A. Bacchieri, M.S. (✉)
Department of Biostatistics and Data Management,
Research and Development, Sigma-tau Industrie Farmaceutiche Riunite S.p.A.,
Via Pontina km 30.400, Pomezia, Rome, Italy
e-mail: antonella.bacchieri@sigma-tau.it

G.D. Cioppa, M.D.
Global Clinical Research and Development,
Vaccines, Novartis Vaccines and Diagnostics,
1 Via Fiorentina, Siena, Italy
e-mail: giovanni.della_cioppa@novartis.com

R.L. Richesson, J.E. Andrews (eds.), *Clinical Research Informatics*, Health Informatics,
DOI 10.1007/978-1-84882-448-5_4, © Springer-Verlag London Limited 2012

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. 19).

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 thousands of active compounds by means of biological assays.
2. Selection of the lead compound.
3. Synthesis and 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 many non-clinical areas, from toxicology to production, becoming increasingly complex, 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 twofold:

1. Obtain indications on the safety and tolerability of the compound over a wide range of doses.
2. Study its pharmacokinetics in humans.

Whereas traditionally Phase I is conducted in healthy volunteers, increasingly Phase I is carried out directly in patients. Then, a third objective is added:

3. Obtain preliminary pharmacodynamic indications.

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 frequency of administration 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; IIb, for dose-finding.

The aim of *Phase III* is to demonstrate the therapeutic (or preventive or diagnostic) efficacy, 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. The last 25 years have seen a large international effort 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 Harmonization (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. 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, this may be done 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. On the other hand, the lack of such harmonization will necessitate 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 (also known as 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, the so-called cytotoxic drugs used in oncology have many peculiarities in their clinical development, mainly concerning Phases I and II. These differences are determined mostly by the toxicity of these compounds, even at therapeutic or subtherapeutic doses, combined with the life-threatening nature of the diseases in question.

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 *acceleration tools* are the following:

- Simulations, which are 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, 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 even 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 pandemic situations, such as the recent 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 non-cyclical (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 simulates 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

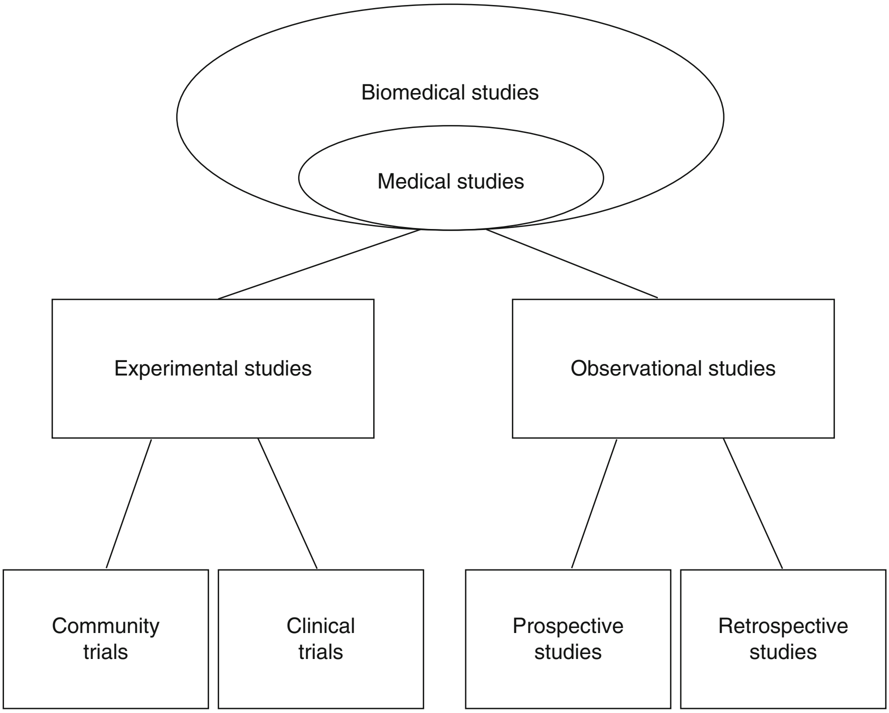


Fig. 4.1 Classification of biomedical studies (Adapted from Bacchieri and Della Cioppa [1], with kind permission of Springer Business + Science Media)

the phenomenon under 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.

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 between which comparisons are made must be homogeneous, 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 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), 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), the other without (controls). In order to ensure homogeneity of the study groups, 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 would be subjects with a diagnosis of lung cancer; each case would 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 Studies

In experimental studies, 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 it by the researcher, generally by means of a procedure called *randomization* (see below). The assignment of the intervention to the study subjects 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 subexperimental 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 subexperimental 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) use of specific study designs (see below). Nevertheless, subexperimental 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.

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. Clearly, investigators would 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 third and 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 in any case, the choice of the threshold for clinical significance (superiority 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 formidable 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 last validation acceptable? Furthermore, the measurement procedure should be described in detail. Our decision is as follows: mechanical sphygmomanometer; three acceptable models (to be reported precisely in protocol); calibration of instruments no more than 6 months before study starts; 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 again undetectable).
- Finally, the question of *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. This last requirement is often ignored by the researchers who design the study, with potentially disastrous outcome.

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 l 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, 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.

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 l 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 differences 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; (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 statistical analysis.

Surrogate and composite end points are special types of end-points often used in clinical trials. *Surrogate end-points* [5] 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 important exceptions, most noticeably oncology, in which Phase I trials are also conducted in patients. The criteria qualifying a person as *healthy* are far from obvious: if a long battery of clinical and laboratory tests are conducted and *normality* is required for every single one, almost everybody would fail. 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 Phase II 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 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. However, 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 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, the eight key ingredients necessary for the sample size calculation are as follows (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 and equivalence studies require more subjects than corresponding crossover designs and 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) and a false-negative result (type II error), the greater the sample size. One can reduce the type I error at the expense 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 increase.
3. The magnitude of the signal (threshold of clinical relevance or equivalence margin): the smaller the clinically relevant difference between treatments we want to detect in a superiority trial or the smaller the equivalence margin we want to allow for in an equivalence/non-inferiority trial (see below), the greater the number of subjects we need.
4. The number of primary end-points and signals: in general, the more primary end-points and signals we have in our protocol, the greater the sample size, as we need to *adjust* 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 equivalence 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 study 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 expected discontinuations affecting 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 fixed combinations 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 represent 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 very 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 pivotal Phase III trials, and 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), etc. 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 (see below). Dose-finding studies (Phase II) require multiple study treatments. 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 (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, some patients may respond better to one treatment than to another, some may be allergic to a particular treatment, some may develop tolerance to one specific compound, and so on.

Equivalence studies play a particularly 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 study as if it were a superiority study. Instead, different methods must be used.

When planning a superiority study, the investigator must select a threshold of clinical relevance (superiority margin), i.e., the smallest difference between treatments, judged *a priori* 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) or the largest difference between treatments, judged *a priori* as clinically irrelevant. A guidance document under the patronage of the EHA Committee for Human Medicinal Products (CHMP) on the choice of the equivalence/non-inferiority margin is available at www.ume.europe.eu.

In superiority studies, the null hypothesis, which we seek to reject in the traditional statistical testing, is that there is no difference between treatments. Oppositely, 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 not due to chance, while if the test is not statistically significant, we can conclude that the difference is likely generated by chance. In equivalence/non-inferiority studies, the statistical test is useless. A statistically significant result does not necessarily imply that the treatments are not equivalent, because the difference between the treatments could be clinically irrelevant and therefore fall within the equivalence margin. A statistically non-significant outcome does not allow accepting that there is no difference between treatments, because the statistical test may not have enough power to detect differences that are bigger than the threshold of equivalence.

The analysis of equivalence studies must be based on confidence intervals. Assuming we use the mean as the group indicator, and the difference between means as the signal the 95% confidence interval on the observed mean difference between the treatments must be calculated (note that the 95% level for the confidence interval is set conventionally, as well as the 5% level for the statistical test). Equivalence between the treatments is demonstrated if such confidence interval is entirely included within the equivalence margin. To grasp the meaning of this rule, it helps to recall that the 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.

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 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 than in superiority studies. This is one of the main reasons why regulatory authorities are often reluctant to allow pivotal Phase III trials with an equivalence or non-inferiority approach for new molecular entities (i.e., non-generic drugs) and request the addition of a placebo arm as well.

The two treatments under comparison could be equivalent or one could be non-inferior to the other simply because both are ineffective. This is another reason why in equivalence/non-inferiority studies it is recommended to include a comparison with placebo aimed at showing superiority of the presumed active treatments to the inactive compound (see guideline ICH E12). With a placebo arm included in the study, the equivalence 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 the treatments under comparison are both efficacious or both non-efficacious. Unfortunately, this reasoning is theoretical and often far 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

In the section above, the experimental and subexperimental factors have been defined. The experimental design is the logical structure of an experiment. By means of the experimental design, the researcher controls the experimental factors, typically the study treatments being compared and some of the most important subexperimental factors, typically key protective and/or risk factors. There are two main objectives of the experimental design: (1) minimize the systematic error or bias, between the 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 subexperimental factors). The remaining part of variability, i.e., the part that cannot be attributed to any known factor, is attributed to accidental factors. These all together are named *chance*. This unexplained variability (due to chance), often referred to as *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 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 no systematic errors are present, and *precise*, which means that the part of variability not explained by the considered factors must be as small as possible. The different designs must be evaluated as for these two main characteristics. In addition, in choosing the study design, the researcher must always keep simplicity in mind, i.e., 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 and After Comparisons in a Single 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* refer to the beginning and the end of treatment, respectively.

This study design is simple and very 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 the context of the study.

- 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 [6, 7].
- 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 occur 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 prevail. 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.

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 trough 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 component 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 component, 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, of which we will introduce the following: simple randomization, randomization in blocks, stratified randomization, and variants that allow allocation of patients to treatments based on information collected during the study (adaptive 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 the same 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.

The so-called *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 so-called *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 result of a preestablished end-point, 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).

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 treatment factor with T and assume it has k levels, i.e., $T_1,\ldots, T_k$. These 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.

In this design, it is possible to estimate only the treatment effect. Accordingly, the total variability is decomposed into two components: the part explained by the treatment and the part unexplained by the treatment, which is 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:

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 subexperimental 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:

Table 4.1 The parallel group design matrix

T_1	T_2	…	T_k
Y_{11}	Y_{21}		Y_{k1}
…	…	…	…
Y_{1n1}	Y_{2n2}		Y_{knk}

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 subexperimental 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 subexperimental factor(s) are categorized in classes (*strata*). Let us assume we chose 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 treatment factor as above 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, i.e., without considering the treatment.
- 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, like wise, 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 decomposed into four parts: the part explained by the treatment, the part explained by the subexperimental factor(s), the part explained by the interaction between the treatment and the subexperimental factor(s), and the residual variability attributed to chance (computed by

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_{11n11}	Y_{21n21}		Y_{k1nk1}
S_2 adolescents	Y_{121}	Y_{221}		Y_{k21}
	…	…	…	…
	Y_{12n12}	Y_{22n22}		Y_{k2nk2}
S_3 non-elderly adults	Y_{131}	Y_{231}		Y_{k31}
	…	…	…	…
	Y_{13n13}	Y_{23n23}		Y_{k3nk3}
S_4 elderly adults	Y_{141}	Y_{241}		Y_{k41}
	…	…	…	…
	$\mathrm{Y}_{\mathrm{14n14}}$	$\mathrm{Y}_{\mathrm{24n24}}$		$\mathrm{Y}_{\mathrm{k4nk4}}$

averaging the estimates of the variability calculated within each stratum). If the factor used for the stratification is a real prognostic factor, the residual (chance) 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 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. This special case is referred to as the *matched-paired design*, which is the variant of randomized block design most often used in clinical trials. Often the randomized block design is used in clinical trials when the time of enrolment is one of the factors that should be controlled for. Time can be a known prognostic factor (e.g., asthma, Reynaud syndrome) or just a subexperimental factor with unknown prognostic value (e.g., 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), 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 cross-over design (i.e., a design with the same sample site in each period and each sequence) is shown in Table 4.3.

Table 4.3 The crossover design matrix

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

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

In this 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; it should be noted that 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.
- Subject 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.

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. In the crossover design, the subject and the sequence effects have a very limited interest per se. These factors are useful for reducing the residual variability.

The statistical analysis typically starts with the test of the carry-over 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., they 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 often 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 other words, if the measurements on the same subject are highly correlated, the crossover design generates a test on the treatment effect which is more powerful as compared to the one for a design based on between-subject comparisons (parallel group). Therefore, it requires a smaller sample.

In summary, 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 disadvantages:

- The crossover design is more complex for the logistical aspects and more problematic on methodological grounds than the parallel group design.
- The subjects, before receiving the second treatment, must be back to their baseline conditions, i.e., the treatment effect must be fully reversible.
- 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.

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 four, 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 four 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 any 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).

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

Table 4.4 The Latin square design matrix

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

square designs shown above. This design maintains some level of balance: each treatment appears in the same number of sequences, each pair of treatments 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. Only once these are addressed will the classical methodological criteria for design selection be used. 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.

Adaptive Clinical Trials

The classical clinical development of new pharmacological compounds is based on a chain of subsequent studies, where the researcher has to wait until the end of the previous trials before planning the next. At the end of a long clinical development process the registration decision is mostly made on the data collected in Phase III studies.

As mentioned at the beginning of this chapter, in most therapeutic areas, the costs and times associated with clinical development plans are becoming prohibitive, and this, in turn, has ultimately a tremendous impact on the price of the drugs and the time that the patients should wait before new therapies can reach them.

It is obvious that a lot of effort is devoted to shorten the time and reduce the cost of the classical research programs. To this end, one of the most promising approaches is that of implementing *adaptation* strategies, i.e., programs that are not fixed from the very beginning but can be changed based on interim looks at the data. This strategy can be used at the individual study level to modify ongoing trials that could not continue successfully without appropriate changes or, at a more general level, to allow subsequent trials to overlap partially, i.e., the next trial to start before the previous one is finished.

The adaptation strategy is based on the so-called *adaptive designs*, the use of which is relatively new in clinical research. Adaptive designs allow the flexibility to modify sample size, terminate early one or more treatment arms, change the study duration, or make other changes, based on the evidence generated by accumulating data.

The adaptive methodology is applicable in those situations where the enrolment is relatively slow, the efficacy end-point can be evaluated rapidly, and the data can be collected and analyzed quickly. A landmark paper is [8], which proposes a general method for combining samples obtained before and after a preplanned interim analysis. This is adaptive in the sense that all information available at the time of performing this analysis may be used for planning the subsequent steps of the study(ies). One important stream of research in the adaptive design area has been sample size reassessment [9–12].

Various other design modifications have been considered, for example, redefining multiple end-points [13, 14] or changing the study design [15]. At present, an important role of the adaptive designs is played in Phase II of drug development. In fact, the high rate of study failure in Phase III (45% according to [16]) is to a large extent determined by a wrong choice of the dose, which is made in Phase II. The adaptive designs may be very useful in Phase II because, for a given sample size, they allow to explore a bigger number of doses than the fixed designs, to collect more data on those doses that provide meaningful information on the dose-response curve (i.e., those in the steep part of the curve) and that are more promising in terms of efficacy and safety. Interesting papers on methodology for adaptive designs in Phase II are [17, 18].

In Phase III, an interesting approach is that of the integration of Phase II and Phase III into one adaptive trial. This of course creates major issues, both operational and methodological.

The conventional approach is to conduct one or more Phase II trials in which several doses and/or dose schedules of an experimental drug are evaluated in terms of a preestablished primary end-point. The Phase II results are then used for deciding whether subsequent Phase III trials should be conducted and at what dose(s). The final analysis relies only on the data collected in the Phase III trials, individually considered. Generally, the Phase II and III primary end-points are different, being the former biomarkers or surrogate end-points and the latter clinical end-points, i.e., variables able to capture the clinical benefit for the patient, if any. Sometimes, however, the Phase II and III end-points are the same variable evaluated at different times, namely, the Phase II end-point is assessed at an early time, while the Phase III end-point requires a much longer follow-up. Only occasionally can the same end-point be used in Phase II and III trials: this happens when it is compatible with a relatively short observation period and small sample size (an example is the forced expiratory volume in 1 s [FEV1] in trials evaluating bronchodilators in asthma or chronic obstructive pulmonary disease). The drawbacks of the traditional sequential approach are essentially two: development is delayed due to the pause in patient enrolment between the end of Phase II and the start of Phase III, and the Phase II data are not incorporated in the final analysis for efficacy, which is based on Phase III data only.

The adaptive 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 experimental treatments (generally more doses of the same treatment) and a control, and at the end 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 sometimes also statistical methods. At the end of the study, data from both stages are combined for the final analysis. Interesting publications on the issue of combining the two phases are [19–22].

Different research approaches and study designs are appropriate in selected situations. New technologies (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. Milano: Springer; 2007.
2. Rubin AA. New drugs; discovery and development. New York: Marcel Dekker; 1978.
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. Pretince R. Surrogate end-points in clinical trials: definition and operational criteria. Stat Med. 1989;8:431–40.
6. Bland JM, Altman DG. Regression toward the mean. BMJ. 1994;308:1499.
7. Bland JM, Altman DG. Some examples of regression toward the mean. BMJ. 1994;309:780.
8. Bauer P, Kohne K. Evaluation of experiments with adaptive interim analyses. Biometrics. 1994;50:1029–41.
9. Jennison C, Tumbull BW. Mid-course sample size modification in clinical trials based on the observed treatment effect. Stat Med. 2003;22:971–93.
10. Proscham M, Liu Q, Hunsberger S. Practical mid-course sample size modification in clinical trials. Control Clin Trials. 2003;24:4–15.
11. Shun Z. Sample size re-estimation in clinical trials. Drug Inf J. 2001;35:1409–22.
12. Gould AL. Sample size re-estimation: recent developments and practical considerations. Stat Med. 2001;20:2625–43.
13. Kieser M, Bauer P, Lehmacher W. Inference on multiple endpoints in clinical trials with adaptive interim analyses. Biom J. 1999;41:261–77.
14. Hommel G. Adaptive modifications of hypotheses after an interim analysis. Biom J. 2001;43:581–9.
15. 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.
16. Kola I, Landis J. Can the pharmaceutical industry reduce attrition rates? Nat Rev Drug Discov. 2004;3:711–6.
17. Gaydos B, Krams M, Perevozskaya I, et al. Adaptive dose-response studies. Drug Inf J. 2006;40:451–61.
18. Bauer P, Rohmel J. An adaptive method for establishing a dose-response relationship. Stat Med. 1995;14:1595–607.
19. 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.
20. Liu Q, Pledger GW. Phase 2 and 3 combination designs to accelerate drug development. J Am Stat Assoc. 2005;100:493–502.
21. Liu Q, Proscham MA, Pledger GW. A unified theory of two-stage adaptive designs. J Am Stat Soc. 2002;97:1034–41.
22. Bauer P, Kieser M. Combining different phases in the development of medical treatments within a single trial. Stat Med. 1999;18:1833–48.

Chapter 5
Informatics Approaches to Participant Recruitment

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

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

C. Weng, Ph.D. (✉)
Department of Biomedical Informatics, Columbia University,
622 W 168th Street, VC-5, New York, NY 10332, USA
e-mail: cw2384@columbia.edu

P.J. Embi, M.D., M.S.
Department of Biomedical Informatics and Internal Medicine,
The Ohio State University Medical Center,
3190 Graves Hall, 333 W. 10th Avenue, Columbus, OH 43210, USA

R.L. Richesson, J.E. Andrews (eds.), *Clinical Research Informatics*, Health Informatics,
DOI 10.1007/978-1-84882-448-5_5, © Springer-Verlag London Limited 2012

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] 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 trials 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, to the use of government and privately sponsored websites (Fig. 5.1). 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 matching tools, clinical data repositories or warehouses, or clinical registries [11–16] (Figs. 5.2, 5.3, and 5.4). Figures 5.2, 5.3, and 5.4 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

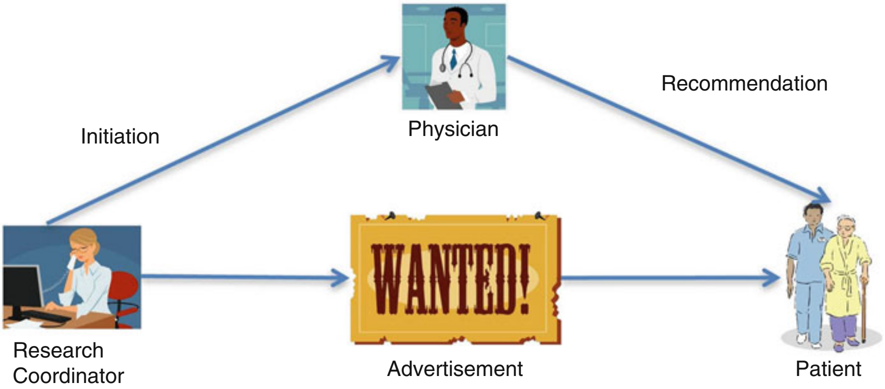


Fig. 5.1 Traditional researcher-initiated trial recruitment workflow

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. 5.2), 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. 5.3) 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. 5.4) 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. 5.2, 5.3, and 5.4).

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

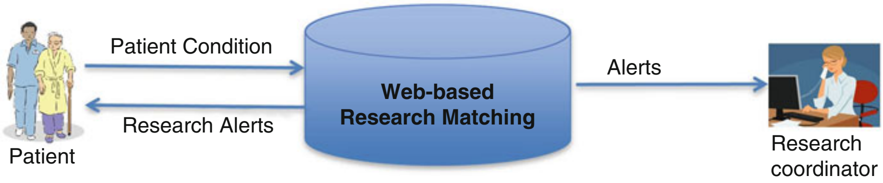


Fig. 5.2 Patient-initiated research recruitment workflow

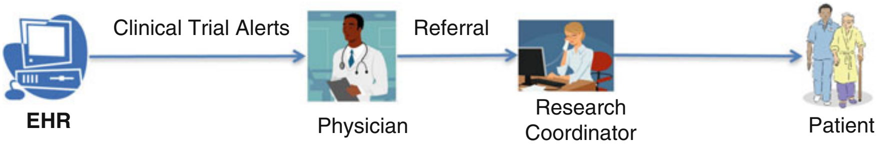


Fig. 5.3 Healthcare provider–initiated research recruitment workflow

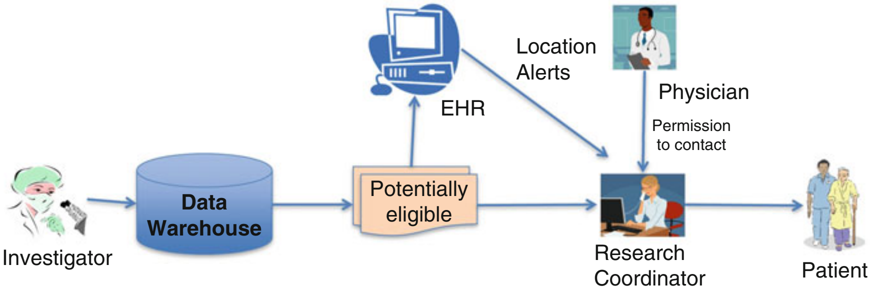


Fig. 5.4 Data warehouse–aided research screening and recruitment model

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 United States [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.

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 [38]. 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 10-fold 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 [39] 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 [40–42].

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 [43]. 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% [44]. 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 [45]. 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 [46]. 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 [47, 48]. 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 [49]; 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 [50], 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 [51]. 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 [52, 53], 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 [54], 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 [54]. 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 [55]. 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 [56]. 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) [57], 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 [58]. 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 [59]. MLP has demonstrated its broad uses for a variety of applications, such as extracting knowledge from medical literature [60, 61], indexing radiology reports in clinical information systems [62–64], and abstracting or summarizing patient characteristics [65]. One of the widely used tools is MetaMap Transfer (MMTx) [66], which is available to biomedical researchers in a generic, configurable environment. It maps arbitrary text to concepts in the UMLS Metathesaurus [67]. Chapman demonstrated in her studies that MLP is superior to ICD-9 in detecting cases and syndromes from chief complaint reports [68, 69]; this finding was also confirmed by Li et al. in a study comparing discharge summaries and ICD-9 codes for recruitment uses [53]. The most mature MLP system is MedLEE [70]. In numerous evaluations carried out by independent users, MedLEE performed well [71]. 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 [57]. 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 identification underscore a great need for a long-range research plan to design and evaluate different methods to surmount the social, organizational, and technical challenges facing clinical trial recruitment, the key components of the plan being: (1) to improve the data quality and completeness for EHR systems, (2) to design better data presentation techniques for EHR systems to enable patient-centered, problem-oriented data views, (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, separate 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." [72] 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.

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. June 1, 2006.
5. Sinackevich N, Tassignon J-P. Speeding the critical path. Appl Clin Trials. January 1, 2004.
6. Sullivan J. Subject recruitment and retention: barriers to success. Appl Clin Trials. April 1, 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.
11. Breitfeld PP, Weisburd M, Overhage JM, Sledge Jr G, 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, Theocharous 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. Califf 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, D.C.; 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, D.C.; 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, D.C.; 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 Recuiting. 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, Oct 1998.
34. Cohen E al. e. 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, CA, http://crisummit2010.amia.org/files/symposium2008/S14_Niland.pdf, Accessed on December 13, 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.vanderbilthealth.com/clinicaltrials/13133. J Am Med Inform Assoc. 2005;12(6):608–13.
38. 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.
39. 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.
40. 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.

41. 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.
42. 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.
43. 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.
44. 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.
45. 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.
46. Sung NS, Crowley Jr WF, 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.
47. 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.
48. 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.
49. Parker CG, Embley DW. Generating medical logic modules for clinical trial eligibility criteria. AMIA Annu Symp Proc. 2003;2003:964.
50. 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.
51. Lin J-H, Haug PJ. Data preparation framework for preprocessing clinical data in data mining. AMIA Annu Symp Proc. 2006;2006:489–93.
52. Carlo L, Chase H, Weng C. Reconciling structured and unstructured medical problems using UMLS. Proc AMIA Fall Symp. 2010:91–5.
53. 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.
54. Kahn MG. Integrating electronic health records and clinical trials. 2007. http://www.esi-bethesda.com/ncrrworkshops/clinicalResearch/pdf/MichaelKahnPaper.pdf. Accessed Aug 2011.
55. Lewis JR. IBM computer usability satisfaction questionnaires: psychometric evaluation and instructions for use. Int J Hum-Comput Interact. 1995;7(1):57.
56. Ruberg S. A proposal and challenge for a new approach to integrated electronic solutions. Appl Clin Trials. 2002;2002:42–9.
57. Chute C. The horizontal and vertical nature of patient phenotype retrieval: new directions for clinical text processing. Proc AMIA Symp. 2002:165–9.
58. Weng C, Tu SW, Sim I, Richesson R. Formal representation of eligibility criteria: a literature review. J Biomed Inform. 2010;43(3):451–67.
59. Friedman C, Hripcsak G. Natural language processing and its future in medicine. Acad Med. 1999;74:890–5.
60. Friedman C, Chen L. Extracting phenotypic information from the literature via natural language. Stud Health Technol Inform. 2004;107:758–62.
61. 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.
62. 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.

63. 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.
64. 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.
65. Baud R, Lovis C, Ruch P, Rassinoux A. Conceptual search in electronic patient record. Medinfo. 2001;84:156–60.
66. 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.
67. Brailer 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.
68. 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.
69. 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.
70. Friedman C. Towards a comprehensive medical language processing system: methods and issues. Proc AMIA Annu Fall Symp. 1997:595–9.
71. 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.
72. Conway PH, Clancy C. Commentary. Transformation of health care at the front line. JAMA. 2009;301(7):763–5. doi:10.1001/jama.2009.103.

Chapter 6
The Evolving Role of Consumers

J. David Johnson and James E. Andrews

Abstract The culmination of the changes in health care, 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. For those who are incapable (or unwilling) to develop the requisite health literacy skills, there is also a tendency to turn to intermediaries for advice. This chapter explores these developments from various perspectives and looks at some foundational issues in health communication as related to health consumerism. The overarching concern is the information environment within which health consumers are immersed and some of the 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 behavior research have relevance for our current understanding and future studies of the evolving role of the consumer.

Keywords Health consumerism • National Library of Medicine • Consumer health information • Web-based information resources • Consumer health movement • Patient empowerment • Public access technologies • Personalization of medicine • Public Internet access

> 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

J.D. Johnson, Ph.D., M.A., B.S. (✉)
Department of Communication, University of Kentucky,
242 Grehan Building, Lexington, KY 40506, USA
e-mail: jdj@uky.edu

J.E. Andrews, Ph.D.
School of Information, College of Arts and Sciences, University of South Florida,
4202 E. Fowler Ave, CIS 1040, Tampa, FL 33620-8100, USA
e-mail: jimandrews@usf.edu

R.L. Richesson, J.E. Andrews (eds.), *Clinical Research Informatics*, Health Informatics,
DOI 10.1007/978-1-84882-448-5_6, © Springer-Verlag London Limited 2012

> 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

The role of patients as consumers has been evolving for over a generation. Patients are central to clinical research, and there are a number of challenges and issues stemming from the health consumerism movement that stand to impact the conduct and ultimate success of clinical research. Accelerating the trends in health consumerism are a host of technology and decision-support-related advances, including enhanced access to authoritative web-based information resources, social networking capabilities, and personal decision aids. Generally, the goal is 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]. Consumerism and empowerment, however, assume and require a level of health literacy on the part of consumers and an understanding of challenges and implications of this by developers, researchers, and healthcare providers.

Increasingly, the responsibility for health-related matters is passing to the individual, partly because of legal decisions, which have entitled patients to full information access. Ever since the 1970s, patients have increasingly become more active participants in decisions affecting health care [4]. The overload of information on today's health professionals forces decentralization of responsibilities, placing the onus on individuals to inform themselves if they are going to receive up-to-date treatment, identify clinical studies for which they are suited and eligible, and utilize the variety of consumer health tools now available. In effect, patients must often do the traditional work of doctors, who cannot possibly keep up with the breadth and depth of information related to specific research advances and studies and requisite eligibility criteria for their patients. Recognition of the limitations on health professionals also requires individuals to be able to confirm and corroborate information by using multiple sources in various formats, in other words, to develop effective health information literacy skills. Their efforts now can be more easily pooled because of advances in health information technology, social networking, and related information and communication technologies (ICTs) and tools to produce informed consumers who share information with each other and whose collective knowledge may even exceed that of some health professionals, especially in terms of issues related to everyday life with a disease.

In the U.S., the Federal government, particularly the National Library of Medicine (NLM), has played a significant role in increasing the public's access to authoritative information [5]. Early government efforts focused on telephone hotlines such as the Cancer Information Service [6], and more recently, the government has focused on providing information through access to authoritative databases such as NLM's PubMed, which first became free to the public in 1997 [7]. Many NLM sites now provide users with the capability to develop profiles that keep them abreast of self-defined areas of interest as relevant material is added to a database by using RSS feeds. The Federal government is thus promoting a rich infrastructure from which individuals can draw information, enhancing their access to richer information fields [8].

The culmination of the changes in health care, motivated in many ways by the rapid evolution of 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. For those who are incapable (or unwilling) to develop the requisite health literacy skills, there is also a tendency to turn to intermediaries for advice. This chapter explores these developments from various perspectives and looks at some foundational issues in health communication as related to health consumerism. The overarching concern is the information environment within which health consumers are immersed and some of the 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 behavior research have relevance for our current understanding and future studies of the evolving role of the consumer.

Traditional Perspectives: Health Campaigns and Information Behaviors

Public communication campaigns represent "purposive attempts to inform, persuade, or motivate behavior changes in a relatively well-defined and large audience, generally for noncommercial 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" [9]. We know a lot about how formal organizations (e.g., the National Cancer Institute, the American Cancer Society) conduct campaigns to change individual behaviors [10]. Increasingly, however, individual action, embodied most clearly in information seeking, determines what messages individuals will be exposed to and how they will behave. In our view, actors operate in "information fields" (covered in greater depth later in the chapter) 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 information campaigns that tend to view the world as rational, known, and which concentrate on controlling individuals to seek values of efficiency and effectiveness [9].

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 retrieval. 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 [11]. 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 within social networks [12], with growing importance attached to social media.

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. A true 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 [13, 14]; for our purposes, however, 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 have learned that these classic approaches are not very effective unless the needs of the audience and their reaction to messages are considered [15, 16]. 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 [17–20]. Indeed, campaigns tend to reach those who are already interested and typically bypass those who are most in need of their messages [19]. In effect, campaigns reach the 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 [17].

One of the areas where the limitations of public campaigns is most clearly revealed is in the difficulty and considerable expense involved in recruiting people into clinical research studies, which has prompted initiatives like the Army of Women addresses (www.armyofwomen.org). According to Allison [21], less ~3% of eligible cancer patients enroll in trials, and roughly one in five of NCI-sponsored trials fail to meet their necessary enrollment [22]. The challenge of trial recruitment becomes especially pronounced in the area of 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 [23], and researchers must consider whether they are getting representative samples given that those seeking trials might disproportionately represent certain demographics [24]. 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.

The Social World of Health Consumers

The most obvious and compelling development in consumer health over the past 15 years or more has been the emergence of a dynamic social world facilitated by the Internet and social media. The interactions and relationships among people, the evolving healthcare environment, technology, and information resources and carriers are incredibly complex and continually in flux. The frequently cited Pew Internet report on the social life of health information showed that 61% of adults seek health information online [25]. 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 [25]. In addition to seeking health information that ranges in complexity, there are increasing online support groups that are now showing signs of fostering patient empowerment or management [26] and participation tools that may lead to more positive outcomes, especially for rare diseases [27]. The consumer health environment is expanding at a breakneck pace, but the underlying theoretical issues and social dynamics are not terribly different from those that preceded the Internet age.

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 of one kind or another, is important as we consider the technologies and approaches that currently populate the landscape of consumerism 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. The role of third parties is also discussed, including brokers and advocacy groups.

Information Fields

As suggested above, one conception of an information environment is that of the information field within which the individual is embedded [28]. An individual's information field provides the more static context for their information seeking, containing resources, constraints, and carriers of information [4, 29] It provides the starting point for information seeking [30] representing the typical arrangement of information stimuli to which an individual is regularly exposed [11] and the information resources they routinely use [31]. Individuals are embedded in a physical world that involves recurring contacts with an interpersonal network of friends and/or family. 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 [32] with interesting recent variants such as the information horizons approach [31]. 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 is greatly facilitated by the Internet and explosion of choices among even traditional media such as cable television and magazines.

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, videotapes, and more tailored electronic communication—described later—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 [33]. 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 [34].

In some ways, the total of someone'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, they also share a context which provides the information grounds for further interaction [35]. 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 recent studies that demonstrate a clear link between individuals' positioning in social networks and their health [36, 37]. 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 [38, 39]. These networks, in effect, constitute elaborate feedback processes through which individual behavior is regulated and maintained [38, 39].

Social networks are often viewed as the infrastructure of social support with social support seen as "…inextricably woven into communication behavior" [38, 39]. 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 [40], but they also need authoritative professional guidance in the institution of proper treatment protocols, selection of 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" [41]. 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 [8]. This concept has been intimately tied to the flow of information. Weak ties' notions are derived from the work of Granovetter [42] 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 [38]. 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 [26]. 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

There are a number of ways that use of third parties, particularly knowledge brokers, can complement clinical practice and, by extension, research. First, individuals who want to be fully prepared before they visit the doctor often consult the Internet [43, 44]. In fact, Lowery and Anderson [45] 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 [25, 44]. While the Internet 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 [46]. 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.

Although intermediaries play an important role despite more consumer health information on the Internet, increasing health literacy by encouraging autonomous information seekers also should be a goal of our healthcare system [47]. While it is well known that individuals often consult a variety of others before presenting themselves in clinical settings [4] outside of HMO and organization 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, diagnosis, drug regimens) may be established that will be unfulfilled in the clinical encounter.

The emergence of the Internet as an omnibus source of information has apparently changed the nature of opinion leadership; 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 [48]. This is part of a broader trend that Shapiro [49] refers to as "disintermediation," or the capability of the Internet 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 essence, 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 searches. 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 [18] and gatekeepers [50] and instantiates weak ties [51]. 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 [52], reinforcing messages by their social influence over them [18], and validating the authoritativeness of the information [53]. 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 [52]. 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.

Self-help Groups

Increasingly, more formal groups, acting as crowd-sourced medicine, are serving as opinion leaders and information seekers for or supporting the everyday health information needs of individuals. Self-help groups are estimated to be in the hundreds of thousands across a wide variety of diseases with members numbering in the millions [26]. 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? etc. 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. Driving this movement has been the notion that self-help groups have the potential to affect outcomes by supporting patients' general well-being and sense of personal empowerment [26], and the diversity of tools now available have the potential to further enable this.

The Internet 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. Perhaps the most prominent 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 [54]. As noted by a few of its developers, 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?" [55]. 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 [56]. Computer-mediated support groups (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 [57].

Advocacy Groups

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, exchange knowledge that is recognized as different from that of health professionals, and to speak with a more unified voice to impact policy and related matters [58]. 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 [59]. 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. They also may expose their members to risky experimental treatments prematurely, though they would argue that these individual costs are often for the greater good of their members and that the patriarchal attitude of cancer (or other) researchers has shrouded scandal in secrecy and led to an orientation toward treatment instead of prevention [59]. Advocacy groups are particularly active when there are no clear treatment options or when they are perceived to be ineffective [59]. Thus, at times individual and group interests coincide, and at times, of course, they do not.

Advocacy groups for cancer (e.g., Army of Women, Breast Cancer Action, Patient Advocates for Advanced Cancer Treatments) serve as increasingly important lobbyists for the provision of information. For instance, these groups have been very successful in increasing research funds for breast cancer research. Much of their influence is due to how deeply cancer has pervaded our society and to the vast numbers of individuals who have either been personally diagnosed or have been impacted by family members and friends who have suffered from some form of cancer. These groups actively seek more money for research that leads to information for databases, to enhanced access to and availability of information, and so on. In short, they lobby for an information infrastructure.

The international growth of advocacy groups for rare diseases has been particularly interesting. Because rare diseases affect small numbers of people who are geographically dispersed, there is a need, indeed an intense desire, for these patients and families to connect. A few decades ago, prior to the explosion of health information accessible on the Internet, many of the stories of how these groups began were similar: A family has a sick member with a disease that perhaps only a few dozen others have; they seek more answers and often become experts in the disease; and, ultimately, they are able to find others with similar challenges and needs and are able to work together to promote change and gain attention. The notion of "expert patient" is one that has gained increased attention, particularly in rare diseases, given the amount and variety of information available on the Internet, the concern by some health professionals that patients may disintermediate existing power structures (though many others encourage knowledge acquisition by patients), and the increased ability to form communities [58]. Most compelling, however, has been how these groups are driving research. There are a number of examples, such as the Army of Women, where they have been highly successful at generating funds for clinical studies, enhancing participation in trials (particularly challenging in rare diseases), and garnering support for more trials on orphan drugs.

Emerging Technologies and Models

We asserted at the beginning of this chapter that a major impetus of the consumer health movement is patient or consumer empowerment. We know that health professionals remain the primary resource for authoritative information, although many augment this with information seeking prior to or following consultations. Still, there continues to be a shift from traditional models of medicine and medical research to ones where patients have a greater role in their own decisions, from treatment options to involvement in clinical research to actually initiating and conducting research. The core issues relate to more than simple choice, but rather choice for achieving more personalized care, for increasing safety in research and care, and for accomplishing other altruistic purposes that require social networks that can enable knowledge transfer, greater voice, and concerted action evoking the wisdom of crowds.

Saying that the Internet is what has heralded so many changes in our world is almost cliché now, particularly in emerging social networks. However, this has been exactly the case with Internet sites and the technologies and functionality that they employ to meet the information needs of consumers. The term *Web 2.0* has been widely used to describe social computing and other technologies that have gone beyond more static (or at least less interactive) Web applications. As correctly noted by Eysenbach [60], however, neologisms such as this may be easily dismissed; however, to dismiss the overall impact of the technologies this implies and the optimism for what these will yet bring, particularly in the area of health, would be erroneous.

As noted earlier, PatientsLikeMe is a robust and prominent harbinger of what is to come. The site essentially furthers the notion of the "patient researcher" and serves many of the same positive elements of social interaction, as described above (e.g., emotional, social, and informational support). In fact, there is some work showing promise in how PLM is positively impacting outcomes [27]. Yet unlike social groups of the past, sites such as PLM are much more dynamic. For example, thousands of patient 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/visualized self-tracking, and opportunities for other treatments or research participation. Of interest here is how such a technology has seemed to accelerate the kind of networks and patient interactions that have benefited consumers and how this acceleration seems to be helping to strengthen and 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 also 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, there are clear challenges and issues. While much of the traditional medical literature has focused on very real concerns about poor health literacy and the growing gaps in knowledge/awareness of large segments of the public [61, 62], 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 [63]. In response to these trends, governmental agencies are adopting policies to promote information equity among various segments of our society [64], but some question whether access to information resources can ever truly be universal, in spite of the best intentions of our policy makers [65].

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 [66]. 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 [67].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 [54]. 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 [68]. 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 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.

Clinical Trial Recruitment

Another attractive aspect of health-related social networks is the potential to overcome the discouraging barriers to patient recruitment into clinical trials [67] and other research projects. Projects like the Army of Women can greatly facilitate researcher access to willing populations for those who go through their elaborate approval process. There are certainly 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 [21].

As noted earlier, social networking sites 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 [67]. Again, the nature of the participants in many of these sites seems to 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 trial recruitment sites failing to recruit a single patient [21], online patient communities offer a far more refreshing outlook. 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 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. 13).

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. 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. 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. While the prospects are exciting, there continue to be serious concerns over the information literacy levels of most people and the resulting risk of major segments of the population being left behind. 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 [67, 69].

All this also begins to raise the question of whose information is it anyway? The social norms that cast doctors and public health officials as the brokers of medical information are yielding to an era in which individuals actively seek information and in which a balance of power is sought in the patient-provider relationship. Information that is necessary to a client for coping with cancer may be seen by doctors as an intrusion into their prerogatives. Exacerbating this problem is the fact that doctors and patients may not share similar outcome goals. Traditionally, doctors have viewed the ideal patient as one who came to them recognizing their authority and who were willing to comply totally (and with enthusiasm) with recommended therapies [70]. So, for example, most doctors believe in treating cancers aggressively, even those with low cure rates; however, increasingly some more harmful aspects of chemotherapy and other treatments are weighed against the likelihood of success, the quality of life, and costs. So, while doctors typically engage in narrow problem-solving relating to the disease, patients often view a disease as but one component of a complex social system of which they are a part. What good does it do to save me if I will be but a shell of my former self and my family is bankrupted in the process? These issues are particularly salient for often highly experimental 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. doi: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. doi:10.1016/j.healthpol.2008.03.006.
4. Johnson JD. Cancer-related information seeking. Cresskill: Hampton Press; 1997.
5. Lindberg DAB, Humphreys BL. Rising expectations: access to biomedical information. Yearb Med Inform. 2008;3:165–72.
6. Johnson JD. Innovation and knowledge management: the Cancer Information Science Research Consortium. Cheltenham: Edward Elgar; 2005.
7. Miller N, Tyler RJ, Backus JEB. MEDLINE Plus: the National Library of Medicine brings quality information to health consumers. Libr Trends. 2004;53:375–88.
8. Johnson JD. Managing knowledge networks. Cambridge: Cambridge University Press; 2009.
9. 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.
10. Atkin C, Walleck L, editors. Mass communication and public health. Newbury Park: Sage; 1990.
11. 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. doi:10.1016/j.ipm.2004.12.001.
12. 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. doi:10.1080/10810730500461059.
13. 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.
14. Noar SM. Challenges in evaluating health communication campaigns: defining the issues. Commun Methods Meas. 2009;3:1–11. doi:10.1080/19312450902809367.
15. Freimuth VS. Improve the cancer knowledge gap between whites and African Americans. J Natl Cancer Inst. 1993;14:81–92.
16. Freimuth VS, Stein JA, Kean TJ. Searching for health information: the Cancer Information Service model. Philadelphia: University of Pennsylvania Press; 1989.
17. Alcalay R. The impact of mass communication campaigns in the health field. Soc Sci Med. 1983;17:87–94. doi:10.1016/0277-9536(83)90359-3.
18. Katz E, Lazersfeld PF. Personal influence: the part played by people in the flow of mass communications. New York: Free Press; 1955.
19. Lichter I. Communication in cancer care. New York: Churchill Livingstone; 1987.
20. Rogers EM, Storey JD. Communication campaigns. In: Berger CR, Chaffee SH, editors. Handbook of communication science. Newbury Park: Sage; 1987. p. 817–46.
21. Allison M. Can web 2.0 reboot clinical trials? Nat Biotechnol. 2009;27:895–902. doi:10.1038/nbt1009-895.
22. 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.
23. 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. doi:10.2196/jmir.9.2.e13.
24. Marks L, Power E. Using technology to address recruitment issues in the clinical trial process. Trends Biotechnol. 2002;20:105–9. doi:10.1016/S0167-7799(02)01881-4.
25. 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.
26. Barak A, Boniel-Nissim M, Suler J. Fostering empowerment in online support groups. Comput Hum Behav. 2008;24:1867–83. doi:10.1016/j.chb.2008.02.004.

27. 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. doi:10.2196/jmir.1549.
28. Cool C. The concept of situation in information science. Annu Rev Inf Sci Technol. 2001;35: 5–42.
29. Johnson JD. Information seeking: an organizational dilemma. Westport: Quorom Books; 1996.
30. Rice RE, McCreadie M, Chang SL. Accessing and browsing information and communication. Cambridge: MIT Press; 2001.
31. 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 socioeconomic students' information seeking behavior. New Rev Inf Behav Res. 2001;2:65–85.
32. Scott J. Social network analysis: a handbook. 2nd ed. Thousand Oaks: Sage; 2000.
33. Kuhlthau CC. Inside the search process: information seeking from the user's perspective. J Am Soc Inf Sci Technol. 1991;42:361–71. doi:10.1002/(SICI)1097-4571(199106)42:5<361::AID-ASI6>3.0.CO;2-#.
34. 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. doi:10.1016/S0740-8188(98)90004-4.
35. 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. doi:10.1002/asi.20019.
36. 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. doi:10.1016/j.socnet.2008.08.003.
37. Cornwell B. Good health and the bridging of structural holes. Soc Netw. 2009;31:92–103. doi:10.1016/j.socnet.2008.10.005.
38. 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.
39. 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.
40. 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.
41. Rogers EM, Kincaid DL. Communication networks: toward a new paradigm for research. New York: Free Press; 1981.
42. Granovetter MS. The strength of weak ties. AJS. 1973;78:1360–80.
43. 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.
44. 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.
45. 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.
46. Johnson JD. Dosage: a bridging metaphor for theory and practice. Int J Strateg Commun. 2008;2:137–53. doi:10.1080/15531180801958204.
47. 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, Inc; 2003. p. 637–50.

48. 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. doi:10.1002/asi.20000.
49. Shapiro AL. The control revolution……: how the internet is putting individuals in charge and changing the world we know. New York: Public Affairs; 1999.
50. Metoyer-Duran C. Information gatekeepers. Annu Rev Inf Sci Technol. 1993;28:111–50.
51. Burt RS. Structural holes: the social structure of competition. Cambridge: Harvard University Press; 1992.
52. Katz E. The two step flow of communication: an up to date report on an hypothesis. Public Opin Q. 1957;21:61–78.
53. Paisley WJ. Knowledge utilization: the role of new communications technologies. J Am Soc Inf Sci. 1993;44:222–34.
54. 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. doi:10.1177/1479972309348655.
55. Brownstein CA, Brownstein JS, Williams III DS, Wicks P, Heywood JA. The power of social networking in medicine. Nat Biotechnol. 2009;27:888–90. doi:10.1038/nbt1009-888.
56. 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. doi:10.1111/j.1460-2466.2008.00383.x.
57. 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.
58. Aymé S, Kole A, Groft S. Empowerment of patients: lessons from the rare diseases community. Lancet. 2008;371(9629):2048–51.
59. 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.
60. Eysenbach G. Medicine 2.0: social networking, collaboration, participation, apomediation, and openness. J Med Internet Res. 2008;10:e22. doi:10.2196/jmir.1030.
61. 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.
62. McCray AT. Promoting health literacy. J AHIMA. 2005;12:152–63. doi:10.1197/jamia.M1687.
63. Siefert 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.
64. 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.
65. Fortner RS. Excommunication in the information society. Crit Stud Mass Commun. 1995;12:133–54. doi:10.1080/15295039509366928.
66. 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.
67. Ferguson T. e-patients: how they can help us heal health care. e-patients.net. 2007. http://e-patients.net. Accessed Aug 2011.
68. 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.
69. Sarasohn-Kahn J. The wisdom of patients: health care meets online social media. Oakland: California Healthcare Foundation; 2008.
70. Hibbard JH, Weeks EC. Consumerism in health care. Med Care. 1987;25:1019–32.

Chapter 7
Clinical Research in the Postgenomic Era

Stephane M. Meystre, Scott P. Narus, and Joyce A. Mitchell

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

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

S.M. Meystre, M.D., Ph.D. (✉) • S.P. Narus, Ph.D. • J.A. Mitchell, Ph.D.
Department of Biomedical Informatics, University of Utah,
26 South 2000 East, Salt Lake City, UT 84112-5750, USA
e-mail: stephane.meystre@hsc.utah.edu; scott.narus@hsc.utah.edu;
joyce.mitchell@hsc.utah.edu

R.L. Richesson, J.E. Andrews (eds.), *Clinical Research Informatics*, Health Informatics,
DOI 10.1007/978-1-84882-448-5_7, © Springer-Verlag London Limited 2012

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. 7.1). 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 interactome. 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

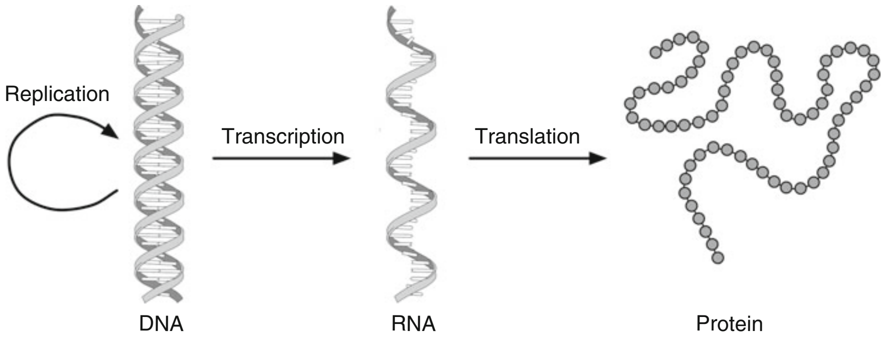


Fig. 7.1 Central dogma of molecular biology

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.

Molecular Biology and Genomics Data

Molecular biology produces vast amounts of data. Currently, more than a 1,000 public molecular biology databases are available. Prominent examples and their Web addresses are listed in Table 7.1.

Table 7.1 Selected molecular biology databases

Nucleotide sequence databases	
GenBank	www.ncbi.nlm.nih.gov/Genbank
EMBL-Bank (European Molecular Biology Laboratory)	www.ebi.ac.uk/embl
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	
Amino acid sequence and proteomics databases	
UniProt (Universal Protein Resource)	www.uniprot.org
PDB (Protein Data Bank)	www.rcsb.org/pdb
PRIDE (PRoteomics IDEntifications database)	www.ebi.ac.uk/pride
SGKB (Structural Genomics KnowledgeBase)	kb.psi-structuralgenomics.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.	
Genes databases	
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	
Gene and protein functional databases	
OPHID (Online Predicted Human Interaction Database)	ophid.utoronto.ca
GEO (Gene Expression Omnibus)	www.ncbi.nlm.nih.gov/geo/

(continued)

Table 7.1 (continued)

Gene and protein functional databases	
ArrayExpress	www.ebi.ac.uk/microarray-as/ae/
HMDB (Human Metabolome Database)	www.hmdb.ca
MMCD (Madison Metabolomics Consortium Database)	mmcd.nmrfam.wisc.edu
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"	
Databases combining different types of molecular biology data	
Entrez cross-database search	www.ncbi.nlm.nih.gov/sites/gquery
HGPD (Human Gene and Protein Database)	www.hgpd.jp
KEGG (Kyoto Encyclopedia of Genes and Genomes)	www.genome.jp/kegg
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	

The flood of data (one RNA analysis, for example, can produce an uncompressed image of more than 2,000 MB) and its nature require 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 [4]. 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, the 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 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 sequence annotation, for the exchange of annotation data, and for the description of 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 2009, it contained data on over 106 billion nucleotide pairs and about 108 million sequence records [9] 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 Database 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 Structures Classification [10] and is represented in XML format as PDBML [11]. The data can be analyzed with the aid of viewers that create three-dimensional representations of the proteins. Good examples are RasMol [12] and PyMOL [13]. The Structural Genomics Knowledgebase (SGKB) 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.

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 over 12,000 associated genes. OMIM is linked to NCBI's Entrez Gene [14], which contains information on about 45,000 human genes or loci (i.e., fixed positions on a chromosome that may or may not be occupied by one or more genes). 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 [15], the classification of the Human Genome Organization Gene Nomenclature Committee (HGNC) [16], and a growing body of information about epigenetic factors (factors that modify genes without changing their DNA sequence) [17] 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 [18], GeneCard [19], 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 Microarray and Gene Expression Data (MGED) Society has defined the minimum information needed to document a DNA microarray experiment (Minimal Information About a Microarray Experiment, or MIAME) [20] 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 [21]. Gene and protein expression results are stored in a MIAME-compliant format in public repositories such as the Gene Expression Omnibus at NCBI [22] and the Array Express 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 [23]. 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 [24] 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 [25], 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 [26] and the Madison Metabolomics Consortium Database [27] 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. The most common source of genetic differences 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 is replaced, respectively, by a nucleotide other than an A, T, C, or G. SNPs are the main reason that people differ in their susceptibility to common diseases. The International HapMap Project [28] 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. As more haplotypes are studied, the database, dbSNP [29], will more accurately reflect the specific types and extent of human variation.

While many variations are associated with health problems, many other variations are advantageous and many are neutral. Genome-wide association studies (GWAS) consider the statistical association between specific genome variations from the HapMap and human health conditions and analyze specific chromosome regions or whole genomes for those health-associated sites. Since 2002, when the HapMap project began, the list of human health conditions associated with patterns of genetic variation has grown rapidly [30].

Structural variants are another source of genetic variation among humans. They include sequence inversions, insertions, deletions, copy number variations, and complex rearrangements. The 1,000 Genomes Project [31] (which is actually sequencing 2,000 genomes) is investigating structural variants as well as SNPs in human population samples from Europe, Africa, East Asia, and the Americas. The whole genome of three individuals have been recently sequenced: a Caucasian man of European descent from the HapMap project, a Yoruban woman from HapMap whose genome is also being sequenced as part of the 1,000 Genomes Project, and a Caucasian man from the Personal Genome Project [32]. The team found around three million SNPs in each Caucasian genome and four million SNPs in the African genome, and 10% of the former were new, while 19% of the latter had not been identified in past studies [33].

A catalog of GWAS studies and their disease-gene associations is maintained by the National Human Genome Research Institute [34]. As of December 2011, it listed 1106 publications and 5481 SNPs. A 2009 review of the HapMap project and GWAS indicated that over 150 risk loci had been associated with over 60 common diseases and traits [35]. Those findings suggested potential new therapies based on molecular genetic information.

Many projects arise from an interest in applying genetic data clinically and the need to keep track of the variations shown to be associated with health problems. The Human Gene Mutation Database maintains a catalog of germline (sperm and egg) mutations in nuclear genes that are associated with human inherited diseases, and new entries are accruing at the rate of 9,000 per year [36]. Somatic mutations are covered by the COSMIC system, which is especially relevant for cancer [37], and mitochondrial mutations are covered by the MITOMAP database [38].

The Human Variome Project is an overarching initiative focused on collecting and curating all human genetic variation affecting human health [39]. It is considered the successor to the Human Genome Project [40] and to the HapMap project. The Human Variome Project is creating a full catalog of the genome sequence and variations in the human species, and will focus on the development of standards associated with the use of genetic information in the health care and clinical research communities.

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, a somatic genetic disease, is caused by a series of mutations in a single cell that provide that 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) 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). At the genomic level, the Cancer Genome Project [41] 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 [42], and at the proteomics level, the Clinical Proteomics Program of the National Cancer Institute and the US Food and Drug Administration [43] is searching for and characterizing new circulating cancer biomarkers.

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-Sanchez and colleagues [44] and demonstrated in examples cited below.

Integration of Molecular and Clinical Data

Researchers have made significant advances in the use of genomic data to describe the genetic makeup of organisms and are investigating how genes are expressed under various conditions. As mentioned earlier, however, whether a gene is turned on under a given set of conditions varies between individuals – even if they have the identical gene – and expression of that identical gene may manifest different physical or behavioral characteristics in different people [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. New informatics tools and techniques are being employed to address the growing need for integration between molecular and clinical data. Some prominent examples are presented below.

A National Center for Biomedical Computing research initiative based at Brigham and Women's Hospital (Boston, MA) called "i2b2" (Informatics for Integrating Biology and the Bedside) [49] is seeking to "build an informatics framework that will 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 is developing 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 evaluated at the University of Utah [50] and is used by research groups to study the genetic mechanisms underlying the pathogenesis of Huntington disease [51] or predict the response to bronchodilators in asthma patients [52].

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 [53]. caBIG's goal is to develop a collaborative information infrastructure that links data and analytic resources within and across institutions connected to the cancer grid (caGrid [54]). caBIG resources proposed or currently available to researchers 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 [55].

To store and search the vast amounts of knowledge generated by genotype-phenotype research, new databases and ontologies are being developed or enhanced. These resources set standards for how genomic and phenotypic concepts are named and defined, how they are associated, and how new knowledge can be modeled, shared, and stored. The PhenoGO database, for example, contains gene-disease annotations that were derived from the literature using several Gene Ontology annotation databases, the Unified Medical Language System, and other specialized ontologies [56]. The Unified Medical Language System Metathesaurus and the National Cancer Institute Thesaurus have been used to map annotation fields within genomic databases to standard concepts in order to integrate data for translational research [48]. The Unified Medical Language System has also been used to map textual annotations across microarray studies in order to join similar phenotypes and automatically construct disease classes [57]. Many ontological resources used in biological 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 provide standard application programming interfaces for more robust access to the underlying terminologies and their concept associations [58].

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 [59]. 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 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 had been developed separately for genomic and clinical data, the HL7 CG SIG's goal was to associate personal genomic data and clinical data. A data storage message encapsulates all the raw genomic data as static HL7 information objects. As this stored information is accessed for clinical care or research purposes, a data access/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 [60]. In parallel to the HL7 CG SIG, the Clinical Data Interchange Standards Consortium was formed in order to develop data standards that enable interoperability of medical research systems [61]. 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 [62]. As of this writing, neither the Clinical Data Interchange Standards Consortium 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.

Integration of Molecular Data into Clinical Research

When genomic data is incorporated into clinical trials, patient selection can be refined so that 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 ("ultrarapid metabolizers") [63]. Thus, selecting patients according to their metabolizing genotype when evaluating drug effects yields more useful information.

Molecular data can also be 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; markers for breast cancer treatment is one example [64]. 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 [65]. 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 [66]. A large amount of the research conducted on the mechanisms of diseases is nonclinical in nature.

Diagnostic Methods and Therapeutic Applications Studies

Single-gene tests are being developed at a very rapid pace and are the bellwether of postgenomic diagnostic development. The GeneTests system [67] for describing disease-gene relationships and available genetic tests, now hosted at the NCBI, was started in the mid-1990s when there were only a handful of DNA-based tests for inherited diseases; in 2009, it listed almost 1900.

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 [68] as well as leukemias [69], bladder cancer [70], sarcomas [71], head and neck cancers [72], kidney cancers [73], ovarian cancers [74], neuroblastoma [75, 76], and melanoma [77]. Many clinical trials involve therapeutic interventions. In breast cancer in particular, several commercial genomic assays for outcome prediction such as the MammaPrint are available [78]. The ongoing TAILORx and MINDACT clinical trials concentrate on outcomes [79]. Molecular therapies for lymphomas are also undergoing clinical testing [80].

Transplantation is another active research area. Heart transplant studies and microarray-based biomarker signatures have been ongoing for a decade, resulting in the CARGO clinical trials using an 11-gene signature called the Allomap genes [81]. 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 [82]. New studies of other organ transplants indicate a similar promise of monitoring the risk of transplant rejection [83].

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 [84, 85]. 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 [86].

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 [87]. 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 who access biomedical information on the World Wide Web, contribute their own health data through online tools such as personal health records, and directly access genetic testing resources, poses both challenges and opportunities for clinical investigators. Today, consumers can send a saliva or cheek swab sample to companies such as 23andMe [88], Navigenetics [89], and deCODEme [90] 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 Web site [91] 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 can tap these sites for highly motivated, well-informed study subjects; they may also able to gain access to the clinical and genetic data and use them to design new research projects.

Tailoring therapy to a specific subject, referred to as "personalized medicine" or "genomic medicine," offers new challenges for both clinical and informatics investigators. For the informatics investigator, linking genomic, phenotypic, and therapeutic data with outcomes presents computational and analytic challenges. Those four dimensions of data demand ever more precise data collection and analysis if meaningful and accurate associations are to develop. For the clinical investigator, personalized medicine research could improve therapy prediction, planning, and monitoring. The studies would be informed by subjects' genomic and phenotypic data with the goal being to increase knowledge about targeted treatments for individuals.

Molecular Epidemiological Data

Molecular epidemiology is the study of how genetic and environmental risk factors, at the molecular level, relate 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 [92].

Genealogical data allows for the study of the familiality 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 hereditability of many diseases, including cancers – some before the genetics was established [93]. 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 [94]. Another demonstrated linkage of rheumatoid arthritis with several chromosomes [95].

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 [96]. 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 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 [97, 98]. Specific particles can interact with tumors of a specific genotype. The future will undoubtedly involve individualized nanoparticle therapy.

References

1. Collins FS, Morgan M, Patrinos A. The human genome project: lessons from large-scale biology. Science. 2003;300:286–90.
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;94:336–42.

4. NCBI. Just the facts: a basic introduction to the science underlying NCBI resources – bioinformatics. Available from: http://www.ncbi.nlm.nih.gov/About/primer/bioinformatics.html. Accessed Aug 2011.
5. Luscombe NM, Greenbaum D, Gerstein M. What is bioinformatics? A proposed definition and overview of the field. Methods Inf Med. 2001;40:346–58.
6. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. GenBank. Nucleic Acids Res. 2008;36:D25–30.
7. Eilbeck K, Lewis SE, Mungall CJ, Yandell M, Stein L, Durbin R, et al. The sequence ontology: a tool for the unification of genome annotations. Genome Biol. 2005;6:R44.
8. Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, et al. The OBO foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol. 2007;25:1251–5.
9. NCBI. GenBank. 2009. Accessed Aug 2011.
10. Cuff AL, Sillitoe I, Lewis T, Redfern OC, Garratt R, Thornton J, et al. The CATH classification revisited – architectures reviewed and new ways to characterize structural divergence in superfamilies. Nucleic Acids Res. 2009;37:D310–4.
11. Westbrook J, Ito N, Nakamura H, Henrick K, Berman HM. PDBML: the representation of archival macromolecular structure data in XML. Bioinformatics. 2005;21:988–92.
12. RasMol. 2009. Accessed Aug 2011.
13. PyMOL. 2009. Accessed Aug 2011.
14. Maglott D, Ostell J, Pruitt KD, Tatusova T. Entrez gene: gene-centered information at NCBI. Nucleic Acids Res. 2005;33:D54–8.
15. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000;25:25–9.
16. White JA, McAlpine PJ, Antonarakis S, Cann H, Eppig JT, Frazer K, et al. Guidelines for human gene nomenclature. HUGO Nomenclature Committee. Genomics. 1997;45:468–71.
17. Yoou MH. Case study of a patient with Parkinson's disease. Taehan Kanho. 1991;30:56–60.
18. Frezal J. Genatlas database, genes and development defects. C R Acad Sci III. 1998;321:805–17.
19. 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;14:656–64.
20. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, et al. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat Genet. 2001;29:365–71.
21. Courtot M, Bug W, Gibson F, Lister AL, Malone J, Schober D et al. The OWL of biomedical investigations. 2008. Available from: http://webont.com/owled/2008/papers/owled2008eu_submission_38.pdf. Accessed Aug 2011.
22. Edgar R, Domrachev M, Lash AE. Gene expression omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 2002;30:207–10.
23. Oh JE, Krapfenbauer K, Fountoulakis M, Frischer T, Lubec G. Evidence for the existence of hypothetical proteins in human bronchial epithelial, fibroblast, amnion, lymphocyte, mesothelial and kidney cell lines. Amino Acids. 2004;26:9–18.
24. Stoevesandt O, Taussig MJ, He M. Protein microarrays: high-throughput tools for proteomics. Expert Rev Proteomics. 2009;6:145–57.
25. Natale DA, Arighi CN, Barker WC, Blake J, Chang TC, Hu Z, et al. Framework for a protein ontology. BMC Bioinformatics. 2007;8:S1.
26. Wishart DS, Tzur D, Knox C, Eisner R, Guo AC, Young N, et al. HMDB: the human metabolome database. Nucleic Acids Res. 2007;35:D521–6.
27. Cui Q, Lewis IA, Hegeman AD, Anderson ME, Li J, Schulte CF, et al. Metabolite identification via the Madison Metabolomics Consortium Database. Nat Biotechnol. 2008;26:162–4.
28. International HapMap Consortium. The international HapMap project. Nature. 2003;426:789–96.
29. NCBI. dbSNP. Available from: www.ncbi.nlm.nih.gov/projects/SNP/. Accessed Aug 2011.
30. Manolio TA, Brooks LD, Collins FS. A HapMap harvest of insights into the genetics of common disease. J Clin Invest. 2008;118(5):1590–605.

31. Kaiser J. DNA sequencing. A plan to capture human diversity in 1000 genomes. Science. 2008;319:395.
32. Personal genome project. 2009. Accessed Aug 2011.
33. Drmanac R, Sparks AB, Callow MJ, Halpern AL, Burns NL, Kermani BG, et al. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. Science. 2009;327:78–81.
34. NHGRI. A catalog of published genome-wide association studies. 2009. Accessed Aug 2011.
35. Manolio TA, Collins FS. The HapMap and genome-wide association studies in diagnosis and therapy. Annu Rev Med. 2009;60:443–56.
36. Stenson PD, Mort M, Ball EV, Howells K, Phillips AD, Thomas NS, et al. The Human Gene Mutation Database: 2008 update. Genome Med. 2009;1:13.
37. Bamford S, Dawson E, Forbes S, Clements J, Pettett R, Dogan A, et al. The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. Br J Cancer. 2004;91:355–8.
38. MITOMAP: a human mitochondrial genome database. 2009 [cited 2009]. Accessed Aug 2011.
39. The human variome project. 2009. Accessed Aug 2011.
40. Cotton RG, Auerbach AD, Axton M, Barash CI, Berkovic SF, Brookes AJ, et al. Genetics. The human variome project. Science. 2008;322:861–2.
41. Institute WTS. Cancer genome project. Available from: http://www.sanger.ac.uk/genetics/CGP. Accessed Aug 2011.
42. NCI. Cancer genome anatomy project. 2009. Accessed Aug 2011.
43. NCI. FDA-NCI clinical proteomics program. Available from: http://home.ccr.cancer.gov/ncifdaproteomics/default.asp. Accessed Aug 2011.
44. Martin-Sanchez F, Iakovidis I, Norager S, Maojo V, de Groen P, Van der Lei J, et al. Synergy between medical informatics and bioinformatics: facilitating genomic medicine for future health care. J Biomed Inform. 2004;37:30–42.
45. Butte AJ, Kohane IS. Creation and implications of a phenome-genome network. Nat Biotechnol. 2006;24:55–62.
46. Chen DP, Weber SC, Constantinou PS, Ferris TA, Lowe HJ, Butte AJ. Clinical arrays of laboratory measures, or “clinarrays”, built from an electronic health record enable disease subtyping by severity. AMIA Annu Symp Proc. 2007;2007:115–9.
47. Butte AJ, Chen R. Finding disease-related genomic experiments within an international repository: first steps in translational bioinformatics. AMIA Annu Symp Proc. 2006;2006:106–10.
48. Shah NH, Jonquet C, Chiang AP, Butte AJ, Chen R, Musen MA. Ontology-driven indexing of public datasets for translational bioinformatics. BMC Bioinformatics. 2009;10:S1.
49. Murphy SN, Mendis ME, Berkowitz DA, Kohane I, Chueh HC. Integration of clinical and genetic data in the i2b2 architecture. AMIA Annu Symp Proc. 2006;2006:1040.
50. Deshmukh VG, Meystre SM, Mitchell JA. Evaluating the informatics for integrating biology and the bedside system for clinical research. BMC Med Res Methodol. 2009;9:70.
51. Lee JM, Ivanova EV, Seong IS, Cashorali T, Kohane I, Gusella JF, et al. Unbiased gene expression analysis implicates the huntingtin polyglutamine tract in extra-mitochondrial energy metabolism. PLoS Genet. 2007;3:e135.
52. Himes BE, Wu AC, Duan QL, Klanderman B, Litonjua AA, Tantisira K, et al. Predicting response to short-acting bronchodilator medication using Bayesian networks. Pharmacogenomics. 2009;10:1393–412.
53. NCI. caBIG: cancer biomedical informatics grid. 2009. Accessed Aug 2011.
54. Saltz J, Oster S, Hastings S, Langella S, Kurc T, Sanchez W, et al. CaGrid: design and implementation of the core architecture of the cancer biomedical informatics grid. Bioinformatics. 2006;22:1910–6.
55. Amin W, Parwani AV, Schmandt L, Mohanty SK, Farhat G, Pople AK, et al. National Mesothelioma Virtual Bank: a standard based biospecimen and clinical data resource to enhance translational research. BMC Cancer. 2008;8:236.
56. Sam LT, Mendonca EA, Li J, Blake J, Friedman C, Lussier YA. PhenoGO: an integrated resource for the multiscale mining of clinical and biological data. BMC Bioinformatics. 2009;10:S8.

57. Liu CC, Hu J, Kalakrishnan M, Huang H, Zhou XJ. Integrative disease classification based on cross-platform microarray data. BMC Bioinformatics. 2009;10:S25.
58. Pathak J, Solbrig HR, Buntrock JD, Johnson TM, Chute CG. LexGrid: a framework for representing, storing, and querying biomedical terminologies from simple to sublime. J Am Med Inform Assoc. 2009;16:305–15.
59. Hewett M, Oliver DE, Rubin DL, Easton KL, Stuart JM, Altman RB, et al. PharmGKB: the Pharmacogenetics Knowledge Base. Nucleic Acids Res. 2002;30:163–5.
60. Shabo A. The implications of electronic health record for personalized medicine. Biomed Pap Med Fac Univ Palacky Olomouc Czech Repub. 2005;149(2):251–8.
61. CDISC. Clinical data interchange standards consortium (CDISC). 2009. Available from: http://www.cdisc.org/. Accessed Aug 2011.
62. BRIDG. Biomedical Research Integrated Domain Group (BRIDG). 2009. Accessed Aug 2011.
63. Schenk PW, van Fessem MA, Verploegh-Van Rij S, Mathot RA, van Gelder T, Vulto AG, 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;13:597–605.
64. Loi S, Buyse M, Sotiriou C, Cardoso F. Challenges in breast cancer clinical trial design in the postgenomic era. Curr Opin Oncol. 2004;16:536–41.
65. Vogel CL, Cobleigh MA, Tripathy D, Gutheil JC, Harris LN, Fehrenbacher L, 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;20:719–26.
66. Jahromi MM, Eisenbarth GS. Cellular and molecular pathogenesis of type 1A diabetes. Cell Mol Life Sci. 2007;64:865–72.
67. Waggoner DJ, Pagon RA. Internet resources in Medical Genetics. Curr Protoc Hum Genet. 2009;7: Chapter 9:Unit 9.12.
68. Leich E, Hartmann EM, Burek C, Ott G, Rosenwald A. Diagnostic and prognostic significance of gene expression profiling in lymphomas. APMIS. 2007;115:1135–46.
69. Codony C, Crespo M, Abrisqueta P, Montserrat E, Bosch F. Gene expression profiling in chronic lymphocytic leukaemia. Best Pract Res Clin Haematol. 2009;22:211–22.
70. Chan KS, Espinosa I, Chao M, Wong D, Ailles L, Diehn M, et al. Identification, molecular characterization, clinical prognosis, and therapeutic targeting of human bladder tumor-initiating cells. Proc Natl Acad Sci USA. 2009;106:14016–21.
71. Hoffman AC, Danenberg KD, Taubert H, Danenberg PV, Wuerl P. A three-gene signature for outcome in soft tissue sarcoma. Clin Cancer Res. 2009;15:5191–8.
72. Gold KA, Kim ES. Role of molecular markers and gene profiling in head and neck cancers. Curr Opin Oncol. 2009;21:206–11.
73. Petillo D, Kort EJ, Anema J, Furge KA, Yang XJ, Teh BT. MicroRNA profiling of human kidney cancer subtypes. Int J Oncol. 2009;35:109–14.
74. Yoshihara K, Tajima A, Komata D, Yamamoto T, Kodama S, Fujiwara H, 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;100:1421–8.
75. Volchenboum SL, Cohn SL. Are molecular neuroblastoma classifiers ready for prime time? Lancet Oncol. 2009;10:641–2.
76. Vermeulen J, De Preter K, Naranjo A, Vercruysse L, Van Roy N, Hellemans J, 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.
77. Ugurel S, Utikal J, Becker JC. Tumor biomarkers in melanoma. Cancer Control. 2009;16(3):219–24.
78. 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;133:855–9.
79. Sotiriou C, Pusztai L. Gene-expression signatures in breast cancer. N Engl J Med. 2009;360:790–800.

80. Rabson AB, Weissmann D. From microarray to bedside: targeting NF-kappaB for therapy of lymphomas. Clin Cancer Res. 2005;11:2–6.
81. Fang KC. Clinical utilities of peripheral blood gene expression profiling in the management of cardiac transplant patients. J Immunotoxicol. 2007;4:209–17.
82. XDx. XDx's AlloMap(R) gene expression test cleared by U.S. FDA for heart transplant recipients. 2008 [cited 2009]. Accessed Aug 2011.
83. Khatri P, Sarwal MM. Using gene arrays in diagnosis of rejection. Curr Opin Organ Transplant. 2009;14:34–9.
84. van Baarsen LG, Bos CL, van der Pouw Kraan TC, Verweij CL. Transcription profiling of rheumatic diseases. Arthritis Res Ther. 2009;11:207.
85. Bauer JW, Bilgic H, Baechler EC. Gene-expression profiling in rheumatic disease: tools and therapeutic potential. Nat Rev Rheumatol. 2009;5:257–65.
86. Lin B, Malanoski AP. Resequencing arrays for diagnostics of respiratory pathogens. Methods Mol Biol. 2009;529:231–57.
87. Roche. Individualize drug dosing based on metabolic profiling with the AmpliChip CYP450 test. 2009 [cited 2009]. Accessed Aug 2011.
88. 23andMe. 23andMe: Genetics just got personal. 2009. Accessed Aug 2011.
89. Navigenics. There's DNA. And then there's what you do with it. 2009. Accessed Aug 2011.
90. deCODEme. deCODE your health. 2009. Accessed Aug 2011.
91. PatientsLikeMe. PatientsLikeMe: patients helping patients live better every day. Available from: http://www.patientslikeme.com/. Accessed Aug 2011.
92. Kaput J, Rodriguez RL. Nutritional genomics: the next frontier in the postgenomic era. Physiol Genomics. 2004;16:166–77.
93. Cannon-Albright LA, Thomas A, Goldgar DE, Gholami K, Rowe K, Jacobsen M, et al. Familiality of cancer in Utah. Cancer Res. 1994;54:2378–85.
94. Hamshere ML, Schulze TG, Schumacher J, Corvin A, Owen MJ, Jamra RA, 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;11:610–20.
95. Hamshere ML, Segurado R, Moskvina V, Nikolov I, Glaser B, Holmans PA. Large-scale linkage analysis of 1302 affected relative pairs with rheumatoid arthritis. BMC Proc. 2007;1:S100.
96. Bos JM, Towbin JA, Ackerman MJ. Diagnostic, prognostic, and therapeutic implications of genetic testing for hypertrophic cardiomyopathy. J Am Coll Cardiol. 2009;54:201–11.
97. de la Fuente M, Csaba N, Garcia-Fuentes M, Alonso MJ. Nanoparticles as protein and gene carriers to mucosal surfaces. Nanomedicine (Lond). 2008;3:845–57.
98. Emerich DF, Thanos CG. Targeted nanoparticle-based drug delivery and diagnosis. J Drug Target. 2007;15:163–83.

Part II
Data Management and Systems in Clinical Research

Chapter 8
Clinical Research Information Systems

Prakash M. Nadkarni, Luis N. Marenco, and Cynthia A. Brandt

Abstract Clinical Research Information Systems (CRISs) are a type of specialized software application which are designed to support clinical research. The use of CRISs can reduce the costs of research studies. Information systems can also support a host of functions and activities within clinical research enterprises. In this chapter, we look at the various CRIS vendor models, including new open-source systems. 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 (EMR) Systems. We also discuss the considerations involved in deciding whether to build, lease, or purchase a vendor model. Some significant quality-control issues are also highlighted, including double data entry (DDE) and random audits. We then describe the operations of a CRIS during different phases of a study, including determining patient recruitment and eligibility, protocol management, patient monitoring and safety, and analysis and reporting. Included here is role of informaticians in working with investigators and biostatisticians to make systematic requirements analysis and to provide costs estimates for informatics components of studies. We finally discuss briefly the issues of standards and certification and look at usability and user-centered designed as evaluative criteria for CRISs.

P.M. Nadkarni, M.D. (✉)
Center for Medical Informatics, Yale University School of Medicine,
300 George Street, Suite 501, New Haven, CT 06511, USA
e-mail: prakash.nadkarni@yale.edu

L.N. Marenco, M.D.
Yale Center for Medical Informatics, Yale University,
300 George Street, Suite 501, New Haven, CT 06511, USA
e-mail: luis.marenco@yale.edu

C.A. Brandt, M.D., M.P.H
Yale Center for Medical Informatics, Emergency Medicine, Yale University,
300 George Street, Suite 501, New Haven, CT 06511, USA

Department of Anesthesiology, VA Connecticut Healthcare System,
New Haven, CT, USA
e-mail: cynthia.brandt@yale.edu

R.L. Richesson, J.E. Andrews (eds.), *Clinical Research Informatics*, Health Informatics,
DOI 10.1007/978-1-84882-448-5_8, © Springer-Verlag London Limited 2012

Keywords Clinical research information systems • Evolution of CRISs • Electronic protocol • Electronic data capture • Electronic medical record • HIPAA • Study protocol • Double data entry

Clinical Research Information Systems (CRISs) are a category of software application specialized 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 Medical Record (EMR) Systems, they must typically interoperate with EMRs. Initially, many systems were developed to support 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 in an attempt to provide one-stop shopping to their customers. Nonetheless, despite the proliferation of commercial software, special problems still arise that only custom software development can solve. Protocol-authoring capabilities, while part of CRIS functionality, are discussed in the next chapter.

In this chapter, we provide the reader with a feel for the various issues and processes related to CRISs in order to be able to perform a systematic requirements analysis and to decide whether to build a system, lease, or purchase one. We will also emphasize practical issues in CRIS operation that have little to do with informatics *per se*, but which can be ignored only at one's peril. We will not provide an exhaustive survey of CRIS-related efforts; few such efforts have sustained themselves, and fewer still have resulted in software that has been made freely available.

Clinical Research Information Systems Vendor Models

The larger commercial CRISs require a financial investment almost as formidable as that for an institutional EMR. The size of this outlay, coupled with the software's complexity, virtually mandates their deployment be institution wide. In keeping with their scope, such systems can manage an arbitrary number of clinical studies within a single physical database design, subject only to hardware limits. These are sold outright to the customer, installed at a customer site, and operated by customer personnel.

In addition to expense, such systems require a team of individuals with diverse skills to operate and maintain – database administrators, software developers, and nontechnical individuals familiar with clinical research as well as the CRIS software, which can translate an investigator's study design into an electronic representation. This team's salaries can only be amortized through support of multiple studies, and this is typically feasible only at an institution level.

Institution-level information technology projects have high risk and require organizational commitment. Many customers with smaller budgets, for example, individual departments, may be uncertain about getting such commitment, and also lack the budgets to hire or retain skilled informatics support staff. Therefore, certain

CRIS vendors follow a different approach, installing the software at either the customer or vendor's site (the latter is an option only if Institutional Review Board concerns about data privacy and security can be addressed), but the vendor staff generally perform all administrative and software-development functions remotely. The customer is billed based on factors such as 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 and potentially higher ongoing costs; often, though, the vendor can perform the administrative/developer tasks much more cheaply than the customer: 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, adopt, 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 the 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 capabilities of different systems vary widely, and certain systems support only simple study designs, such as surveys.

Why Have Clinical Research Information Systems Evolved?

EMRs are not entirely suitable for supporting clinical research needs by themselves, mostly due to fundamental differences between clinical research and patient-care processes. We describe these differences below, while emphasizing that workflows involve interoperability with EMR-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 study 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 Clinical Research Information Systems

CRISs differ from EMRs 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.

Regardless of any other functions, a CRIS must provide two essential functions: representing a protocol electronically and supporting electronic data capture. EMRs are not designed for the former objective. Individual CRIS offerings differ in how fully they can model a variety of protocols, and the sophistication of their data capture tools. (Chap. 9 describes issues related to protocol representation in greater detail.)

Clinical Research Information Systems 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. Unlike in the EMR setting, where a patient can be seen by almost any healthcare provider in the organization, 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 coinvestigator in another; therefore, certain administrative-type privileges may be denied in the latter study.

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 placebos) 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, for example, 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; EMR software lacks this capability.

The Scope of a Clinical Research Information System 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, putatively "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 unnecessary PHI such as patient addresses (unless 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 EMRs do not address.

For transinstitutional studies, CRIS solutions must increasingly use Web technology to provide access across individual institutional firewalls. By contrast, EMRs even when used in a geographically distributed setting (as for a network of community-based physicians) are still institutional in scope. Therefore, EMR 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 login (through mechanisms such as Citrix).

When a multisite study is conducted across countries with different languages, the informatics challenges can be significant, as well-described in [2]. Besides challenges in the coordination of the studies, 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 which person has logged in. This feature, called *dynamic localization*, is possible to implement with relatively modest programming effort using Web-based technologies such as Microsoft ASP.NET. The approach relies on *resource files* containing text-string elements of the user interface for each language of use, which are consulted by a user interface where the elements are defined symbolically rather than hard-coded (so that their actual definitions are pulled from a resource file at run time). While several commercial websites implement this capability, to the best of our knowledge no existing commercial CRIS has employed it as of yet.

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

In EMR-supported processes involving patient care, the Joint Commission on Accreditation of Healthcare Organizations (JCAHO) recommends the use of *at least* two personal identifiers [1, 3] to ensure that errors due to treatment of the wrong patient are minimized. In contrast, in certain multisite clinical studies that involve minimal risk to the patient (such as purely observational studies), Institutional Review Boards (IRBs) 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.

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 (e.g., digit-substituted or digit-transposed) Study ID from being accepted. There is often a very real fear of allowing PHI to be entered, even in multisite studies where physical injury (e.g., dose escalations of a toxic drug) would result from decisions accidentally made for the wrong patient. In such circumstances, IRBs need to be gently educated that in their zeal to prevent patient harm due to PHI disclosure, they risk much greater clinical harm.

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

Most research studies are conducted in ambulatory (outpatient) settings simply because most conditions of research interest do not mandate the expense of continuous subject monitoring through admission to a hospital or research center. (Even in major illnesses like cancer, patients are typically in a hospital only for the short duration of chemotherapy.) 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 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 for medical emergencies.

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 Previsit Reminder Event will drive a workflow related to mailing of form-letter reminders. Thus, the Subject Calendar is really a *Calendar of Events*.

Time Windows Associated with Events

One should note that, in order to allow for subjects' convenience, and due to the fact that certain scheduled days may fall on weekends or public holidays, the subject calendar dates in most intermediate- to long-term studies have some built-in slack, and the permissible slack, or *window,* is also predetermined by the investigator/s on a per-event basis. Thus, the event that corresponds to the 1-year follow-up may be allowed to occur between 11 and 13 months. This window, of course, would vary by the protocol and type of study. For example, in a natural history/observational study, the windows might be as broad as 6 months on each end of the event due date, while in a pharmacokinetic study of a fast acting drug, the acceptable window might be measured in minutes.

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

The Event-Case Report Form Cross-Table

At each event, specific actions are performed – for example, administration of therapy – and units of information gathered in individual CRFs, for example, for questionnaires, physical examinations, laboratory tests, and special investigations. The association of individual events with individual case report forms is called the "Event-CRF Cross-Table." For reasons of expense as well as risk to the patient from invasive tests, all investigations are not carried out at all events, or with equal frequency: expensive and/or tests posing risk of physical harm are much fewer than cheaper or routine tests.

In such experimental designs, the supporting software must typically *enforce the Study's Event-CRF cross-table constraints*. That is, it should not be possible for a member of the research team, when entering data in real time or off-line, to accidentally create a CRF for an event where, according to the cross-table, it does not apply. Cross-table constraint enforcement ensures that the values of individual parameters gathered at a specific time point on the Study Calendar for individual patients can be pooled together into summary statistics (and compared across groups of patients, where the study design uses more than one group) because these values reflect the state of that cohort or subcohort at that point in time.

The CRIS should also provide 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) can be planned. The CRIS should ideally also support provision of *advance reminders* to subjects either through form letters, phone messages, or email. (Reminders are one feature that today's EMRs support very well: missed office visits translate into lost revenue because scheduled services reserved for a given patient are not utilized.) Timely alerts about *missed visits* are particularly critical because even if a subject is persuaded to show up for an appointment later, 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 EMR supports:

- EMRs 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 care giving, such as performing cognitive tasks or responding to standard questionnaires.
- In most studies, a large number of potential subjects are screened for recruitment into the study. Many of the screened individuals who show initial interest in participation may, on screening via a questionnaire, fail to meet the study's eligibility criteria. But even among those who are eligible, it often takes persistent persuasion to secure their participation. The process may take several encounters (phone calls or personal interviews) after which certain candidate subjects may ultimately decline once participation risks are explained. All the while, the CRIS must record contact information about potential subjects and log all encounters, if only to keep a work record for the recruiting staff who are paid for their efforts.
- In research studying genetic influences on particular diseases, one type of study design involves study of large groups of subjects who are related to each other through marriage and common ancestors (i.e., *pedigrees* of individuals). In such situations, to increase the power of the ultimate data analysis, one may include "pseudo-subjects": ancestral individuals (e.g., great grandparents) who connect smaller families even though they are long deceased and almost nothing is known about them.

Clinical Research Information Systems Often Need to Support Real-Time Self-reporting of Subject Data

To ease the research staff's data entry burden, some CRISs may support self-entry by subjects or can accept data via bulk import from external systems that support such self-entry. Self-entry is appropriate for certain CRFs in studies involving self-rating. For example, pain intensity is typically self-reported on an ordinal or analog scale, and certain instruments, such as the Center for Epidemiological Studies Depression Scale (CES-D) [4], have been used to self-assess intensity of depressive symptoms following radiotherapy for head-neck cancers.

Many subjects are more than capable of using Web-based computer applications for work or personal purposes, so it is reasonable 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 have mandate a visit or have a staff member interview them over the phone. CRISs that support self-entry by subjects allow informatics staff to provide a limited login 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. To enhance data completion and quality, the use of good interface design and online data collection features is particularly important in patient-directed data collection applications.

Clinical Research Data Capture Is More Structured Than in Patient Care

In clinical care, a patient may present with any disease; even in most clinical specialties, a broad range of conditions are possible. The only 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 undergoing a specific protocol where the required data elements are known in advance, for example, for coronary bypass, cataract surgery, or when partial structure can be imposed (e.g., for a chest X-ray examination).

While narrative text is very flexible, it is extremely challenging to analyze because of issues such as medical term synonymy and the telegraphic, often non-grammatical nature of the notes. By contrast, in most clinical research, the patients are preselected for a specific clinical condition or conditions, so one knows in advance exactly what data elements will be captured. Therefore, CRFs are highly structured, maximizing the use of elements that require numeric or discrete responses (e.g., yes/no responses or values selected from a list of choices).

Occasionally, in studies that have dual objectives – that is, research combined with clinical care – such forms will occasionally contain narrative-text elements like “Additional Comments,” “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. These provide an opportunity to revise the CRF by increasing its structure through specific prompts for such responses. Apart from making the CRF faster to fill (entering narrative text is always slower than clicking check boxes or selecting items from lists), they improve the data’s subsequent analyzability by making more of the questions discrete-valued.

Clinical Research Information Systems Electronic Data Capture Needs to Be Robust and Flexible and Efficient to Set Up

Data capture in many research settings (notably psychiatry/psychology) is typically far more extensive than in EMRs. Numerous questionnaires have been designed specifically for research problems and are too lengthy for convenient use by busy caregivers or by patients who are not compensated for their time in a research study. Because many CRFs are so lengthy, there is a greater risk of the data capture process introducing inconsistency. Consequently, CRISs must provide extensive support for real-time data validation:

- 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, 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 – for example, 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, for example, 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 in addition to being preventive in nature, streamline the data entry process.

- 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 yes to a question about past history of cardiovascular disease) causes certain subsequent questions which would ask for more details about this condition to be disabled or to become invisible if the user responds to the initial question with a "no" because they are now inapplicable.
- *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 subindication. This feature, typically implemented using Web-based technologies such as Asynchronous JavaScript over XML (AJAX) [5], 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.

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 [6], even the questions themselves are not fixed: depending on how the subject has responded to previous questions, different new questions will appear.

While EMRs increasingly allow sophisticated data capture, it would be safe to say that CRISs have defined the state of the art in this regard. Note that a given CRIS may not support every possible feature: the requirement for adaptive designs, for example, has resulted in such systems being developed from first principles, as in the PROMIS consortium [7]. Also, certain experimental designs, as described in [8, 9], 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 here too. 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 forms may be editable only by those responsible for creating their data.

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 reused. 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.

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

EMRs capture patient-encounter data in real time or near real time: CRISs are more adaptable to individual needs, supporting off-line 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. The chapter by Nahm deals with the issue of quality control in greater detail: we will highlight a few significant issues here which that chapter does not address.

To ensure highest quality in terms of minimal missing or unusable data – a major issue in clinical research [10] – off-line transcription should be as little delayed from the original encounter as possible – for example, not more than 4 days later. Missing data can occur because source documents can be misplaced or damaged. Bad data-element values are much more likely with paper source documents than with electronic CRFs that support robust interactive validation. Bad-data errors can be corrected only by querying the source document's human originator, and only if the operator remembers the encounter, which is likely only if the encounter is very recent.

Double data entry (DDE) is a quality-control method based on the principle of comparing identical input created by two different operators who transcribe the same document separately: input that matches exactly is likely to be correct (unless both operators made the same mistake). Originating during the punched-card era, DDE, in our opinion, has outlived its usefulness. In a seminal article, Day et al. [11] point out that DDE is neither necessary nor sufficient for good data quality: it does not catch bad-source-data errors.

Today, best 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 is based on criteria such as the criticality, 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 multisite 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 W. Edwards Deming's (a leading scholar in QC) approach, allows concentration of limited resources in the areas of most potential benefit, as opposed to DDE, which indiscriminately weights every single question on every single CRF equally.

Because timeliness of data entry is so important, 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.

Clinical Research Information System-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 Study Planning and Protocol Authoring, Recruitment/Eligibility

Determination (screening), Protocol Management and Study Conduct (including patient-monitoring and safety), and Analysis and Reporting. Most of the foregoing text has dealt with issues relating to stage 3, though CRF design and Calendar setup are part of stage 1.

Study Planning and Protocol Authoring

While clinical investigators are ultimately responsible for the overall study plan, it is far more productive for the study plan to 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 [12–14], 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 determinations 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. *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 EMRs. An informatics support team should have all these issues worked out in advance. In particular, informaticians must 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 is understandably expensive, but even if custom development 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 clinical investigators may be naive and 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 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 and 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, they can create the documentation in formats such as HTML (and automatically generating a searchable website) 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, for example [15], while others, such as the Cancer Center Participant Registry [16], 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 EMR. 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 EMR 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 EMR vendor's database schema. The issues with integrating CRIS and EMR data are described in more detail in Chapter 17.

When a subject agrees to participate in the study, he or she 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 EMR'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. 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, for example, 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 events (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) [17]. Here, the gradation of each concept is specified unambiguously, typically on a five-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, helps prevent 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 discrimination.

In cases where the study is being conducted in a hospital rather than a clinical research setting, effective interoperability between the EMR 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 EMR 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 [18]. In hospital settings, AEs are also recorded in the narrative text of progress notes. Processing these is much more challenging, but Wang et al. [19] describe an approach for pharmacovigilance based on narrative EMR 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 were terminated from the trial abnormally – for example, 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 CRIS must perform extensive transformation on their data. Also, in practically all cases, the data sets generated for statistical analysis must be *de-identified*, that is, the subjects must be identified only by their machine-generated ID without any PHI being 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, once priced at around $3 million, which 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). As many EMR 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 inelegant, 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" [20] emphasizes, the absence of detected errors detected 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 fully meets all its requirements, it may not be defect-free in actual operation. CRFs, which end-users interact with, are developed by the CRIS support team within an organization. CRF design is essentially a kind of high-level programming, typically using a GUI metaphor (so that nonprogrammers can accomplish most tasks). Errors of both commission – for example, a mistake in a formula – or omission – for example, forgetting to add sufficient validation checks, so that bad data creeps in – are possible. 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 simple computational formulas are represented as text strings that are specific to a particular programming language. Even simple computational formulas are represented as text strings that are specific to a particular programming language. Even if CDISC acquired such capabilities, 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.

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 held for an unacceptably long duration in the area of Enterprise Resource Planning software.)

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;5:75–84.
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;5:253–61.
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. Raloff LS. The CES-D scale: a self-report depression scale for research in the general population. Appl Psychol Meas. 1977;1:385–401.
5. Crane D, Pascarello E, James D. AJAX in action. Greenwich: Manning Publications Co.; 2005.
6. Wikipedia. Computerized adaptive testing. Available at: en.wikipedia.org/wiki/Computerized_adaptive_testing. Last accessed 12/1/09.
7. Cella 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:S3–11.
8. 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;4:572–82.
9. 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;18:23.
10. Wisniewski SR, Leon AC, Otto MW, Trivedi MH. Prevention of missing data in clinical research studies. Biol Psychiatry. 2006;59:997–1000.
11. Day S, Fayers P, Harvey D. Double data entry: what value, what price? Contemp Clin Trials. 1998;19:24.

12. 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 Programs Biomed. 1994;43:283–91.
13. Modgil S, Hammond P. Decision support tools for clinical trial design. Artif Intell Med. 2003;27:181–200.
14. Rubin DL, Gennari J, Musen MA. Knowledge: representation and tool support for critiquing clinical trial protocols. Proc AMIA Symp. 2000;2000:724–8.
15. Gennari JH, Sklar D, Silva J. Cross-tool communication: from protocol authoring to eligibility determination. Proc AMIA Symp. 2001;2001:199–203.
16. National Cancer Institute. Cancer Center Participant Registry. Information available at: https://cabig.nci.nih.gov/tools/c3pr. Last accessed 12/01/09.
17. National Cancer Institute. Common terminology criteria for adverse events (CTCAE) and common toxicity criteria (CTC). 2009. Available from: http://ctep.cancer.gov//electronic_applications/ctc.htm. Last accessed 12/1/09.
18. 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;31:807–22.
19. 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.
20. Kaner C, Falk J, Nguyen HQ. Testing computer software. 2nd ed. New York: Wiley; 1999.
21. Wikipedia. User-centered design. Available at http://en.wikipedia.org/wiki/user_centered_design. Last accessed 11/1/09.

Chapter 9
Study Protocol Representation

Ida Sim and Joyce C. Niland

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.

I. Sim, M.D., Ph.D. (✉)
Department of Medicine, University of California San Francisco,
1545 Divisadero Street, Suite 308, San Francisco, CA 94143-0320, USA
e-mail: ida.sim@ucsf.edu

J.C. Niland, Ph.D.
Division of Information Sciences, Beckman Research Institute, City of Hope,
500 East Duarte Road, Duarte, CA 91010, USA
e-mail: jniland@coh.org

R.L. Richesson, J.E. Andrews (eds.), *Clinical Research Informatics*, Health Informatics,
DOI 10.1007/978-1-84882-448-5_9, © Springer-Verlag London Limited 2012

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, or their biospecimens, in order 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 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*, not study protocols, in text. They file study protocol *documents* for human subjects approval applications, and sponsors sometimes post study protocol *documents* on the web. Study protocol documents are artifacts generated to guide the conduct of clinical research during the course of a study, and while they 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.

Clinical Research Informatics and the Study Protocol

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 clinical research informatics (CRI) tools, the focus has been on facilitating the electronic sharing of such

Table 9.1 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 *shared and standardized* 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")

study protocol documents. However, because study protocol documents are only derivatives of the abstract study protocol, this focus on document management is narrower and less powerful than direct information management of the abstract study protocol itself.

To advance CRI tools, the abstract study protocol must be made directly computable, without the intermediary of textual descriptions in the form of a protocol document. Rather, the convenience of sharable protocol documents should be generated from the creation of a computable study protocol. 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; 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]. 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 true information management in full-spectrum support of the design, execution, analysis, and reporting of clinical research.

Table 9.1 defines several key terms for 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 we review below. Standardization of the underlying computable protocol model into a *common computable protocol model* has been a “holy grail” of clinical research informatics for many years, to facilitate the interoperation of 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 briefly below. This will 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 and call it 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 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 large benefits. Additional 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, which will conclude with a discussion of the many remaining research and policy challenges in study protocol representation.

Elements and Specifications of 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, to guiding study execution. The documents vary greatly in length and content, but generally should include detailed background rationale; carefully stated scientific hypotheses; clear and complete eligibility criteria; well-specified measurements, data collection, and variables; and robust statistical 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 one is in development [3]. The International Conference on Harmonization E3 standard applies to describing the executed protocols of completed studies and is meant for a different audience and purpose than planned protocol documents created before study initiation.

The major elements of e-protocols overlap with, but are of necessity broader reaching and more standardized than the elements contained within study protocol documents. 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, e-protocols must satisfy both domain modeling (content requirements) as well as requirements for computability to satisfy a broad range of clinical research use cases. By considering what is required of the e-protocol to meet particular use cases, we illuminate the abstract common requirements for generic computable protocol models.

Content Requirements

Content requirements for the e-protocol are dictated by the ultimate functionality to be supported. We assert that 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, and (6) facilitate reuse of study data and artifacts (e.g., biosamples). 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 trial within the e-protocol. The minimal content requirements for each of the areas of desired functionality are described in the following sections.

Capture the Complete Study Plan in Computable Form

A first step toward computable study plans is to capture study plan in electronic, if not necessarily computable, form. Absent widely accepted guidelines on study protocol contents, Table 9.2 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 as well 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 suggest in Table 9.2 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. 17), 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.

Table 9.2 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

[a]These data elements are meant to be illustrative, not exhaustive

Provide 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 frontline staff have many complex protocol rules to follow (e.g., who to enroll, when to assess outcomes, and how/when to report adverse events). 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., adverse event reporting), and (3) participant-level and study data referenced by eligibility criteria and decision rules. Following sections discuss the representation of eligibility criteria and the requirements for achieving computability. We focus here on the content requirements for criteria, rules, and clinical data.

Clinical research studies cover the entire range of health and disease, so the broad answer to the question of "what are the content requirements for study protocol decision support?" is "all of medicine." The need for concept representations for all medical concepts is as much a challenge for clinical research informatics as it has been a challenge for health informatics over many decades. The challenge will most certainly require the exchange and use of knowledge from multiple domains. Different controlled terminologies may be used for subdomains in medicine (e.g., RxNorm for drugs; see Table 9.2), but there should be no bounds on the permissible domain content for e-protocols. Indeed, clinical 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 limited by the domain coverage of the clinical data that is 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 if 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 OBO Relations Ontology details ten relations: two foundational ones (is_a, part_of), and other physical (e.g., located_in), temporal (e.g., preceeded_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 about 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 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.

Facilitate 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, along with the ability to draw upon this repository to create standard data elements. Ideally, 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, and these forms would 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, the metadata describing the specifications for data capture should capture the core and full eligibility criteria, treatments received, treatment deviations, routine monitoring results for subject health status, any AEs that may occur, primary and secondary endpoint measurements, and any covariates or adjustment factors for the patient. 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]. 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 could be exported to electronic data capture (EDC) tools in the future to automatically program the up-front data validations into the system.

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. 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."

Support Appropriate Statistical Analysis and Reporting

The purpose of clinical research studies is to collect 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 (HbA1c$\geq$6.5%) should be analyzed using different statistical methods than diabetes as a continuous variable of HbA1c level.

The set of appropriate statistical tests to use depends 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 contents of e-protocols needed to support statistical analysis and reporting include a clear definition of study variables 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., of the primary outcome), 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].

Support 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, therefore, 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 non-randomized 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].

Facilitate 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 9.2). 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 EHR 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 Web Ontology 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 either UML or OWL, but can be 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 is the holy grail of clinical research informatics. An important gap in tooling includes easy-to-use and widely accessible tools that allow distributed editing and harmonization of conceptual models expressed in various formalisms.

Benefits of a Computable Study Protocol

Current Inefficiencies in Study Protocol Informatics

As described above, 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. One of the greatest sources of these inefficiencies 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 different companies and protocol sponsors ascribe different meanings to the same term.

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 through the e-protocol is the semantic foundation of clinical research informatics that adds value through improved clarity and communication. Such standardization can confer benefits even in a paper-based world; however, these benefits are leveraged much more 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 represents 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.

Use Cases 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, as shown in Fig. 9.1. We discuss several illustrative examples.

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 [11]. 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 [12]. 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) [13].

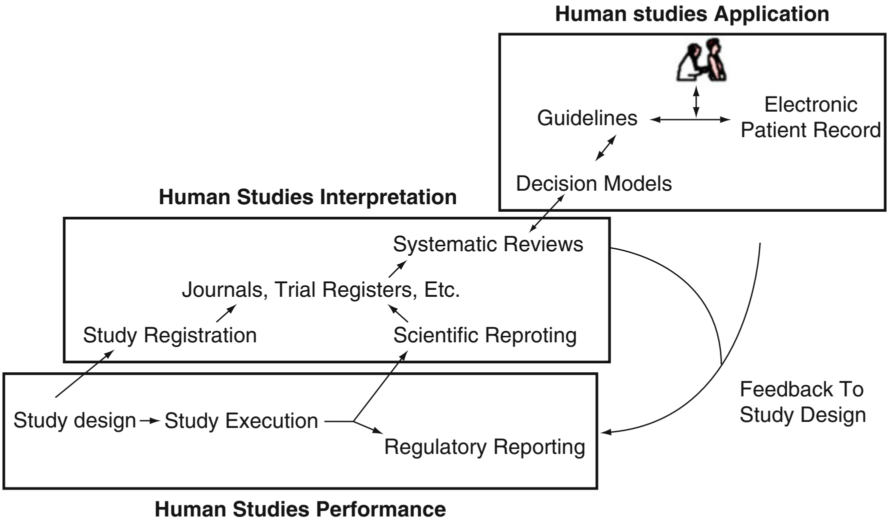


Fig. 9.1 Life cycle of human studies

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 institutions patient data repository for automated cohort discovery [14,15]. At the study design stage, an investigator could tweak 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.

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.

Moreover, many clinical questions are addressed by more than one study, 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 [16].

Benefits of a Common Computable Protocol Model

Figure 9.1 illustrates that a single clinical study involves many different people and many different information systems over its life cycle. For each system to take advantage of the computable e-protocol of the study, each system will have 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 were 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 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 current protocol representation standard activities, and the chapter will conclude with a view to the future.

Protocol Representation Standards

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. HL7 is an American National Standards Institute (ANSI)-accredited standards development organization. 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.

The HL7 Reference Information Model (RIM) V2.0 allows exchange of information on clinical care processes through *technical* interoperability. HL7 V3.0 is a standardized model to represent healthcare information, and 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 through a collaboration of biopharmaceutical, regulatory, academic, and technology partners with a goal toward optimizing clinical research through the creation and adoption of standards. The CDISC Protocol Representation Group (PRG) has developed the

CDISC Protocol Representation Model (PRM) V1 (http://cdisc.org/standards/protocol.html) 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: "To develop a standard structured protocol representation that supports the entire life cycle of clinical research protocols to achieve semantic interoperability (the exchange of content and meaning) among systems and stakeholders." The CDISC 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 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.

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. 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

The Biomedical Research Integrated Domain Group (BRIDG) model strives to be an overarching protocol-driven biomedical model in support of clinical 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. 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 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 [16]. Thus, 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 [17]. 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) [18,19]. 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 [12]. 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 eligible for 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 eligibility criteria are often written in idiosyncratic free-text sentence fragments that can be ambiguous or underspecified (e.g., "candidate for surgery"). Indeed, in one study, 7% of 1,000 eligibility criteria randomly selected from ClinicalTrials.gov were found to be incomprehensible [20]. The remaining criteria exhibited a wide range of complex semantics: 24% have negation, 45% have Boolean connectors, 40% include temporal data (e.g., "within the last 6 months"), and 10% have if-then constructs. Formal representations of eligibility criteria should ideally 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 defines 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) [21]. For each criterion, ASPIRE defines 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. As of 2008, disease-specific standardized criteria have 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 benefit of lower adoption barriers.

The Eligibility Rule Grammar and Ontology (ERGO) project takes a different approach than ASPIRE. ERGO aims 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 [22]. 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 [23]. In preliminary work, natural language processing techniques were used to assist in transforming eligibility criteria from free-text into ERGO Annotation.

Other eligibility criteria representations include caMatch, SAGE, and GLIF [24–26]. 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 [27]. 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.

The Protocol-Model Driven 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 clinical research informatics 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 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 and would pay immense dividends for clinical research and science.

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 Making. 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. German 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. 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 Programs Biomed. 1994;43:283–91.

12. 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.
13. 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?papetoc. Accessed Aug 2011.
14. 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.
15. 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. Springer: New York; 2010.
16. 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.
17. HSDB Wiki. Home page. 2010. Available at https://hsdbwiki.org/index.php/HSDB_Collaborative_Wiki. Accessed Aug 2011.
18. 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. doi: http://doi.acm.org/10.1145/1183568.1183574.
19. Speedie SM, Taweel 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.
20. 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.
21. Niland J. ASPIRE: agreement on standardized protocol inclusion requirements for eligibility. In: An unpublished web resource. 2007.
22. Tu SW, Peleg M, Carini S, Rubin D, Sim I. ERGO: a template-based expression language for encoding eligibility criteria 2008. 2008. http://128.218.179.58:8080/homepage/ERGO_Technical_Documentation.pdf.
23. Tu S, Peleg M, Carini S, Bobak M, Ross J, Rubin D, Sim I. A Practical Method for Transforming Free-Text Eligiblity Criteria into Computable Criteria. J Biomed Inform. April, 2011;44(2): 239–250. Epub 2010 Sep 17 PMID: 20851207.
24. Cohen E. caMATCH: a patient matching tool for clinical trials. In: caBIG 2005 Annual Meeting, Bethesda, MD. April 12-13, 2005.
25. Tu SW, Campbell JR, Glasgow J, Nyman MA, McClure R, et al. The SAGE guideline model: achievements and overview. JAMA. 2007;14:589–98.
26. Boxwala A. GLIF3: a representation format for sharable computer-interpretable clinical practice guidelines. J Biomed Inform. 2004;37:147–61.
27. 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.

Chapter 10
Data Quality in Clinical Research

Meredith Nahm

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 collect and process research data 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 plan, achieve, and control the quality of research data.

Keywords Clinical research data • Data quality • Research data collection • Processing methods • Informatics • Management of clinical data • Data accuracy

Clinical Research Data Processes and Relationship to Data Quality

Data quality is foundational to our ability to human research. Data quality is so important that an Institute of Medicine report [1] was written on the topic. Further, two key thought leaders in the quality arena, W. E. Deming and A. Donabedian, specifically addressed data quality [2–4].

Failing to plan for data quality is an implicit assumption that errors will not occur. Emphasizing that failing to plan for data quality further threatens data quality by inhibiting the detection of errors when they do occur, Stephan Arndt et al. state,

M. Nahm, Ph.D.
Informatics, Duke Translational Medicine Institute, Duke University,
2424 Erwin Road, Durham, NC 27705, USA
e-mail: meredith.nahm@duke.edu

R.L. Richesson, J.E. Andrews (eds.), *Clinical Research Informatics*, Health Informatics,
DOI 10.1007/978-1-84882-448-5_10, © Springer-Verlag London Limited 2012

"Ironically, there is a major difference between a process that is presumed through inaction to be error-free and one that monitors mistakes. The so-called error-free process will often fail to note mistakes when they occur" [5].

Quality is broadly defined by Juran as fitness for use [6]. Unfortunately, for clinical investigators and research teams, the use varies from study to study. In clinical research, data collection processes are often customized according to the scientific questions and available resources, resulting in different processes for individual studies or programs of research. Because data quality assurance and control are largely dependent on how data are collected and processed, they are complicated by this mass customization. (The label *mass customization* used to describe clinical research by Karen Koh in a meeting at Duke Clinical Research Institute.) Given the likely persistence of science-driven customization, an antidote may lie in methods for data quality planning. It is only when a planning framework exists and is used that knowledge gained from work on prior projects can translate to new projects with different data sources, processes, and people.

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 patient-reported items. Each data source is associated with a method 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 quality plan should take into account the data source, precollection processing, the data acquisition method, and, finally, postprocessing. While these elements of the data quality plan 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 treat all data will overlook both errors 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 methods, thus the need to take into account the data source, precollection processing, data acquisition, and postprocessing in data quality planning. This chapter is primarily concerned with how to accomplish this and will give the reader a framework to use to assure and control quality regardless of the data source, acquisition method, or processing.

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 must address data quality considerations, for example, by specifying a consistent suitable collection method, planning interrater reliability assessments for subjective assessments, or other collection of independent data. The resulting degree of data quality affects how data can be used and, ultimately, the level of confidence that can be reposed in research findings or other

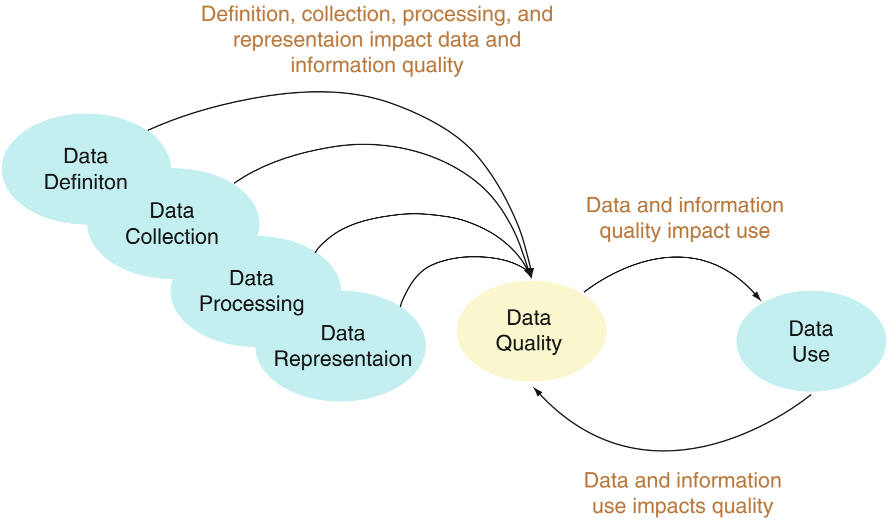

Definition, collection, processing, and representaion impact data and information quality
Data Definiton
Data Collection
Data Processing
Data Representaion
Data Quality
Data and information quality impact use
Data Use
Data and information use impacts quality

Fig. 10.1 Impacts of data generation and handling features on data and information quality. The way data and information are handled impacts the quality of that data and information. The quality of data and information impacts our willingness and ability to use it. Use of data and information causes more care to be taken in their handling, increasing the quality

decisions based on the data. Thus, protocol and Case Report Form (CRF) design, including data capture methods, must be concerned with data quality assurance measures from the start.

Data quality and the discipline of informatics are inextricably linked. Data definition, collection, processing, representation, and use are central to informatics (Fig. 10.1). Definition, collection, processing, representation, and use impact data and information quality, and data and information quality impact use. In turn, data and information that are used are more likely to have higher quality. In clinical research, data can be collected both prospectively and retrospectively, depending on the protocol and local procedures at the clinical investigational site. Therefore, information use in clinical care as well as information use in the study may impact data quality.

Each step in the collection, handling, and processing of data affects 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” [7]. We suggest a less literal interpretation of the ICH E6 guidance document. The gold standard in achieving quality is prevention rather than after-the-fact finding and fixing errors; thus, 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.

Assuring and controlling data quality are largely a focus on the presence of *data errors*, defined here as a data value that does not accurately reflect the true state of the thing being represented. Data represent things (or states of things) in the real world. Since all things change with time, so does the accuracy with which a data value represents the true state of the represented thing. Thus, data that are correct will not necessarily remain so with the passage of time. The qualifiers necessary for a given datum to remain accurate over time are often referred to as context, for example, patient age as of the first study visit; or air temperature in degrees Celsius at latitude 35.620252°N, longitude −82.542933°W, at an elevation of 2,310 ft at noon on May 23, 2009; 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 collection and processing, as demonstrated by the following examples adapted from actual cases. In this chapter, we develop and apply a framework for preventing and controlling data errors in the context of clinical research. The following examples come from the Society for Clinical Data Management [8].

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 rather than human subjects. The data listed on the data collection form were checked and were found to match 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.

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 principle 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; rather, the problem was due to simple human error, possibly from working on multiple studies with differing form formats.

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, these sites enrolled 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.

Each of these scenarios describes a data quality problem, one in device-based data collection, one in data processing, one in measurement procedure. Despite the differences in setting and in the sources of the errors, the end result was the same: inaccurate data.

The 1999 Institute of Medicine Report [1] emphasized the importance of data quality to regulatory decision-making, i.e., drawing conclusions from clinical trials. At the time, there was little in the literature base to synthesize in the report. Further, since the IOM report, there has been scant methodological progress toward data quality assurance, assessment, and control in clinical research. The framework 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 due to calibration drift or measurement error. Nagurney reports that in a recent study, up to 8% of subjects could not recall historical items and up to 30% gave different answers on repeat questioning [9]. 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 [10].

Even 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 [11]. For most types of assessment, observation, or interpretation of test results, reports of error or agreement rates can be found in the literature. These known and real errors and inconsistencies are often not accounted for in data quality planning in clinical research.

Moreover, in every process, nature affects every project every day, increasing disorder. 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 [12] and others [13] 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. Such concern should be taken into account in data quality planning; for example, many organizations perform a review of blinded tables, listings, and figures prior to closing a database, to identify such obvious errors. 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.

It is important to note that cleaner data can save time in programming and data use, but this is likely concomitant with additional costs. As such, and within the context of a given research project, pursuing data quality to a greater extent than needed to support the conclusions 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 for the experienced informaticist to explore.

Defining Data Quality

The Institute of Medicine (IOM) defines quality data as “data strong enough to support conclusions and interpretations equivalent to those derived from error-free data” [1]. Like Joseph Juran’s famous “fitness for use” definition [6], the IOM definition is use dependent. Further, the robustness of statistical tests and decisions to data errors differs. Thus, applying the IOM 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 set the acceptance criterion for data quality.

Further specification of the IOM definition of data quality is necessary for operational application. Other authors who have discussed data quality define it as a *multidimensional concept* [14–20]. In clinical research, the dimensions most commonly considered are *reliability*, *validity*, *accuracy*, and *completeness* [21]. 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 the trial (as in the case of data and safety monitoring boards), *timeliness* has emerged as an important data dimension. Further, regulatory authorities are concerned with trustworthiness of the data and initially identified the following data quality dimensions for clinical research: "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" [22].

These "fundamental elements," *attributable*, *legible*, *contemporaneous*, *original*, and *accurate*, are commonly referred to as ALCOA. Registries commonly report data quality in terms of accuracy and completeness [23]. 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 attributes, or descriptors of data quality, 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 fundamental dimensions of data quality will become a necessary data itself.

The multidimensionality data quality causes ambiguity because any given use of the term might refer to a single dimension or to a 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 stated business rules, while regulatory authorities are concerned with attribution and verifiability [22]. 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, can be used to assess the quality of the data for primary and secondary uses.

All dimensions apply to any use of data, but often the circumstances surrounding a given (or the primary) use include built-in processes that assure a relevant dimension is present and addressed. For example, in a clinical trial, those who use data often have a role in defining it, meaning the *definition* is of little concern. 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 in new situations. In other words, carving data quality up into dimensions helps us design for, measure or assess, control, and increase data quality. A consensus set of dimensions for clinical research does not yet exist. Here, we will primarily address the dimensions of *accuracy*, *completeness*, *timeliness*, *accessibility*, *relevance*, *and volatility*. *Reliability* and *validity* are addressed in Chaps. 4 and 11, as noted, and data *definition* (full specification) is addressed in Chap. 13.

Using multiple dimensions to characterize data quality, and measuring those dimensions to assess data quality, requires both operational definitions and acceptance criteria for each dimension of quality. An approach that will allow 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 a given set of data is needed and the actual date it is 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. Table 10.1 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.

Framework for Data Quality Planning

Over the past decade or more, the number and diversity of both new technology and new data sources have increased. Managing new technology or data sources on a given project is now a normal aspect to clinical research data management. One of the largest problems is preparing data managers to work with new technology and data sources. Simply put, a framework is needed that will enable data managers to assess a given data collection scenario, including new technology and data sources, and systematically evaluate that scenario, apply appropriate methods and processes, and achieve the desired quality level.

A dimension-oriented approach provides a framework that practitioners can rely on when handling data in a novel situation (e.g., data from a different source, in a

Table 10.1 Data quality dimensions for clinical research

Dimension	Type	Natural language definition	Operational definition/ metric
Accuracy	Inherent	*States in the data match the intended state in the real world*	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)	Use/need date minus date data last updated
Completeness	Inherent	*The extent to which every represented real-world state is reflected in the data*	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	*Length of time from a change in the real-world state to the time when the data reflect the change*	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 values applicable to intended use
Granularity	Context dependent	Level of detail captured in data	Percentage of values at level of detail appropriate for intended use
Specificity (nonambiguity)	Inherent	*Each state in the data definition* (metadata) *corresponds to one* (or no) *state of the real world*	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 [18]

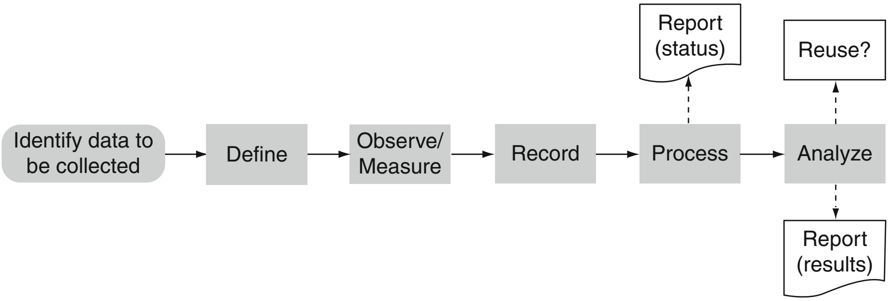


Fig. 10.2 Data-centric view of the research process. A set of general steps for choosing, defining, observing, or otherwise measuring, recording, analyzing, and using data apply to almost all research (From Data Gone Awry [8], with permission)

new environment, or using new technology). Such a framework helps guard against methodological omissions and assures that data will meet specified needs. However, data quality dimensions alone are an incomplete solution. A systematic way to assess data sources and processes on a project is necessary. Figure 10.2 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, (3) observing and measuring values, (4) recording those observations and measurements, (5) processing data to render them in electronic form and prepare them for analysis, and (6) analyzing data. While research is ongoing, data may be used to manage or oversee the project. After the analysis is completed, results are reported, and 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*, *weight*, *size of package*, *and color of paper towels*). When writing a protocol, the investigator be as specific as possible because in multicenter trials, 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 sample type on which the test is to be conducted (e.g., pregnancy test is to be performed on women of childbearing potential).

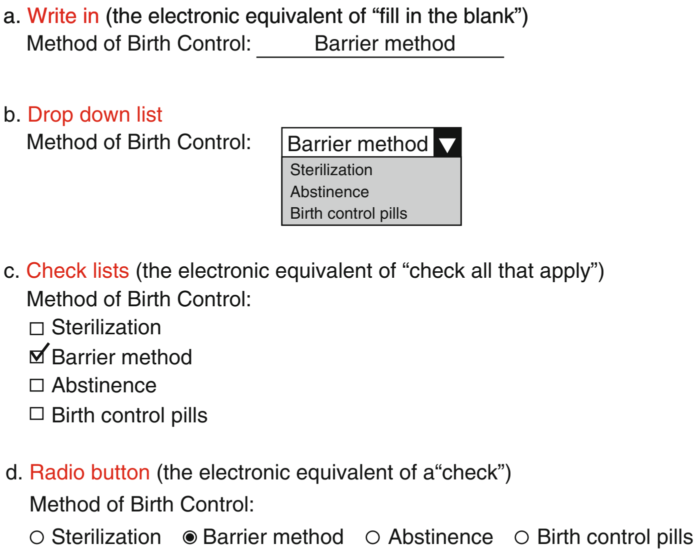
As standards such as the Protocol Representation Standard [24] mature and supporting software becomes available, full specification of protocol elements will become the most efficient method for defining data, as metadata specified in the protocol will be immediately available for generation of data 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 not measuring the same construct. 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, 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 that it is necessary only to collect the data needed to assure patient safety, answer the scientific question(s), and uniquely identify the collected data elements. Jacobs and Studer report that for every dollar spent to produce a data collection form, $20–$100 are required to fill each one in, process it, and store it, emphasizing that "the true cost of a form involves people not paper" [25]. When extensive data cleaning is required, this ratio becomes even more exaggerated. Eisenstein and colleagues report extensive cost savings in clinical trials by decreasing the number of data collection form pages [26, 27]. At the time of this writing, the relationship between form length and data accuracy for online forms remains unprobed [28]. Further, the evidence relating form length to decreased response rate while considered equivocal by some [28] has been demonstrated in controlled and replicated experiments [29, 30]. There is no question, however, that collecting more data increases costs and places additional burden on clinical investigational sites and data centers [26, 27].

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 (1991) 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. A good introductory work is Shneiderman and



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. 10.3 Example data collection structures. For many data elements, more than one data collection structure exists

Plaisant (2004) Designing the User Interface: Strategies for Effective Human-computer Interaction. While this topic is too broad to discuss in depth here, two principles that are directly relevant to clinical research informatics, and 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* [31]. A general assumption is that the more structured the data, the higher the degree of accuracy and ease of processing. We will see, however, that this can be counterbalanced by considerations related to ease of use.

As a general principle, the data collection structure should match the type data. Data elements can be classified according to Stevens’ scales (nominal, ordinal, interval, and ratio) [32], or as categorical versus continuous. 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. 10.3.

Mismatches between data type and collection structure, for example, collecting data in a structure more or less granular than reality, can cause data quality problems. Collecting data at a more granular structure than exists or than 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*, than can be discerned in reality also invites variability and results in information loss. The real granularity 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 Steven's scales. 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 useful. 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 is 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 [33–35]. 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 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, and the BMI should be calculated by a computer. Sometimes, this principle is stated as "collect raw data." 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 [33–35]. An additional application of the compatibility-proximity principle is that all items that the person completing the form needs to complete his or her task should be immediately apparent on the form itself (separate form completion instruction booklets are less effective) [34]. There is evidence that data elements with higher cognitive load on the abstractor or form completer also have higher error rates [35–47]. Adhering to the compatibility-proximity principle, by decreasing cognitive load, may help prevent this.

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 outweigh 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, a real-time solution or tool to support the additional cognitive tasks is needed; such a tool may require raw data as input; and (4) it may not be possible to design forms that match clinical 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 and selection of one over the alternatives is a design decision requiring knowledge of the advantages and disadvantages of each option and how they impact costs and the 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 and full specification, defining 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 types of measurement and observations 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. Some measurements return a concrete number (e.g., temperature) or answer, while others require interpretation (e.g., the trace 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 eliminates the possibility to correct errors in observation. For this reason, error checking processes should be built into measurement and observation whenever feasible. This can be accomplished by building redundancy in to data collection processes [48, 49]. 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. 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 error checking with immediate feedback. Error checking is a comparison of a recorded value against a known standard, for example, valid ranges, or relative comparison to another value. While error checking can identify some errors, it will miss those within the valid value set. Errors within the valid value set can only be identified through redundancy. Secondly, error checking may occur at the point of measurement or recording, but is usually not built in to measurement processes, and thus occurs after the fact, and only serves as an identification mechanism, rather than as a correction mechanism. In summary, measurement discrepancies can be mitigated through careful procedures and training; however, errors are nonetheless inevitable. While error checking near or after measurement can identify errors, immediate independent verification with contemporaneous feedback remains the safest option.

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" example given at the beginning of the chapter reflects an all-too-common problem rooted in the fact that clinical investigational sites each practice 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 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 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 quality, 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 US Food and Drug Administration regulations codified in 21 CFR Part 58 [50], 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 [51, 52]. 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. However, 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 [53]. 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 [53].

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 10.2). While optical scanning methods could provide accuracy comparable to key-entry methods, they were 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

Table 10.2 Accuracy associated with common data processing methods

	Min.	Median	Mean	Max.	Std. Dev.	LCL	UCL
Abstraction	70	647	960	5,019	1,018	510	818
Optical	2	81	207	1,106	338	4	220
Single entry	4	26	80	650	150	21	36
Double entry	4	15	16	33	10	6	24
No batch data cleaning	2	270	648	5,019	946	200	475
Batch data cleaning	2	36	306	1,351	428	23	287

methods [53]. Additionally, other factors have been hypothesized in the literature, but an association has yet to be established, for example, staff experience [53], number of manual steps [54], and complexity of data [51]. For these reasons, measurement of data quality is listed as a minimum standard in the Good Clinical Data Management Practices document [54]. 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 [55], publication of data accuracy with clinical research results should be required.

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 to actively manage 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, capture 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

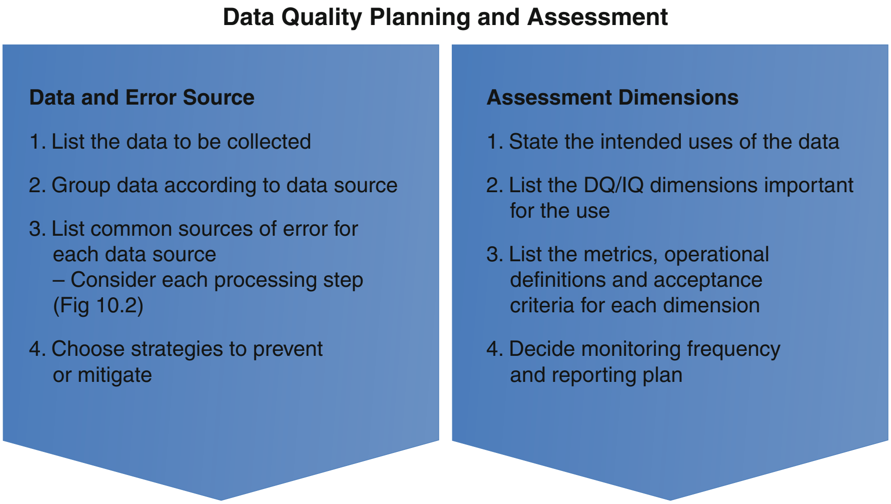
Analyzing and reporting data differ fundamentally from other steps discussed in the preceding sections, as they lack the capacity to introduce error into the data values themselves. Errors in analysis and reporting programming or data presentation, while potentially costly, do not change underlying data. Analysis and reporting programming is typically applied to a copy of the database. However, analysis and reporting do have the potential to misrepresent the data. Assuring and controlling quality at the analysis and reporting stage is achieved through choice of appropriate methods, through validation of programming, and through applying the compatibility-proximity principle to data displays through matching the scale of represented data and representing display.

Using the Framework to Plan 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 framework presented here can be used to structure this task to increase the transparency of decisions to the research team and lessen the likelihood that anything is missed. The first step is to group the data to be collected into categories depending on source, for example, medical history and medications will be abstracted from the medical record, blood pressures will come from a study provided device, lab values will be transferred electronically from a central lab, and so on. Where data sources within the medical record are varied, a more granular treatment may be required. The data sources and process by which the data are obtained can then be diagrammed, and alternative sources, methods, and processes can be considered. For example, some data sources may have undesirable preprocessing steps or known higher variability and may be excluded from consideration. Once the data sources have been chosen and the data gathering process has been specified, the steps in Fig. 10.4 can be applied to identify known error sources, to consider the possibility or desirability of preventing or mitigating the error, and to evaluate the methods for accomplishing the change. Data quality dimensions that are important to the research study are assessed for each type of data and each processing step. The output of this process is then discussed with the research team and incorporated into the plan for data collection and management. Importantly, application of this framework is a tool and mental exercise to use in planning and a tool to promote discussion and informed decision-making by the research team. Use of such a framework should impact the data collection and management plan, ultimately optimizing data quality. Use of this framework to produce an additional written document is explicitly not the intent (Fig. 10.4).

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 dangerous 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.



Data Quality Planning and Assessment

Data and Error Source

1. List the data to be collected
2. Group data according to data source
3. List common sources of error for each data source – Consider each processing step (Fig 10.2)
4. Choose strategies to prevent or mitigate

Assessment Dimensions

1. State the intended uses of the data
2. List the DQ/IQ dimensions important for the use
3. List the metrics, operational definitions and acceptance criteria for each dimension
4. Decide monitoring frequency and reporting plan

Fig. 10.4 Data Quality Planning and Assessment. This framework links data quality planning and assessment with the decisions about which data elements to collect. During planning, data to be collected are listed and grouped by type and or source of data. Known error sources for each are considered and deliberate decisions are made about prevention, mitigation, or doing nothing. At the same time, the data quality dimensions important to the intended use are identified. Metrics, acceptance or action criteria and operational definitions for each are developed as well as reporting plans. Some mitigation strategies may prompt inclusion of metrics and monitoring for known error types

The following criteria are commonly assessed in preaward site visits and audits. 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.

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 capabilities. 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 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 [55–61].

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

International Standards Organization (ISO) quality system standards. Variation 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. 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 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 much of the data collection and processing, and human performance directly affects 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 (Fig. 10.5).

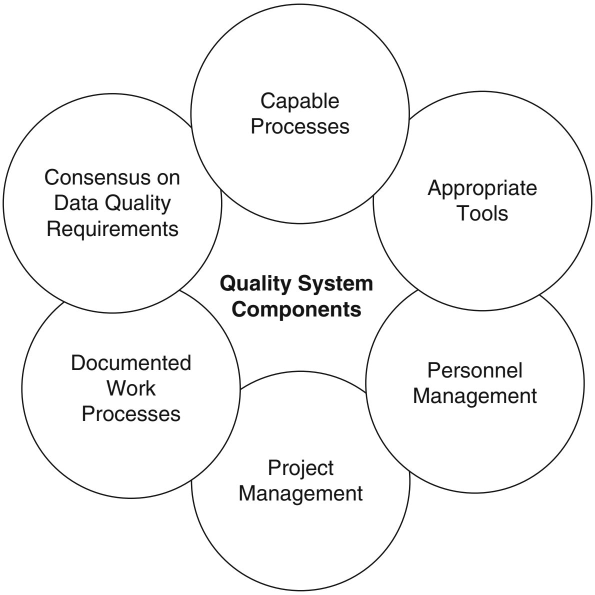

Capable Processes
Consensus on Data Quality Requirements
Appropriate Tools
Quality System Components
Documented Work Processes
Personnel Management
Project Management

Fig. 10.5 Components of a data quality system. The environment in which data are collected and processed impacts data quality. Thus, achieving and controlling data quality usually requires action from entities in the broader environment

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 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 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 through adding variability, potentially decreasing power. As shown conceptually in Fig. 10.6, 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 [62, 63]. 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 [26] and attenuate correlation coefficients to the null [64–66] (i.e., for trials that fail to reject the null hypothesis, data errors rather than a true lack of effect could be responsible) [67]. In the context

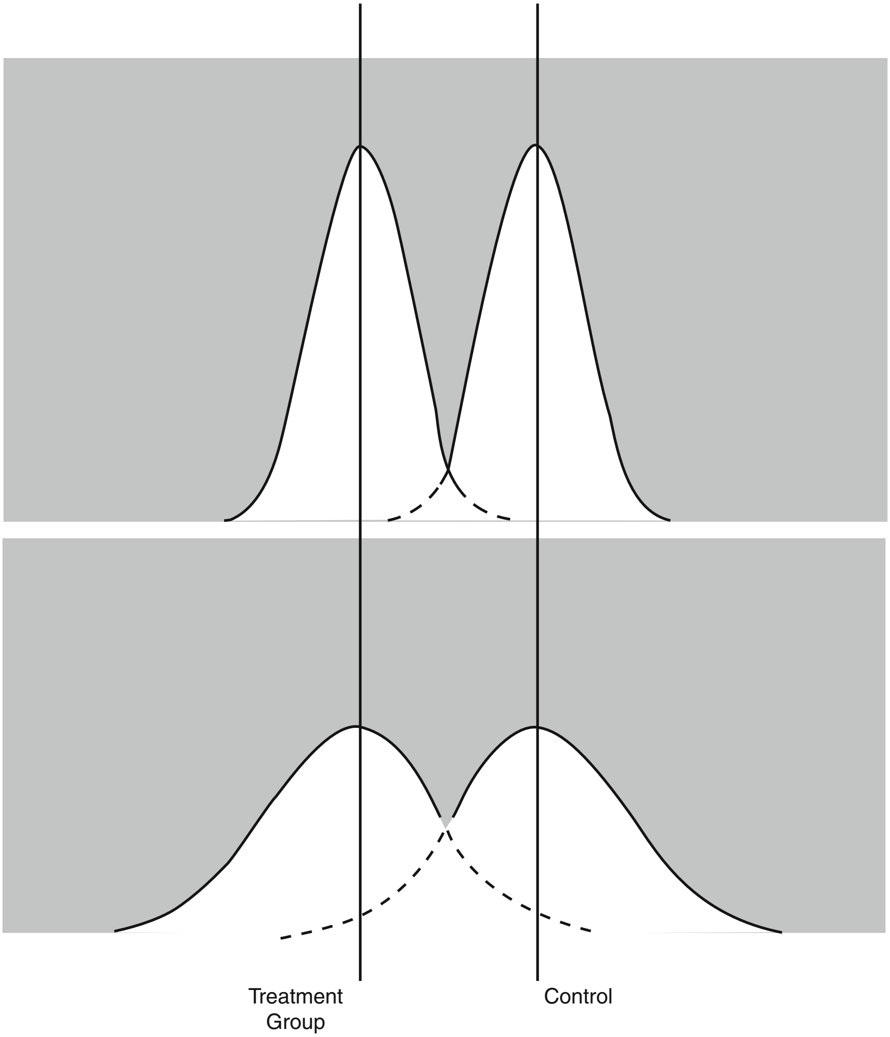


Fig. 10.6 Effect of adding variability. The top two distributions have less variability (are narrower) than the bottom two, making it easier to tell them apart both visually and statistically

of large data error rates, 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 [55, 63, 66]. The adverse impact of data errors has also been demonstrated in other secondary data uses such as registries and performance measures [68–74]. Thus, whether or not data are of acceptable quality for a given analysis is a question to be assessed by the study statistician. The assessment should be based on measured error and completness rates.

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, (3) some errors can be prevented, (4) some errors can be detected, and (5) some errors can be corrected. The sets in 3–5 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 matching methods for prevention, mitigation, detection, and correction where they exist. Further, the skilled practitioner applies this knowledge to design clinical research data collection and management processes that provide the needed quality at an acceptable cost. 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, vol. II: medical care appraisal – quality and utilization. New York: American Public Health Association; 1969. 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. Juran JM, Gryna FM. Juran's quality control handbook. 4th ed. New York: McGraw-Hill; 1988.
7. Guidance for industry E6 good clinical practice: consolidated guidance, ICH E6. April 1996. Available from http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM073122.pdf. Accessed Aug 2011.
8. 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

9. 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.
10. Feinstein AR, Pritchett JA, Schimpff CR. The epidemiology of cancer therapy. 3. The management of imperfect data. Arch Intern Med. 1969;123:448–61.
11. Reason J. Human error. Cambridge: Cambridge University Press; 1990.
12. Nahm M, Dziem G, Fendt K, Freeman L, Masi J, Ponce Z. Data quality survey results. Data Basics. 2004;10:7.
13. Schuyl ML, Engel T. A review of the source document verification process in clinical trials. Drug Info J. 1999;33:789–97.
14. 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.
15. Pipino L, Lee Y, Wang R. Data quality assessment. Commun ACM. 2002;45:8.
16. Tayi GK, Ballou DP. Examining data quality. Commun ACM. 1998;41:4.
17. Redman TC. Data quality for the information age. Boston: Artech House; 1996.
18. Wand Y, Wang R. Anchoring data quality dimensions in ontological foundations. Commun ACM. 1996;39:10.
19. Wang R, Strong D. Beyond accuracy: what data quality means to data consumers. J Manage Inform Syst. 1996;12:30.
20. Batini C, Scannapieco M. Data quality concepts, methodologies and techniques. Berlin: Springer; 2006.
21. Wyatt J. Acquisition and use of clinical data for audit and research. J Eval Clin Pract. 1995;1: 15–27.
22. U.S. Food and Drug Administration. Guidance for industry. Computerized systems used in clinical trials. In: Services DoHaH, editor. Rockville: U.S. Food and Drug Administration; 2007.
23. 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.
24. (CDISC) CDISC. The Protocol Representation Model version 1.0 draft for public comment: CDISC; 2009. p. 96. Available from http://www.cdisc.org
25. Jacobs M, Studer L. Forms design II: the course for paper and electronic forms. Cleveland: Ameritype & Art Inc.; 1991.
26. 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.
27. Eisenstein EL, Collins R, Cracknell BS, et al. Sensible approaches for reducing clinical trial costs. Clin Trials. 2008;5:75–84.
28. Galešic 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.
29. Roszkowski MJ, Bean AG. Believe it or not! Longer questionnaires have lower response rates. J Bus Psych. 1990;4:495–509.
30. 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.
31. Wickens CD, Hollands JG. Engineering psychology and human performance. 3rd ed. Upper Saddle River: Prentice Hall; 2000.
32. Stevens SS. On the theory of scales of measurement. Science. 1946;103:677–80.
33. Allison JJ, Wall TC, Spettell CM, et al. The art and science of chart review. Jt Comm J Qual Improv. 2000;26:115–36.
34. Banks NJ. Designing medical record abstraction forms. Int J Qual Health Care. 1998;10: 163–7.
35. 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.

36. 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.
37. Fritz A. The SEER program's commitment to data quality. J Registry Manag. 2001;28(1): 35–40.
38. 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.
39. 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.
40. 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.
41. 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.
42. 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.
43. 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.
44. 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.
45. To T, Estrabillo E, Wang C, Cicutto 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.
46. Yawn BP, Wollan P. Interrater reliability: completing the methods description in medical records review studies. Am J Epidemiol. 2005;161(10):974–7.
47. 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. Communication Reports. 1 Apr 2007.
48. Helms Ron. Redundancy: an important data forms/design data collection principle. In: Proceedings Stat computing section, Alexandria; 1981. p. 233–237.
49. Helms R. Data quality issues in electronic data capture. Drug Inf J. 2001;35:827–37.
50. U.S. Food and Drug Administration regulations Title 21 CFR Part 58. 2011. Available from http://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/cfrsearch.cfm?cfrpart=58. Accessed Aug 2011.
51. Nahm ML, Pieper CF, Cunningham MM. Quantifying data quality for clinical trials using electronic data capture. PLoS One. 2008;3(8):e3049.
52. Winchell T. The mystery of source documentation. SOCRA Source 62. 2009. Available from http://www.socra.org/.
53. Nahm, M. Data Accuracy in Medical Record Abstraction. Doctoral Dissertation, University of Texas at Houston, School of Biomedical Informatics, Houston Texas, May 6, 2010.
54. SCDM. Good clinical data management practices. www.scdm.org. Society for Clinical Data Management; 2010. Available from http://www.scdm.org
55. 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.
56. Svolba G, Bauer P. Statistical quality control in clinical trials. Control Clin Trials. 1999;20(6): 519–30.
57. 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.

58. 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.
59. McNees P, Dow KH, Loerzel VW. Application of the CuSum technique to evaluate changes in recruitment strategies. Nurs Res. 2005;54(6):399–405.
60. 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.
61. 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 Making. 2009;29(2):247–56.
62. 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.
63. 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.
64. Mullooly JP. The effects of data entry error: an analysis of partial verification. Comput Biomed Res. 1990;23:259–67.
65. Liu K. Measurement error and its impact on partial correlation and multiple linear regression analyses. Am J Epidemiol. 1988;127:864–74.
66. Stepnowsky Jr CJ, Berry C, Dimsdale JE. The effect of measurement unreliability on sleep and respiratory variables. Sleep. 2004;27:990–5.
67. 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.
68. 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.
69. 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.
70. 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
71. 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.
72. Jacobs R, Goddard M, Smith PC. How robust are hospital ranks based on composite performance measures? Med Care. 2005;43:1177–84.
73. Pagel C, Gallivan S. Exploring consequences on mortality estimates of errors in clinical databases. IMA J Manage Math. 2008;20(4):385–93. http://imaman.oxfordjournals.org/content/20/4/385.abstract
74. Goldhill DR, Sumner A. APACHE II, data accuracy and outcome prediction. Anaesthesia. 1998;53:937–43.

Chapter 11
Patient-Reported Outcome Data

Robert O. Morgan and Kavita R. Sail

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 healthcare 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.

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

R.O. Morgan, Ph.D. (✉)
Department of Management, Policy and Community Health,
University of Texas School of Public Health,
1200 Herman Pressler Boulevard, Room 343E, Houston, TX 77030, USA
e-mail: robert.o.morgan@uth.tmc.edu

K.R. Sail, Ph.D.
Department of Health Care Informatics,
McKesson/US Oncology Specialty Care Unit,
10101 Woodloch Forest Drive, The Woodlands, TX 77380, USA
e-mail: kavita.sail28@gmail.com

R.L. Richesson, J.E. Andrews (eds.), *Clinical Research Informatics*, Health Informatics,
DOI 10.1007/978-1-84882-448-5_11, © Springer-Verlag London Limited 2012

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]; however, it is generally considered good practice to use both types of PROs where possible [3].

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. [5]. 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*: For example, 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 [6] provide an alternative classification system for PROs that incorporates many of the same elements presented above.

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 [7]. 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 [8]. 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 [9].

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 healthcare 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 healthcare 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 [9].

Reliability

The reliability of a measure refers to the *stability* or *equivalence* of repeated measurements of the same phenomena within the same patient [10]. 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]) [11] 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 [10, 12].

Test-retest reliability is estimated by the correlation between responses to same measure by 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 [10]. 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 [12]. 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 [12]. 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 [11, 12].

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*) [10, 12].

Table 11.1 Methods of computing validity

	Types of validity		
Methods	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 [12], Table 3.3 (p. 64). Reprinted with permission of John Wiley & Sons, Inc

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 11.1) [11]. 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 [10, 11]. 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 11.1). 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 11.1) [10, 12]. 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.

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 [13]. Although the general concept of responsiveness is straightforward, there is no consensus on the best way to measure it. McDowell provides a summary of different approaches, all of which reflect some form of standardizing the change score [13].

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 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 [14]. 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 [12].

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) [12, 14].

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 [12, 14]. 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 [15], 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, 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 through priority mail

E-Mail and Web Surveys

E-mail surveys are self-administered PRO instruments sent through electronic mail. They are simpler to compose and send than web surveys, but are more limited with respect to their visual stimulation and interaction capabilities. They also provide limited options for structural features of instruments such as skip patterns. The design principles for implementing an e-mail survey are [15]:

- 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 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 an 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. A replacement survey should be included with any follow-up contact.

Web surveys are self-administered surveys accessed through the Internet. Web surveys are not usually sent through electronic mail, although a link to a URL may be. 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 [15]. 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.

E-mail and web surveys offer several advantages over mailed surveys. They are usually 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 [15]. New technology and software have made implementation of e-mail and web-based PROs relatively straightforward – often for little or no cost.

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 [15].

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* [16].

Voice auditory devices: These devices are often referred to as interactive voice response (IVR) and are usually telephone-based, although voice over Internet protocols (VOIP) will likely be increasingly incorporated into their designs. 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 [16]: 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 [16]. However, access to these devices is neither universal nor necessarily representative of the populations of interest [15].

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 [16].

Handheld computer system: These systems use a special pen/stylus to enter the data. The main advantage is the portability of the system due to its light weight. However, the limited screen space leads to smaller fonts and may require the respondent to scroll down to view the entire question and response set [16].

Web-based systems: These systems require access to a computer or device with Internet service [15, 16]. They offer the respondent the convenience of completing the questionnaire in their leisure time or at home, as well as the advantage of capturing the data in the data file as the patient is responding to the questionnaire. These devices can also be adjusted to changes in the protocol during a study period at a lesser cost since modifications are made to the software residing on a central server. Other screen-based systems require software changes to be uploaded to each device, potentially creating logistical and technical challenges [16].

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 [16].

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. The rapid growth of "smart phones" with sophisticated web interfaces (e.g., the iPhone or phones using the Android or similar operating systems), 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 its infancy, this class of mobile devices holds substantial potential for broad application [17]. 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 [18–20].

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, 10, 12, 15]. The following is a summary of the key guidelines presented by DeVellis [10]:

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 is 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 11.2.
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?

Table 11.2 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 scale
Likert scale	It is 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 non redundancy

FDA [1]

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.
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 [12].

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 exclusion; 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).

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. King's College London 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; http://www.proqolid.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 June, 2010, the PROQOLID site provided information on over 600 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). 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 [21]. 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 healthcare 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, are 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. 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.
6. 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.
7. Skinner J, Teresi J, et al. Measurement in older ethnically diverse populations: overview of the volume. J Ment Health Aging. 2001;7:5–8.
8. Anastasi A. Psychological testing. 6th ed. New York: Macmillan Publishing Company; 1998.
9. Morgan R, Teal C, et al. Measurement in VA health services research: veterans as a special population. Health Serv Res. 2005;40:1573–83.
10. DeVellis R. Scale development: theory and applications. 2nd ed. Thousand Oaks: Sage Publications; 2003.
11. Vogt W. Dictionary of statistics and methodology: a nontechnical guide for the social sciences. 2nd ed. Thousand Oaks: Sage Publications; 1999.
12. Aday L, Cornelius L. Designing and conducting health surveys: a comprehensive guide. 3rd ed. San Francisco: Jossey-Bass; 2006.
13. McDowell I. Measuring health: a guide to rating scales and questionnaires. 3rd ed. New York: Oxford University Press; 2006.
14. Bowling A. Mode of questionnaire administration can have serious effects on data quality. J Public Health. 2005;27:281–91.
15. Dillman D. Mail and internet surveys: the tailored design method. 2nd ed. New York: Wiley; 2000.
16. 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.
17. Cole E, Pisano ED, Clary GJ, Zeng D, Koomen M, Kuzmiak 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.
18. 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.
19. Freedman M, Lester K, et al. Cell phones for ecological momentary assessment with cocaine-addicted homeless patients in treatment. J Subst Abuse Treat. 2006;30:105–11.
20. 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.
21. 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.

Chapter 12
Biobanking Challenges and Informatics Opportunities

Elizabeth J. Horn and Sharon F. Terry

Abstract Biobanking is the science and practice of storing biological specimens for future use. Biobanking is an emerging field with the potential to improve our understanding of disease and develop better, more targeted treatments for many conditions. Data associated with the specimens must include information about the specimens, the donor, and the conditions (including informed consent) under which the samples were collected, processed, and stored. Biobanking is based upon the premise that the storage of biologic specimens will enable future research, including the use of advanced technologies and methods beyond what currently exists, and without associated data, samples cannot be leveraged for the future. With the completion of the human genome and the promise of personalized medicine and diagnostics, biobanking is being embraced by a variety of stakeholders, including academic institutions, government, industry, and patient advocacy groups. This wide-ranging adoption has led to the development of many biobanks for various purposes. These different categories of biobanks, from population biobanks to disease-specific biobanks, collect a variety of human specimen types, each requiring different descriptive data and associated standards for collection, processing, and storage. In this chapter, we discuss the challenges inherent in biobanking and opportunities for informatics to resolve some of these challenges.

Keywords Biobanking • Informatics • Biobank data • Biospecimen science • Best practices • Biorepository • Regulatory issues • Biobanking standards • Biomolecules

E.J. Horn, Ph.D., MBI • S.F. Terry (✉)
Genetic Alliance Registry & BioBank,
Genetic Alliance, 4301 Connecticut Avenue, NW, Suite 404,
Washington, DC 20008, USA
e-mail: lhorn@geneticalliance.org; sterry@geneticalliance.org

R.L. Richesson, J.E. Andrews (eds.), *Clinical Research Informatics*, Health Informatics,
DOI 10.1007/978-1-84882-448-5_12, © Springer-Verlag London Limited 2012

A tremendous amount of resources and energy is being invested in biobanking worldwide, and the research potential of biobanks has impressed scientists, policy makers, and the public. In 2009, *Time* Magazine named biobanking as one of ten ideas changing the world [1]. While there is great excitement surrounding the science of biobanking, considerable challenges do exist. The most significant challenges relate to adequate acquisition and preservation of biosamples and to the collection of descriptive data associated with the sample and with the individual donating the sample. Samples must be collected and preserved appropriately and consistently, with relevant data collection to support valid and reproducible research. As scientific research techniques improve, increasingly smaller quantities of materials are needed for very sophisticated techniques. While this modernization has the potential to produce unprecedented research advances with high-quality samples, it also has the potential to produce vast amounts of data from low-quality samples. In other words, we now have the ability to get answers that are potentially inaccurate at unprecedented speed.

Standards

Important concerns in the field of biobanking lie in the realm of standardization. Standards are needed for sample collection methods, specimen processing and storage, and associated clinical data. Standards for each of these areas are limited, and for the few that exist, widespread adoption has not yet occurred. Ongoing efforts in biobanking and biospecimen science are aimed at initiating the development and adoption of standards. The application of informatics tools and techniques can support best practices for sample acquisition and preservation, and thereby reduce variation and increase reproducibility. Informatics can also enable explicit representation of sample characteristics and associated clinical data, enabling innovative and novel use of samples in the future. Therefore, informatics applications have the potential to improve not only the information associated with samples, but also the quality of biobanking resources and their relevance to the clinical research process.

Significance, Relevance, and Challenges of Biobanking

The Need for High-Quality Biospecimens

Following the completion of the sequencing of the human genome, personalized medicine has been heralded as the future of medicine [2]. Tailoring therapies to individuals requires the collection, analysis, and molecular characterization of biological specimens. High-quality, well-annotated biospecimens are needed for identification of targets for drug development, treatment, and prevention; for identification of biologic variations that determine drug efficacy and drug toxicity;

for defining biomarkers for susceptibility, screening, and reoccurrence of cancer; and for the elucidation of molecular mechanisms and validation of therapeutics [3]. For these and other scientific endeavors, accurate, reproducible data derived from patient samples in the clinical setting will be essential.

Biobank Landscape

Traditionally, biobanks have been created for genetic research with relatively simple sample acquisition and storage and with straightforward hypothesis testing. As science advances, more sample types are being collected for more sophisticated analysis by a variety of stakeholders and organizations. Academic institutions have developed biobanks of pathology archives to support clinical care and research, usually within that institution. They have also begun to build large, institution-wide, repositories enabling activities from biomarker development to drug discovery [4]. Some have also created collaborative collections used in clinical trials for both diagnostics and drugs and have even created new technologies based on these collections [5]. Commercial vendors have developed collections of samples and data that can be purchased by investigators. Some collections also include robust, annotated data with available biospecimens. Large, national, population biobanks are being developed in Estonia, Sweden, the UK, and the USA [6]. Population biobanks follow cohorts of people, including healthy individuals, over many years. For example, the UK Biobank is expected to follow 500,000 citizens over several decades, continually updating their health history information [7]. Other biobanks are disease specific, following individuals with a specific condition over time. In some instances, disease advocacy organizations have established and are managing biobanks [8]. The Genetic Alliance BioBank is one example where seven advocacy organizations representing rare or common diseases utilize shared infrastructure in a cooperative model [9]. The biobanks mentioned collect a variety of sample types, including blood, urine, saliva, tissue, and organs from full-body harvests. Many also process and store these sample types as derivatives such as DNA, RNA, and protein.

The Challenge of Acquiring High-Quality Samples

The lack of high-quality human specimens has become the limiting factor for post-genomic biomedical science and is a major roadblock to translational research and personalized medicine [10, 11]. By and large, researchers continue to work in silos where little specimen sharing exists, and each institution has its own program comprised of individual studies with small numbers. Dr. Carolyn Compton, Director of the Office of Biorepositories and Biospecimen Research (OBBR), described the challenges of acquiring high-quality samples at a scientific conference sponsored by the NIH Office of Rare Diseases Research in January, 2010 [12]. In the current culture, many investigators do not believe they can get the quantity or quality of samples needed for their research. Scientists also reported questioning their data due to the unreliable quality of available specimens, and many limit the scope of their work due to the low of quality of biospecimens. If biospecimens of poor or unknown quality are utilized, the accuracy and validity of the research data can be compromised.

Acquiring high-quality specimens and data is challenging because the collection, processing, and storage procedures are not standardized and because the degree and type of data annotation varies. Quality control may not be built into the collection process, and biobanks may overestimate the quality of samples they have collected. This is further complicated by the many different kinds of samples collected for distinct purposes, with each sample type having its own unique best practices. Collection practices are highly variable within and among institutions, and specimen quality is determined by the medical system, not scientific users. Most tissues procured for research purposes are leftover from surgery or autopsy, and misclassification and degradation of the sample are often issues of concern. Diagnosis may not be known until after pathology assessment, compromising the opportunity to collect a sample prospectively for research. The size of samples needed for diagnosis is also becoming smaller, and the more difficult and complicated the diagnosis, the greater the depletion of the sample for testing and the less remaining for subsequent research. Finally, the collection of normal control samples is not routine, making any comparison between diseased and unaffected tissue problematic.

Tissue collection itself is complex, in part because molecules that reflect the disease state are extremely labile. In an ideal collection environment, tissue is frozen within minutes of blood supply disruption, reducing the opportunity for biomolecules such as RNA and proteins to degrade. In standard clinical procedure, tissue is not usually frozen rapidly, but procured after clinical requirements have been satisfied. This makes determining the molecular quality of biospecimens difficult, and histology is not a good indicator of molecular quality. Formalin-fixed paraffin-embedded tissues (FFPE) are often available, but analyses performed on FFPE are limited, and biomolecules are highly degraded by processing; however, technologies are emerging to overcome these challenges [5]. There is also a cost to the institution to procure these samples, and many of the activities required to collect biospecimens are nonreimbursable.

Governance

To add to the complexity, governance practices vary by biobank, with variable levels of transparency. In general, patients give consent with differing levels of permission for sample storage, use, and recontact for additional information or samples. Policies for data access, sample access, incidental findings, and returning results to participants may not exist, may be inadequate, or may be unknown to participants. In addition, material transfer agreements often differ between institutions, and medical institutions may be hesitant to release specimens for research if they were procured for diagnostic use. Complying with regulations can be burdensome for individual investigators, investigators new to human subjects research, and institutions without robust infrastructure. Supporting biorepository IT structures differ in capacity and functionality with little standardization, and there are not uniform, agreed upon standards for clinical data. Extraction and transfer of associated clinical data may be extremely laborious, quite costly, and inhibited by regulatory issues including those concerning privacy. The variability in consent information makes using samples from multiple institutions problematic, as consent is controlled at the institutional level. Historically collected samples are a pervasive problem, since it is not always clear what secondary uses the samples might support. Efforts are underway to standardize consent practices to allow for future use and sharing of samples [13].

Timing

Finally, the timing of biospecimen collection poses unique challenges. Diseases manifest themselves over time, and obtaining early and late stage biospecimens is often difficult. Recruitment can also be difficult when there are limited individuals with the condition or a limited number of investigators studying the disease, such as in rare disease research. It is impossible to anticipate all the data needed for biospecimens when collected prospectively. Getting medical centers to comply with data requests is time-consuming, and while participants can be helpful in obtaining clinical records, it can be challenging if requests are highly technical or if the donor is deceased. Taken together, these conditions lead to a wide variation in quality of biospecimens and associated clinical data.

Informatics Opportunities for Improving Biospecimen Quality

In most instances, the scientific community cannot anticipate the scientific questions that might be asked in the future on the samples being collected today. It is difficult to prepare now for a future that is unknown. Regardless, planning and standardization can facilitate the development of resources to support future research and discovery. While a daunting task, there are great opportunities for informatics applications to improve the state of biobanking and biospecimen science, particularly through the adoption best practices and standards described below.

Biobanking Best Practices and Supporting Initiatives

Best practices are needed to provide state-of-the-science guidance for biobanking and to harmonize procedures for collection, processing, storage, and distribution of biospecimens. Multiple best practices exist but there is not yet uniform adoption of these protocols. The National Cancer Institute (NCI) has developed comprehensive best practices, first published in 2007 and revised in 2010, that examine the scientific evidence for collection, annotation, processing, and storage of biospecimens [14]. Included in NCI's best practices are technical and operational best practices and ethical, legal, and policy best practices. The International Society for Biologic and Environmental Repositories (ISBER) has also developed best practices for biorepositories, focusing on the collection, storage, retrieval, and distribution of biological materials for research [15].

A number of initiatives have been developed to improve standardization of biospecimen quality and informatics to support best practices. In the USA, the Cancer Human BioBank (caHUB) was initiated as a national infrastructure for translational research with evidence-based collection strategies [16]. It was developed to provide a centralized source for both cancer specimens and normal human specimens, as well as for tools, training opportunities, and other biospecimen resources. In mid-2011, caHUB became a center for Biospecimens and Standards. The Tissue Banks and Pathology Tools Workspace (TBPTW) of the Cancer Biomedical Informatics Grid, caBIG [17], includes the following tools relevant to biobanking: caTISSUE Core for managing biospecimens, caTISSUE Clinical Annotation to annotate biospecimens with clinical data, and cancer Text Information Extraction System (caTIES) to extract concepts from free-text pathology reports into a structured data model. It is unclear as of this writing how or to what extent these tools will be supported and available, but certainly they represent relevant and important areas of thought and standardization for biospecimens and future research. In Europe, Biobanking and Biomolecular Resources Research Infrastructure (BBMRI) is a coordinated infrastructure that provides access to a Europe-wide collection of biomedical samples and data [18]. BBMRI is composed of a network of biobanks with different formats and biomolecular resources. BBMRI utilizes standards for sample collection, storage, preanalytics and analysis, and a harmonized database and computing infrastructure, and it also includes guidance on ethical, legal, and social issues. Another useful resource is the ISBER self-assessment tool, a confidential, 158-question assessment designed to help biorepositories strengthen their practices by identifying areas that need improvement [19].

Standards Surrounding Specimen Collection and Storage

The use of a biospecimen in the future requires that the user understand the basis of its collection. How was the sample collected? What were the patient characteristics at the time of collection? How was the sample transported? These are examples of preanalytic variables. These variables start with the patient and include the medical/surgical procedures and acquisition of samples. Preacquisition variables include antibiotics, other drugs and treatments, type of anesthesia, duration of anesthesia, and arterial clamp time. Postacquisition variables include time at room temperature,

temperature of the room, type of fixative, time in fixative, rate of freezing, and time of aliquots. Once the sample has been acquired, handling/processing, storage, distribution, scientific analysis, and restocking unused sample may all affect the integrity of the sample. For example, changes in specific transcript levels may be based on the ischemic time and not the disease. The inability to reproduce protein biomarkers has been seen in discovery research as well as inconsistent immunohistochemistry (IHC) results in research and clinical labs. In metabolomics, the potential for error is greater, where inconsistencies in small molecule readouts may yield results that point to incorrect pathways. Ideally, every piece of relevant data needs to be collected to support future users who may have no connection to or understanding of the specimen collection protocols. Research is also needed to better understand how these variables affect molecular integrity, as some variables will have great influence on molecular pathways and others will not.

The Office of Biospecimen and Biorepositories Research (OBBR) has developed Biospecimen Reporting for Improved Study Quality (BRISQ) to guide researchers to capture information about the source and handling of biospecimens with a goal of making research results more reproducible [20]. BRISQ elements have three tiers of reporting. Tier 1 includes items necessary to report, such as organs from which the biospecimens were derived and the manner in which biospecimens were stabilized and preserved. Tier 2 levels are items advisable to report but are less crucial or less likely to be available in the annotations, such as demographics of the patient population and methods of enrichment for relevant components. Tier 3 includes additional items about conditions that are not as likely to influence research results or are unlikely to be available to researchers. These include environmental factors to which patients were exposed and the type of storage containers in which biospecimens were kept. BRISQ elements are captured as preacquisition, acquisition, stabilization/preservation, storage/transport, and quality assurance measures. Critical unknown elements should also be fully acknowledged.

The ISBER Biospecimen Science Working Group has also been working to improve classification schemas for biospecimens. The Standard PREanalytical Coding for biospecimens (SPREC) was developed to improve biospecimen research experimental protocols and to provide information about the biomolecular quality of samples [21]. SPREC codes are seven element-long biospecimen characterization codes that give details on preanalytical sample processing. They are available for primary samples (those specimens directly collected from the donor) and for simple derivatives (samples prepared by simple laboratory manipulation). SPREC codes are flexible and easy to implement. Complex derivatives are out of scope including cell disruption, cell selection, and multistep chemical manipulation such as acidification, digestion, precipitation, deproteinization, and desalting.

The Standardisation and Improvement of Generic Pre-Analytical Tools and Procedures for In Vitro Diagnostics consortium, better known as SPIDIA, has also been working in this area, developing pan-European quality assurance schemes and guidelines for preanalytical procedures such as sample collection, handling, transportation, processing, and storage of clinical samples [22]. SPIDIA is developing guidelines for processing of blood, tissue, RNA, DNA, and proteins; developing new tools and technologies that integrate and standardize preanalytical steps; and identifying appropriate biomarkers for monitoring changes in clinical samples, including RNA, DNA, proteins, and metabolites. Other European initiatives of interest include the European Network to Promote Research into Uncommon Cancers in Adults and Children: Pathology, Biology, and Genetics of Bone Tumours (EuroBoNeT) and BBMRI [23].

Standards in Clinical Data Associated with Samples

Robust clinical information enhances the value of biological samples, enabling the correlation of phenotype with the research being conducted on a particular sample. The availability of associated clinical information is often variable, ranging from basic demographics to very detailed patient histories. Data collection is expensive. The variability of the available data can be compounded if questions and measured values change over time in subtle ways or if participants are lost before the necessary follow-up. Efforts are ongoing for standardizing how questions are asked and how responses are obtained. Standardization will minimize some of the work of data collection, will increase the likelihood that data collected today will be compatible with data collected in the future, and will facilitate the pooling of data between related registries and studies.

One way of standardizing data is to use scientifically validated survey instruments. Because people's health and disease change over time, participant information cannot be static and must be collected and updated at multiple time points. If validated survey instruments do not exist for the information that needs to be collected, using questions that others have used on a large scale is recommended. The National Health and Nutrition Examination Survey (NHANES) [24], the Patient-Reported Outcomes Measurement Information System (PROMIS) [25], the consensus measures for Phenotypes and exposures (PhenX) project [26], and the database of Genotypes and Phenotypes (dbGaP) [27] are possible sources for questions and answers. The Patient Registry Item Specifications and Metadata for Rare Diseases (PRISM) project is collecting and cataloging questions that have been used by others to develop a library of standardized questions across a broad spectrum of rare diseases [28]. Utilizing messaging standards, such as Health Level 7 (HL7) [29], Logical Observation Identifiers Names and Codes (LOINC) [30], or Systematized Nomenclature of Medicine-clinical Terms (SNOMED CT) [31], is another important aspect of standardizing data. HL7 is the standard messaging for the delivery of many types of clinical and laboratory results. LOINC provides codes for lab tests, clinical measures, diagnostic reports, and surveys, and SNOMED CT provides a comprehensive clinical terminology. The National Library of Medicine (NLM) supports LOINC and SNOMED CT, along with RxNorm [32], which provides codes at the clinical drug and ingredient level. The use of such data standards will be essential for sharing data across multiple platforms, including registries and electronic health records (EHRs). Data standards will also support sharing information from different institutions or over time.

Laboratory Information Management Systems

Laboratory information management systems (LIMS) are necessary to track all aspects of the workflow in a laboratory, including instruments, samples, personnel, and quality assurance and quality control procedures. LIMS are the primary method of managing samples in a biobank, and the record produced by LIMS provides important information about what happens to the sample while it is in the laboratory. When a sample is received, it is accessioned, and a barcode is affixed to the tube for identification. This is used to track the chain of custody, the location of the sample, and all events associated with the sample, such as how the sample is processed, the number of freeze-thaw cycles, and shipment of the sample to investigators. ISBER provides guidance on the information that should be collected [15], and a variety of commercial and open source LIMS are available. Standards used with LIMS include Title 21 CFR Part 11 from the United States Federal Drug Administration, Good Laboratory Practice (GLP) from the Organisation for Economic Co-operation and Development, and ISO/IEC 17025 from the International Organization for Standardization and the International Electrotechnical Commission.

Managing Access to Data and Samples

Information systems are also needed to manage access to samples and data for future research. Biospecimen locators and online catalogs are needed for investigators to identify samples and data needed for their research. Tools are also needed for biobanks to manage sample and data access requests and permissions, and in some instances, return results to participants. These systems must be used in cooperation with governance models to track how existing samples can be used in the future, as dictated by the original informed consent. In some instances, the participant has given blanket consent for all future use, but in other cases, there are restrictions in how the sample can be used. Biobanks' understanding how the samples they have collected can be used and their honoring any restrictions of future use are imperative for public trust. Information systems can be used to facilitate data and sample stewardship, keeping the record of rules and permissions surrounding future use for data and samples. Some of these systems may also track patient preferences and consent; however, this type of information is usually kept separate from information about sample collection and processing discussed above.

Future Directions in Informatics and Biobanking

Biobanking is an emerging field that has the potential to usher in the era of personalized medicine and diagnostics, improving individualized health outcomes.

Informatics tools are needed for biobanking to reach its full potential. Scientists and policy makers have learned that the way in which studies are designed, specimens are handled, assays are performed, data are analyzed, and conclusions are stated *does* matter and that much of this essential information is not readily available. Biospecimen science is helping to identify the best methods to collect and manipulate samples, and to identify standards for sample collection, storage, and representation of associated clinical data – all of which are essential for users to assess the quality of biospecimens and data. The biobanking community's adoption and implementation of agreed upon standards to improve the quality of samples and data and the reproducibility of scientific results are imperative. Much work has already been done in identifying appropriate standards, but more work is needed to encourage adoption of these standards. Shared information technology infrastructure and open source platforms are needed to aggregate sample and associated clinical data from multiple sources. There are also opportunities to use informatics to manage ethical, legal, and social issues. With cooperation and common vision, standards for high quality biospecimens can be developed, adopted, and implemented. These standards can help merge the separate silos that currently exist in research and, ultimately, advance the field of personalized medicine and diagnostics.

References

1. Park A. 10 ideas changing the world right now. Time. 2009. http://www.time.com/time/specials/packages/article/0,28804,1884779_1884782_1884766,00.html. Accessed 15 Oct 2010.
2. Hamburg MA, Collins FS. The path to personalized medicine. N Engl J Med. 2010;363:301–4. doi:10.1056/NEJMp1006304.
3. Troyer D. Biorepository standards and protocols for collecting, processing, and storing human tissues. Methods Mol Biol. 2008;441:193–220. doi:10.1007/978-1-60327-047-2.
4. Pendergrass S, Dudek SM, Roden DM, Crawford DC, Ritchie MD. Visual integration of results from a large DNA biobank (biovu) using synthesis-view. Pac Symp Biocomput. 2011;265–275. doi:10.1142/9789814335058_0028.
5. Paik S, Kim CY, Song YK, Kim WS. Technology insight: application of molecular techniques to formalin-fixed paraffin-embedded tissues from breast cancer. Nat Clin Pract Oncol. 2005;2:246–54. doi:10.1038/ncponc0171.
6. Austin MA, Harding S, McElroy C. Genebanks: a comparison of eight proposed international genetic databases. Community Genet. 2003;6:37–45. doi:10.1159/000069544.
7. Hawkins AK. Biobanks: importance, implications and opportunities for genetic counselors. J Genet Couns. 2010;19:423–9. doi:10.1007/s10897-010-9305-1.
8. Terry SF, Terry PF, Rauen KA, Uitto J, Bercovitch LG. Advocacy groups as research organizations: the PXE international example. Nat Rev Genet. 2007;8:157–64. doi:10.1038/nrg1991.
9. S.F. Terry, E.J. Horn, J. Scott, P.F. Terry. Genetic Alliance Registry and BioBank: a novel disease advocacy-driven research solution. Personalized Medicine, 8(2):207–13, March 2011
10. Compton CC. The surgical specimen is the personalized part of personalized cancer medicine. Ann Surg Oncol. 2009;16:2079–80. doi:10.1245/s10434-009-0526-1.
11. Lim MD, Dickherber A, Compton CC. Before you analyze a human specimen, think quality, variability, and bias. Anal Chem. 2011;83:8–13. doi:10.1021/ac1018974.

12. Compton CC. Rare disease biospecimens: quality and accessibility challenges. Advancing rare disease research: the intersection of patient registries, biospecimen repositories, and clinical data. 2010. http://rarediseases.info.nih.gov/PATIENT_REGISTRIES_WORKSHOP/. Accessed 15 Oct 2010.
13. Beskow LM, Friedman JY, Hardy NC, Lin L, Weinfurt KP. Developing a simplified consent form for biobanking. PLoS One. 2010;5:e13302. doi:10.1371/journal.pone.0013302.
14. Office of Biorepositories and Biospecimen Research. NCI best practices for biospecimen resources. National Cancer Institute. 2010. http://biospecimens.cancer.gov/bestpractices/. Accessed 15 Oct 2010.
15. Pitt KE, Cambell LD, Skubitx APN, Somiari SB, Sexton KC, Pugh RS. Best practices for repositories: collection, storage, retrieval and distribution of biological materials for research. Cell Preserv Technol. 2008;6:5–58. doi:10.1089/cpt.2008.9997.
16. Office of Biorepositories and Biospecimen Research. caHUB: the cancer human biobank. National Cancer Institute. 2010. http://cahub.cancer.gov/. Accessed 15 Oct 2010.
17. National Cancer Institute. caBIG: cancer Biomedical Informatics Grid. 2010. http://cabig.cancer.gov/. Accessed 10 Dec 2010.
18. BBMRI: Biobanking and Biomolecular Resources Research Infrastructure. Managing resources for the future of biomedical research. 2010. http://www.bbmri.eu/. Accessed 15 Oct 2010.
19. International Society for Biological and Environmental Repositories. ISBER self-assessment tool (SAT). 2009. https://secure.asip.org/CVWEB_ISBER/Library/sat.htm. Accessed Aug 2011.
20. Kelly A. Biospecimen reporting for improved study quality: BRISQ. caBIG: cancer Biomedical Informatics Grid. 2010. https://cabig.nci.nih.gov/workspaces/TBPT/Meetings/TBPT_Workspace/19July10_TBPT. Accessed 15 Oct 2010.
21. Betsou F, Lehmann S, Ashton G. Standard preanalytical coding for biospecimens: defining the sample PREanalytical code. Cancer Epidemiol Biomarkers Prev. 2010;19:1004–11. doi:10.1158/1055-9965.EPI-09-1268.
22. SPIDIA Consortium. SPIDIA: standardisation and improvement of generic pre-analytical tools and procedures for in-vitro diagnostics. 2011. http://www.spidia.eu/. Accessed 15 Oct 2010.
23. Riegman PH, de Jong BW, Llombart-Bosch A. The Organization of European Cancer Institute Pathobiology Working Group and its support of European biobanking infrastructures for translational cancer research. Cancer Epidemiol Biomarkers Prev. 2010;19:923–6. doi:10.1158/1055-9965.EPI-10-0062.
24. Office of Information Services. National health and nutrition examination survey. Centers for Disease Control and Prevention. 2010. http://www.cdc.gov/nchs/nhanes.htm. Accessed 10 Dec 2010.
25. National Institute of Health. PROMIS: patient-reported outcomes measurement information system. 2010. http://www.nihpromis.org/default.aspx. Accessed 10 Dec 2010.
26. RTI International. PhenX: consenus measures for phenotypes and exposures. 2010. https://www.phenx.org. Accessed 10 Dec 2010.
27. National Center for Biotechnology Information. dbGaP: database of genotypes and phenotypes. 2011. http://www.ncbi.nlm.nih.gov/gap. Accessed 10 Dec 2010.
28. PRISM: Patient registry item specifications and metadata for rare diseases. http://prism.epi.usf.edu/prism/. Accessed 10 Dec 2010.
29. HL7: Health Level Seven International. 2011 http://www.hl7.org/. Accessed 10 Dec 2010.
30. Regenstrief Institute, Inc. LOINC: logical observation identifiers names and codes. 2011. http://loinc.org/. Accessed 10 Dec 2010.
31. International Health Terminology Standards Development Organisation. SNOMED CT. 2010. http://www.ihtsdo.org/snomed-ct/. Accessed 10 Dec 2010.
32. U.S. National Library of Medicine. RxNorm. Unified Medical Language System (UMLS). 2010. http://www.nlm.nih.gov/research/umls/rxnorm/. Accessed 10 Dec 2010.

Chapter 13
Patient Registries*

Rachel L. Richesson and Kendra Vehik

Abstract Patient registries are fundamental to the research process. Registries provide consistent data for defined populations and can support the study of the distribution and determinants of various diseases. One advantage of registries is the ability to observe caseload and population characteristics over time, which might facilitate the evaluation of disease incidence, disease etiology, planning, operation and evaluation of services, evaluation of treatment patterns, and diagnostic classification. Registries can be developed for many different needs, including research recruitment, study planning, public health, and observational research. Any registry program must collect high-quality data to be useful for its stated purpose. We describe the methodological issues, limitations, and ideal features of registries to support various purposes. The future impact of registries on our understanding and interventions for many diseases will depend upon technological and political solutions for global collaborations to achieve consistent data (via standards) and regulations for various registry applications. The development, implementation, interpretation, and evaluation of registries are areas that can benefit from informatics expertise and coordination.

Keywords Registries • Public health surveillance • Clinical research • Epidemiology • Rare diseases • Patient registry • Observational research

*Adapted from Richesson RL, Vehik K. Patient Registries: Utility, Validity and Inference. In: Posada M and Groft SC (ed). Rare Diseases Epidemiology, The Netherlands, Springer, 2010, with kind permission of Springer Science + Business Media.

R.L. Richesson, Ph.D., MPH (✉)
Division of Informatics, Duke University School of Nursing,
311 Trent Drive, Durham, NC 27710, USA
e-mail: rachel.richesson@dm.duke.edu

K. Vehik, Ph.D., MPH
Department of Pediatrics, Pediatrics Epidemiology Center, University of South Florida,
3650 Spectrum Boulevard, Suite 100, Tampa, FL 33612, USA
e-mail: kendra.vehik@epi.usf.edu

R.L. Richesson, J.E. Andrews (eds.), *Clinical Research Informatics*, Health Informatics,
DOI 10.1007/978-1-84882-448-5_13, © Springer-Verlag London Limited 2012

Patient registries are a fundamental part of research and have been for centuries. Although gathering, observing, and following populations of individuals have long been the important steps in the understanding of any disease etiology, the use of registries is taking on new importance in clinical trials support, effectiveness research, and patient safety. The FDA and health advocacy groups in the United States and the European Union, including those explicitly targeting improved effectiveness and efficiency of the drug development process, are advising patient advocacy organizations to create registries as a primary step in advancing research for understudied diseases. Even before drug development, patient registries can facilitate research by clearly documenting the natural course of disease and providing metrics for comparisons between alternate therapies. In rare and neglected diseases, registries are a critical research step and necessity to identify potential patients for clinical studies. Such "preresearch" registries provide a qualitative picture of disease burden and complications and are often sponsored by patient advocacy groups, rather than academia or industry. In common diseases, the registry data can truly support evidence-based medicine and comparative research. Cystic fibrosis (CF) has had a registry for 40 years that has supported various studies which collectively contributed to tripling the life expectancy for patients with CF. Diabetes has an extensive network of registries in Europe that have been in existence for over 30 years; more recently, the United States has started to develop such a network, primarily for monitoring diabetes in the young. The cumulative effect of these registries has provided evidence of a rapid increasing trend in both type 1 and type 2 diabetes worldwide with the greatest increases in the very young. This has provided impetus to increase research funding to both determine preventative strategies and therapeutic interventions to reduce the complications of this debilitating disease.

A search on ClinicalTrials.gov (which likely excludes most nonresearch registries hosted by patient support and advocacy organizations) will show more than 700 registries, and a PubMed search on treatments in many diseases will show the role of registries is identifying genes, comorbidities, life expectancy, and quality of life. The numbers of registries are expected to grow over the next few decades – largely due to relative ease of both design and maintenance owed to innovative information technology. Changing roles of patients and successful business models for patient advocacy groups are facilitating the development of patient registries to support research activities and funding. While registries are an important tool for clinical research, the diversity of sponsors and objectives has lead to confusion regarding the legal and operational definitions, their subsequent evaluation, and the best practices for the use and interpretation of registry data.

Definition

A patient *registry* is defined as an organized program for the collection, storage, retrieval, and dissemination of a clearly defined set of data collected on identifiable individuals for a specific purpose; the collected data are termed *patient registry*

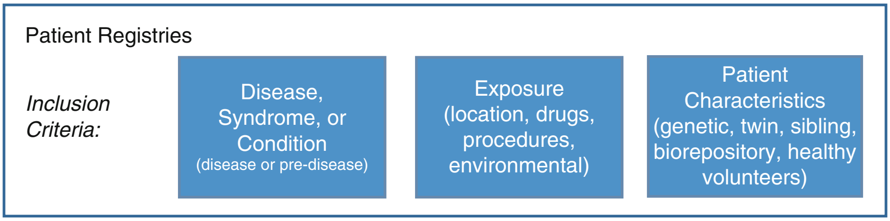

Patient Registries
Inclusion Criteria:
Disease, Syndrome, or Condition
(disease or pre-disease)
Exposure (location, drugs, procedures, environmental)
Patient Characteristics (genetic, twin, sibling, biorepository, healthy volunteers)

Fig. 13.1 Types of patient registries

data. This extends previous registry definitions by viewing a registry as not only a database but also as a systematic data collection program [1–3]. Although there are various definitions for patient registries in the public health literature, there is a general consensus that the term registry implies follow-up and change in status of cases over time [4, 5].

As shown in Fig. 13.1, patient registries have three broad types of inclusion criteria: disease (or condition or syndrome), exposure (e.g., medical or surgical treatment, medical devices, environmental), and patient characteristics (e.g., genetic, twin, sibling, healthy controls). Disease and exposure registries are the most common types of registries, but the number of patient characteristics–based registries is increasing each year due to a surge of new genetic registries. The annotated data records associated with biological repositories ("biobanks") also can be thought of as registries of patient characteristics (usually genetic) with a biological data collection component – and the presence of these collections is growing rapidly [6–9].

History

The first known disease registries go back several hundred years with registries in leprosy and tuberculosis [10–12]. The emergence of chronic diseases has sparked a persistent proliferation of patient registries since the 1950s [4]. A recent review article found over 43,000 articles in the scientific literature (2000) referring to registries [13]. Cancer-specific registries grew explosively from 32 registries in 1966 to 449 cancer registries representing five continents in 2006 [14].

Patient registries are often the first step in estimating prevalence or incidence and building a cause for future research and facilitating enrollment in trials. Genetic sequencing has lead to the identification of new diseases, which in turn has spawned the creation of numerous disease-specific patient advocacy groups that demand funding new disease-specific registries in rare diseases. While registries were first born from government departments to support core public health functions, many successful and large registries have since been established by patient-driven organizations. The creation of a registry is not merely a rite of passage to get a disease "on the map" for funding priority but has become a fundamental early step in the understanding of the natural history of disease, development of clinical endpoints, monitoring trends, patient-reported outcomes, and baseline data to support formal evaluations of therapeutic interventions. In the United States, the NIH and the FDA are actively recruiting special interest groups to develop registries in parallel to the identification of disease assays and drug compounds [15].

Concerns about the safety of new drugs (especially biologics with uncertain long-term outcomes – e.g., thalidomide, human growth hormone), and desire for large-scale, real-world safety and efficacy data on marketed drugs (as well as combination therapies), have fueled the growth of patient registries for use in postmarketing activities. The use of registries for postmarket monitoring (phase 4 studies) of approved drug products has increased in recent years. Under the Food and Drug Administration Amendments Act of 2007 (FDAAA) in the United States, the FDA can mandate postapproval requirement studies and risk mitigation and evaluation systems (REMS) as a condition of approval for new products with potential safety issues [16].

Characterization of Registries, Their Uses, and General Requirements

The unending proliferation of registries and the need for global research cooperation create a situation desperate for standards and best practices for patient registry projects [17]. The large number of registries, and the various purposes and stakeholders for each, complicates any attempts to inventory, standardize, or prescribe good design features for patient registries in general. There have been a few attempts to characterize types of registries by their data source [local hospital, regional (~multiple hospitals), and population-based (~multiple data sources) [18]] or by the database and data characteristics [13]. Characterizations of registries by purpose may simply delineate registries as either clinical or research [4], or by more detailed purposes [19] which inspired the characterization we present in this chapter. Others consider the manifold impact of registries as supporting the classic medical school triad of research, service, and teaching [20].

We present a characterization of registries by purposes and suggest some essential requirements to support various purposes. As displayed in Table 13.1, registry uses can fall into six (nonexclusive) categories of usage: public health, health services research, health promotion, patient care, clinical research, and regulatory (public safety). Based upon the primary purpose of the registry, the columns depict whether the selected primary registry function necessarily dictates an *absolute* requirement for: completeness of case ascertainment, extensive clinical data, verification of data validity, and follow-up. The table is designed to indicate which types of registries need this to fulfill stated functions in any capacity. For example, while clearly verification of data validity and completeness of case ascertainment is a desirable feature for any registry, for some purposes (e.g., the use of registry for scientific or epidemiologic investigation), the verification of data and assurance of complete case capture is of utmost importance, where as in other applications, such as advertising for clinical trials, the lack of data verification or incomplete case

Table 13.1 Purpose of registry and essential requirements

Essential requirements				
Purpose	Completeness of case ascertainment	Clinical data[a] (beyond diagnosis or procedure)	Verification of data validity	Follow-up data
Public health ("population-based")				
Population surveillance	Yes	No	Yes	No
Contact notification	Yes	No	No	No
Patient compliance (for management of infectious diseases)	Yes	Yes	Yes	Yes
Planning (community and service)	Yes	No	No	No
Policy	Yes	No	No	No
Health services research				
Evaluation of healthcare/education delivery	Yes	Yes	Yes	No
Facilitate health utilization treatment patterns	Yes	No	Yes	Yes
Monitoring health services	Yes	No	No	No
Measuring healthcare quality	No	Yes	Yes	Yes
Health promotion tools and education				
Patient education notifications	No	No	No	No
Physician education notifications	No	No	No	No
Aggregate data for patient education/support	No	No	No	No
Patient care				
Chronic disease management	No	Yes	Yes	Yes
Vaccination	Yes	No	No	Yes
Clinical research – funding and support				
Research funding decisions	No	No	No	No
Research planning and design	No	No	No	No
Cohort selection	No	Yes	Yes	No
Recruitment – outreach to patients	No	No	No	No
Clinical research – scientific inquiry				
Cross-sectional	Yes	Yes	Yes	No
Longitudinal	Yes	Yes	Yes	Yes
Regulatory				
Safety of agents (postmarketing)	Yes	Yes	Yes	Yes
Efficacy of agents (postmarketing; phase 4)	Yes	Yes	Yes	Yes

[a] Clinical data – additional data beyond the data elements required for determining eligibility for the registry. Eligibility is determined either by disease, exposure, or patient characteristics

ascertainment does not impede the registry objectives. There are many data quality and bias issues, mostly related to case ascertainment, data validity, and follow-up, which limit the utility of registry data for various purposes. In the next section, we describe the major limitations and biases associated with patient registries.

Data Quality, Bias, and Limitations of Patient Registry Data

Developers of registries and potential users of registry data must be keenly aware of the inherent limitations of certain registry designs for certain functions, particularly in the exploration of research questions involving treatment evaluation. A registry must have high-quality data to be useful for any research purpose. Two fundamental concerns related to gauging the quality of registry data include completeness of case ascertainment and validity of values for each data point [4]. Timeliness of data has also been noted as a quality indicator [14]. For registries requiring follow-up data, the proportion of follow-up obtained and the nature of cases lost to follow-up must be provided and considered for any interpretation of registry data.

Completeness of Registry

Disease registries for epidemiologic purposes are largely designed to ascertain cases of a specific disease for public health surveillance and planning. The primary metrics used are incidence and prevalence of a disease. Completeness of case ascertainment for infrequent or rare disorders is an essential measure to determine the accuracy of the true incidence or prevalence in a population. The idea behind any registry endeavor is that the registry is a tool to either count or characterize health or disease characteristics in a *sampled* population, with the intent to extrapolate those results back to a larger or different population. The completeness of case ascertainment (i.e., the inclusion of all cases in the sample area time or place) therefore has implications for the conclusions and the extrapolations made to the general population.

The capture-recapture methodology has long been the “gold standard” for determining completeness of case ascertainment. Originally, this method was first used in wildlife biology to study fish and wildlife populations [21, 22]. The simplistic model was used to estimate the unknown size of ecological populations. In human populations, the capture-recapture methodology still utilizes the two-mode ascertainment model (e.g., physician provider versus hospital data), although multiple models can be employed [23]. Cases are identified from multiple sources, where a source is defined as any location where a case was reported. Using the various sources, cases are matched to identify duplicate ascertainment across sources. The sources are grouped into “modes” of ascertainment. The capture-recapture method is used to estimate the size of the unknown total population with a specific disease (condition or exposure) by capturing them in one mode and recapturing them in another mode(s). Based on the assumption that the probability of capturing cases in both modes would be equal to the probability of capturing cases in each mode, the number of missed cases can be estimated and the completeness assessed [24]. The percent completeness of ascertainment is defined as the number of observed cases divided by the estimated number by capture-recapture methodology.

Types of Error and Biases Associated with Registry Data

There are two types of measurement error that can affect the accuracy in estimation: random and systematic. Random error is unpredictable and is associated with precision. It often leads to inconsistency in repeated measures. This type of error is usually due to chance alone. Systematic errors are biases in a measurement that distort the measured values from the actual values. There are many sources of systematic error, such as instrument calibration, environmental changes, and procedure changes. Methods for the collection of certain data, such as anthropometric or genetic, may change over time and introduce error based on a specific time period of collection. Alternatively, if a specific genetic test with inaccuracy is used to determine a case for all subjects in a registry, there still would be error, but the error would be constant. In any registry application, it is important to identify the possible sources of error and assess the impact the error will have on interpreting the results.

Bias is "any systematic error in the design, conduct or analysis of a study that results in a mistaken estimate of an exposure's effect on the risk of disease" [25]. Selection and information bias are the two main biases that affect registries. Selection biases are distortions that result from procedures used to select subjects for the registry or from factors that influence participation or inclusion [26]. One example is *self-selection bias* (also called healthy-worker/volunteer effect), where "healthier" participants disproportionally enroll in the registry, creating a false impression that the burden of disease is less or that the survival is increased. For epidemiological purposes, it is difficult to use registries to estimate population-based rate estimates – especially for rare diseases – because most rare disease registries are based on self-selection or hospital-based data collection. In this situation, it is difficult to determine a denominator of "at-risk" subjects because only those cases seen at the hospital or through self-selection are included in the registry. This type of bias will create distorted characteristics of the case population when looking at registry data. *Information bias* results from systematic errors in the measurement of either the exposure or the disease. Sources of this bias include poor questionnaire/survey design, data collection procedures ("interviewer bias"), selective recollection of exposures ("recall bias"), and imprecise diagnostic procedures.

Although both selection and information biases impact the estimates produced from registry data, the degree of their effect depends on how the data were collected. If the degree of inaccuracy of the registry selection (i.e., inclusion) or the data collection is uniform across the sample, then it is nondifferential in that it affects the entire monitoring process rather than just a specific piece of the process. This type of bias predominately underestimates the result. However, if the inaccuracy of the data differs across the population, such that for example, those who are selected differ from those that are not included in the registry, then the bias is differential and can impact any interpretations of the registry data as a whole. These impacts are difficult to disentangle without using methods to control for confounding.

Misclassification is a type of bias generally associated with categorical or discrete variables. This type of bias is usually introduced into registries by inaccuracies or variation in methods of data acquisition and case or exposure definitions, as mentioned above. This bias can be differential or nondifferential depending on how it affects the values of other variables associated with the variable of interest. Differential misclassification is dependent upon the values of other variables (e.g., a case defined in a hospital would not be defined the same in an outpatient setting). This type of bias can skew any summary data from the registry. Nondifferential misclassification does not depend on the values of other variables, such that the misclassification of an exposure, for example, is not dependent upon the disease status.

Changes in the diagnostic criteria for a disease can affect the comparability of cases in a registry over time. *Lead-time bias*, for example, results from advances in testing (e.g., disease-specific genetic screening and testing) that lead to an earlier identification of disease. Patients can theoretically join a registry before symptoms even begin and represent "healthier" individuals than in previous years. Any examination of data characteristics (types of treatment, symptoms, survival time) could show an improvement over time that is not necessarily attributable to effective medical care but rather to the fact that the cases are being identified earlier in the disease process. Similarly, technologies that can identify diseases noninvasively or earlier in the course of disease can influence the number of cases detected and markedly inflate the number of new cases, creating the false impression that the incidence of the disease is increasing.

With increases in genetic determination of disease and improvements in testing quality and sensitivity, the comparability of registry cases over time is a persistent issue for patient registries. Collection of information specific to the method of diagnosis, including detailed testing information, can facilitate future analyses of the data. Of course, the volume of data collection comes with a cost in terms of resources and rate of participation, so the data elements that can preserve context of each registry case must be chosen carefully. Some of these elements are generalizable across diseases (e.g., date of test, nature of test (clinical or molecular), type of person making diagnosis (clinical specialty)), but most elements of value in understanding disease diagnosis over time are, predictably, disease specific. Therefore, it is difficult to ascribe a standard set of data elements. Certainly, the data collection elements for a patient registry must be chosen carefully and reviewed often.

Variability is a random bias that may attenuate true associations in epidemiologic measures but is not intrinsically fatal to certain registry objectives. Within-subject variability tends to average out for repeated measures (e.g., blood ammonia test for urea cycle disorders), whereas observer/measurement variability can vary on its overall effect on the measure of interest. This variability is usually random but can be systematic if different observers or instruments are introduced or not properly trained or calibrated. To reduce systematic bias, it is important to make sure that observers or data collection instruments observe or measure data consistently from all subgroups of the sampled population. Thoughtful design of data collection elements and protocols – driven by multidisciplinary team of disease experts, measurement experts, and informatics professionals – can at least collect enough data to identify, characterize, and control for variability in registry data over time.

Sensitivity estimates how successful a registry is at identifying all of the events, cases, or exposures in the target population. Sensitivity is the probability that a subject who is truly diseased (or exposed in case of exposure registries) will be classified as such by the method used for ascertainment. The level of sensitivity is based on the purpose of the registry. If the registry is purely to monitor trends in disease, then a low sensitivity is satisfactory. However, if the purpose is to assess the distribution or impact of a therapy, then high sensitivity is needed.

Best Practices for Patient Registries in Rare Diseases Research

The unending proliferation in registries is driving a need for registry best practices. Based upon the limitations mentioned above, we can adopt some general guidelines, mostly from the public health practice literature [1], for first determining the appropriateness of a patient registry for a given purpose and best practice for developing and maintaining various types of patient registries. Foci should be on methods that maximize and quantify the level of case ascertainment and limit (or measure) the presence of the biases discussed earlier.

Evaluate Alternatives

Before even considering a registry, the motivations and long-term commitment must be thoroughly explored. Costs for even a simple administrative registry can be expensive. Long-term, multinational registries that capture clinical data can employ dozens to hundreds of people at tremendous expense. More efficient and cheaper alternatives to registries, such as cross-sectional surveys or short-term or limited catchment studies, should always be considered before establishing a new patient registry. Particular caution should be exercised in opening new registries when the primary motivation is epidemiological. The epidemiologic usefulness of a registry increases the longer it has been in existence, often meaning that data collection, documentation, and quality control activities be conducted for many years before a register becomes fully productive for epidemiological purposes [2]. As a general rule, patient registries require continual funding and long-term commitment and should be undertaken only with strong assurance that the registry will be needed and will be funded for years or decades into the future. As summarized by Wedell in a 1973 review: “The critical question is: can this be done any other way?” If the answer is “yes,” then registry planners should consider them heavily [27].

General Methodology and Best Practices

Based upon the intended purpose, certain functionality and best practices will be required. Broadly, the functionalities relate to those presented in Table 13.1: completeness of case ascertainment, type of data collected, verification of data validity, and patient follow-up. The development of registry procedures and data specifications depend upon the goals of the registry and the stakeholders involved. General stages in the development of a registry projects are presented below.

Develop and Document Explicit Goals for the Registry

The ideal design and scope of a registry data collection system is determined by its intended purpose and funding. Once decided that the development of a new registry is warranted, the first step is to develop clarity and consensus on the goals for the registry. Any registry endeavor should start with a clear description of the purpose, which should be vetted through and consensually agreed upon by various stakeholders. Stakeholders for patient registries include patients and families, clinicians, genetic counselors, industry, patient advocacy groups (often multiple), and regulatory agencies – especially if the registry is being developed to support future drug development and approval. The US FDA has encouraged researchers and patient groups to incorporate “regulatory sufficiency” into registry design, with the assumption that the data collected in registries will support the evaluation of treatments in therapeutic trials. Of particular importance is the development of clinical end points that will be acceptable to regulatory agencies at the time of premarketing drug research. Therefore, it is beneficial to engage in dialogue with regulators regarding the appropriateness of various proposed registry data points for future phase 2 and 3 trials in a given disease area.

It is particularly important to note the differences between etiological and therapeutic research, as well as the inherent limitations of registries and observational research designs for the latter [28]. For any comparison of treatment effectiveness, the randomized clinical trial remains the ideal, and perhaps the only credible, means for conclusion – despite the logistic and ethical challenges [29, 30]. The need for randomization emerges from the likely presence of patient-or care-related characteristics that are subtle, complex, and unknown and not easily subject to quantification. These characteristics, then, act as confounders and potentially mask any attempts at comparison. In practice, whenever a rational indication for intervention exists, confounders are likely [30].

However, the use of registries for observational research is often the only practical research strategy. In these cases, it is important to consider the sampling and bias issues, carefully select appropriate comparison groups, and collect adequate data to compare relevant differences between groups. This is particularly vital in exposure registries [31]. Although registry data can be a good source for patient identification to conduct a randomized clinical trial or observation study, it is important that the registry design is clearly understood before patients are selected and associated medical data are used.

Develop Leadership Structure and Policies for Data Storage, Protection, and Access

Once the purpose and goals of the registry are clearly defined, then issues of data ownership and security need to be addressed. These issues affect the enrollment of individuals in the registry and need to be clearly disclosed to all potential registry participants, as part of the informed consent process. Before any data are collected, a data sharing and release policy needs to be developed and documented, and a governance structure for the registry will be required. It is critical to have this in place to ensure that registry data are protected but also disseminated to trusted parties for review and action. Technical solutions for registry transactions and relevant data security should be driven by the policies and requirements set forth by registry leaders.

This issue of leadership and governance is particularly important now as registries are being hosted and marketed by commercial interests outside of traditional research models (e.g., 23 and Me, Inc., PatientsLikeMe) with business models and patients-as-customers driving them [32].

Develop Adequate Infrastructure

A registry should be conceptualized as a multidisciplinary endeavor, and the skills of a multidisciplinary team are crucial. Registry efforts should include active involvement of epidemiologists and biostatisticians as well as technical and informatics specialists. The multidisciplinary team should engage in discussions on the best approach to capture the most valid data on the most cases (or the most representative cases) possible. The goals of the registry will be both the driver and the benchmark for measuring success and will drive iterative discussions on the design and operation of the registry.

Identify Data Sources

The scope, purpose, and funding commitment of the registry also influence decisions about data source (e.g., medical record abstraction, patient self-report) and the aggressiveness of follow-up. The limited resources that are true of any registry project are weighted against the strengths and weaknesses of various data sources. All possible data sources, including existing sources such as death records, related registries or epidemiologic studies, and healthcare records, should be listed and considered at this phase. Small pilot investigations or review of previous work can help determine the suitability of the data source to meet the purpose of the registry will be required. Some data sources that are suitable for applications in prevalent diseases will have particular limitations for rare diseases. For example, although mortality data are often a good data source for chronic disease epidemiology, these data are not suitable for rare diseases, many of which are undiagnosed or "lost" in the death certificate coding system that lumps various rare diseases under a more general heading "other." For epidemiological prevalence studies in general, the use of multiple data sources is preferred to fully understand the disease activity in a given region and might be required for many rare diseases. For most rare disease registry projects, there may not be any existing data collection sources that are appropriate, and new organizational mechanisms for recruitment, enrollment, data collection, and follow-up will need to be devised. If the registry data are to be used as if they were collected from a prospective longitudinal hypothesis–driven study, then the rigor, documentation, enforcement, and validation of registry data collection should be subject to the same methodological consideration as a rigorous natural history study. In this regard, registry developers should consult established clinical research methods and best practices [33]. A detailed research protocol is required for registries developed specifically for postmarket approval studies [16].

Identify Inclusion/Exclusion Criteria, Including Case Definitions

Standardization of data definitions and clinical diagnostic criteria is critical to ensure valid and reliable data for all registry purposes. More detailed examination of representative subsamples might be conducted to validate large survey results, and feedback of the results of validity tests are the primary objectives for registry developers. For exposure-based registries, the length or circumstances of the exposure and the method for determining it (e.g., patient report, public records, pharmacy data) will need to be outlined. With genetic registries, the test method needs to be specified clearly. It is particularly important to standardize and clearly document inclusion criteria and data collection. As diagnostic methods change over time, combining cohorts becomes difficult because the case populations have changed. In these cases, analysts are forced to use the "weakest" case definition that can be derived for all registry cohorts.

Sampling and Surveillance Methods

Passive and active surveillance are two alternative approaches to identifying cases. Passive surveillance is the approach where the registry does not contact possible reporters directly but rather leaves the reporting to others, such as mandated or systematic monitoring system (i.e., physicians are mandated to report cases of influenza or cancer). Active or epidemiological surveillance is an approach where the party conducting ascertainment initiates procedures to obtain data through telephone calls, mailers, or visits with physicians or hospitals. Based on the method of surveillance, bias can be introduced. Passive surveillance is most likely affected by systematic error due to its standard monitoring process, whereas active surveillance is most affected by selection bias. Internet-based registries where patients self-select to enroll are considered passive surveillance and are affected by both systematic bias and selection bias. When doing surveillance, whether active or passive, it is important that the approach used is consistent and documented in detail.

In registries designed for epidemiologic research, it is necessary to check regularly the completeness of case ascertainment – both to evaluate the effectiveness of the outreach and to understand any biases that will affect data interpretation. Eligibility and data collection from each registry case must be collected in a standard manner. Observations on the characteristics of (diseased) cases should be compared with data on the general population (from census, special population surveys, or by matched control studies) [2]. With genetic registries, the test method will need to be specified clearly, with the understanding that tests will change, metrics of the tests are questionable, and variability between labs will exist.

Design Data Collection Instruments

The most basic and important piece of all registries is the design of the data collection tool, which usually is a data collection form or patient-directed survey. The content of the form (i.e., the data collected) is, of course, driven by the goals and resources of the registry. Most registries capture disease, exposure, demographic, severity, and treatment information, as well as some identification number or means to uniquely identify patients and prevent duplicate records in the registry. Important data to include for rare diseases are genetic factors to establish genotype-phenotype correlations, family history, concomitant medications, and medical or surgical interventions. The data to be collected in a survey tool must be specific to the objectives of the study and associated analyses to be conducted. One tendency that investigators should be deterred from is trying to collect or measure too much. Data collection is a tedious and time-consuming process, so it is important to limit metrics that are of secondary importance.

As discussed in detail in Chap. 11, each variable included for measurement should have an operational definition and documented procedures for collection. This will reduce bias and increase validity and repeatability of the findings from an analytical standpoint. It is very important that data collection instruments are standardized across settings or regions (e.g., countries) and that the definitions used to identify a case represent the “standard” or conventions used in the reporting community.

Procedures for data quality and completeness should be developed before data collection begins and evaluated regularly. This might include training and testing of observers/data entry staff and the use of standard or clinical reference material which all data collection centers can calibrate to. Periodic review of the data can identify data elements or system features that need refinement to produce quality or complete data.

The costs for clinical data collection are huge, and many registries are considering patient-reported data as an alternative. Future studies will illuminate which types of data can be reliably reported by patients (e.g., quality of life, functioning, family history). Additionally, future studies might provide insight regarding methods for verifying patient-reported data, thereby increasing the validity of the data while still utilizing economically viable data sources [34].

Plan Follow-Up Data Collection Procedures

Perhaps one of the most expensive registry activities is the collection of follow-up data. The frequency and method of follow-up are influenced by both the purpose of the registry and the resources available. A statistical analysis plan should be developed at the design phase of the registry. Inconsistent follow-up procedures and success can lead to significant bias and affect the interpretation of registry data. In addition to aiming for complete patient follow-up, registry developers will need to characterize those lost to follow-up.

Continually Reevaluate Purpose and the Registry

Communication between registry stakeholders and registry leadership (both governance and implementers) is vital to a registry, and there should be continuous dialogue between all interested parties throughout the life of a registry. There is an inevitable trade-off between limited time and resources and the amount and quality of the data, and this must be recognized by registry stakeholders and leadership. The value of a registry must be reexamined periodically to ensure that the objectives are still relevant and obtainable [27]. A plan or criteria for closing the registry should be specified at the start of the project [31].

Data Standards

Standards should certainly be given priority and consideration at the design of any registry project. Because data standards are continually evolving, there are ongoing opportunities for disease investigators and activists to engage in standards development activities. There are currently no standards for developing registry programs, systems, or data collection instruments, although existing registries, particularly cancer registries, can provide valuable experience. The US Agency for Healthcare Quality and Research commissioned a comprehensive report on the role of patient registries for scientific, clinical, and policy purposes [17]. This report, recently updated as of this writing, 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. A critical and largely unaddressed problem for registries is the need for tools that allow registry data collection forms and their component questions and answers to be encoded in such a way that they can be retrieved for reuse (e.g., to support the rapid development of another related rare disease registry) or that the collected data can be interoperable with other data sources (e.g., personal health records or electronic medical records). This is the focus of several data element repository applications, including the Patient Registry Item Specification and Metadata (PRISM) project for standardizing data elements for rare disease registries and the Consensus Measures for Phenotypes and Exposures (PhenX) funded by the National Human Genome Research Institute (NHGRI) to contribute to the integration of genetics and epidemiologic research – specifically to support standards for genotype-phenotype correlation studies.

Broad areas of standardization that need to be considered when developing a registry include the choice of data content and structure. Specifically, a data model (~data fields) and associated controlled terminologies must be selected. These of course must address the objectives of the registry, but also enable any interoperability needs that might conceivably emerge in the future, and follow standard regulations where applicable. Both of these requirements are vague and dynamic, so it is impossible to prescribe a universal set of standards. The dominant discussion forums for moving toward clinical research data standards that support applied uses are the Clinical Data Standards Interchange Consortium (CDISC) and the Regulated Clinical Research (RCRIM) Technical Committee of Health Level Seven (HL7). Compelling use cases for shared clinical and research data drove the development of the BRIDG domain analysis model as a shared model to harmonize both sets of standards [35]. New and forthcoming pilot projects sponsored by HL7 and CDISC that demonstrate the use of common data elements and the BRIDG for specific therapeutic areas (e.g., cardiovascular, tuberculosis, and diabetes) should be monitored and explored as a source of standardized questions for rare disease registries [35–37]. Similarly, the most recent CDASH recommendations are promising in terms of standardizing form and section names (e.g., patient characteristics form, concomitant medication form, medical history form) [38].

Useful standardization of registry data collection forms should enable unambiguous, consistent, and reliable reuse of questions, answers, and groups of question/answer sets among different registries. Standards for the representation of common sets of questions and answers are maturing (e.g., CDISC/CDASH, caDSR/caBIG), though implementation is still not common and their encoding with standard terminologies is not done consistently [36, 37]. Semantic encoding of data elements (i.e., question + answer + definition) is very prone to intercoder variability [38, 39] and makes consistent querying based on these "standard" codes difficult and unreliable.

Previous research and the current US federal standard for standardized assessment instruments have shown that a combination of standards (specifically LOINC + SNOMED CT) is ideal to represent first the structural and generic features of questions and then the clinical content [40]. Promising feasibility studies have been conducted on small samples of questions in nursing, mental health, and public health [41–43]. Other standards, such as a recent (December 2008) standards recommendations put forth by the American Health Information Community's (AHIC) Family Health History Multi-Stakeholder Workgroup to the Office of the National Coordinator (ONC) can identify data elements for family history data collection, although controlled terminology such as SNOMED CT has not yet been incorporated into the standard [44].

One of the most important constraints for rare disease registries is coding and classification – both for finding related registries and linking them to other relevant data sources. There is no global "master index" of registries, so it is hard to know if a new registry is duplicating work or could be an extension of an existing program. Registry participation could be increased if people/physicians could be aware of all registry opportunities and not asked to submit data to separate but related registries. There is a need for standards to "organize" or inventory registries. The Orphanet project in Europe (http://www.orpha.net/consor/cgi-bin/ResearchTrials_ParticipateClinicalResearch.php?lng=EN) maintains a database of rare disease registries in European and surrounding countries, although it is unclear how perfect the inventory is considering the need for continuous data collection and the fact that the system is voluntary. Registries can be included in various trial registers (discussed in detail in Chap. 20), but the coverage is not complete.

Ethical and Policy Issues

There are several important ethical and policy considerations that need to be explored for registries that will operate in a multinational context. This is especially clear in the EU, where a mix of policies – at the regional, national, and European level – regarding consent and data sharing are difficult to navigate [3]. The variety of disparate regulations not only govern general consent, research, data collection methods, and privacy issues, but also dictate which data elements can be collected and how patients can be recruited. Because of the confusion, several groups have gotten together to assimilate these resources [3, 45–47], though we must point out this is a dynamic area in need of continuous reexamination.

Rare diseases, some with very visible phenotypes and small numbers of affected individuals, are especially vulnerable to possible identification. The increasing availability of electronic data with which to link to individuals in a registry has enabled the capture of data beyond the registry data set. This has been demonstrated in cancer by combining registry data with treatment and clinical data from health insurance records [48] and hospital data [49], and socioeconomic status from census data [50]. In *deidentified data*, all explicit identifiers, such as Social Security number, name, or address, are removed or replaced with an alternative. Deidentifying data does not guarantee that the result is anonymous, however. *Anonymous data* imply that the data cannot be manipulated or linked to identify any individual [51]. "Privacy" is emerging as a scientific discipline that includes mathematics and computer science to help address today's privacy-technology conflicts – including the prevention of reidentification from combining multiple seemingly innocent data sources [51, 52]. The creation and use of special algorithms, techniques, and qualified oversight are especially critical for rare diseases to prevent the identification of cases by association with other data sets.

An outstanding question that remains unanswered is the identification of who is best suited to host a registry and control the data. The notion of patients (via patient advocacy organizations) "owning" their data collection is gaining popularity and with some good reason. However, the resources, expertise, and governance structure of these groups vary tremendously, and all might not be ready for the demands and responsibilities of data stewardship. There has been little work to explore the nature and organizational characteristics of different patient organizations. Patient organizations are exempt from some regulations such as HIPAA, although the use of registries for research purposes does constitute research involving human subjects and is subject to those regulations. The summary of a recent multiperspective and EU-wide meeting on the topic of registries (funded by the European Commission Public Health Directorate) called for a code of conduct for patient organizations, academic researchers, policy makers, and the industry regarding the use of health information in biomedical research [3].

Future of Registries

The future of registries will continue to be shaped by the changing models of registry sponsorship, the availability of technical tools to support registry development and patient access to registry participation, and the development of policy to support global cooperation in medical research. The availability of computer technology has contributed to the proliferation of registries and influenced their evolution. Over the past decades, there has been more direct use of registries for patient care including chronic disease management, delivery of best practice guidelines to both patients and providers, and quality care on both institutional and community levels [17]. In addition, we are seeing computer technology impact the nature and scope of registry data by affecting the collection (i.e., new sources), the volume, the quality (e.g., verification by using multiple sources), the promotion (e.g., social networking), and follow-up (e.g., customized reminders for data updates or corrections).

The transparency of systems and processes enabled by information technology can enable patients to consent to their information being part of a registry and allow them to specify preferences regarding how their data are used over time. Implied in that consent, and enabled by information technology, is the monitoring and control of the data. 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 electronic health record (EHR) data direct from healthcare providers or from patient-managed personal health records (PHR). PHRs might someday 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 – arguably of central importance to rare disease research. One assumption of PHRs is that they provide data that are complete and closest to the patient. Data streams from physiologic or device measures could also be incorporated.

Social networking tools (e.g., MySpace, Facebook) are playing a growing role in the promotion and recruitment of registries. In rare diseases, coping with multiple languages will be a growing challenge. New applications are enabling patients to view aggregate data from similar patient communities, creating emergent needs for guidance on the presentation and appropriateness and utility of these ventures. Patient advocacy groups and vendors can analyze data to share with patient communities but should be cautioned as to how the data are displayed or used. As mentioned earlier, registry data are generally inappropriate for comparing treatments, and any presentation of registry data for this purpose could be misleading and perhaps dangerous.

Advances in technology, standards, global communication, and policy will be needed to support expanded use and functionality of patient registries in the future. Technology and tools are needed to enable the rapid development of registries and to maximize participation by reducing response burden and enabling high-quality data collection. Standards are required to enable sharing of content and technology across registry efforts and to enable the reuse of data from clinical settings or patient reports. In that sense, registry standards must be compatible with healthcare, though we are likely to see a certain synergy of standards as the eligibility criteria for clinical trials begin to drive the type and strategy of data collection in EHRs and healthcare settings. As noted in other research activities, multinational cooperation is needed for consistent or complementary policies for data stewardship and patient privacy and data collection so that registries can enable global research and patient safety. There is broad support for registries as a means to complement clinical research and patient care, and with rigorous design and informatics involvement, registries can gain prominence within the hierarchies of evidence that are used to support patient guidelines and policy.

References

1. Solomon DJ, et al. Evaluation and implementation of public health registries. Public Health Rep. 1991;106:142–50.
2. WHO. Epidemiological methods in the study of chronic diseases. Eleventh report of the WHO expert committee on health statistics. Geneva: World Health Organization; 1967.
3. EPPOSI. EPPOSI workshop on patients registries for rare disorders. 2009. Available from http://rbdd.org/index.php?option=com_content&view=article&id=102. Cited 22 July 2009.
4. Goldberg J, Gelfand HM, Levy PS. Registry evaluation methods: a review and case study. Epidemiol Rev. 1980;2:210–20.
5. Bellows MT. Case registers. Public Health Rep. 1949;64:1148–58.
6. Muilu J, Peltonen L, Litton JE. The federated database – a basis for biobank-based post-genome studies, integrating phenome and genome data from 600,000 twin pairs in Europe. Eur J Hum Genet. 2007;15:718–23.
7. Nakamura Y. The BioBank Japan project. Clin Adv Hematol Oncol. 2007;5:696–7.

8. Ollier W, Sprosen T, Peakman T. UK Biobank: from concept to reality. Pharmacogenomics. 2005;6:639–46.
9. Sandusky G, Dumaual 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;46:2–9.
10. Irgens LM, Bjerkedal T. Epidemiology of leprosy in Norway: the history of The National Leprosy Registry of Norway from 1856 until today. Int J Epidemiol. 1973;2:81–9.
11. Groth-Petersen E, Knudsen J, Wilbek E. Epidemiological basis of tuberculosis eradication in an advanced country. Bull World Health Organ. 1959;21:5–49.
12. Wirth HE, Locke BZ. Analyzing the tuberculosis case register. Public Health Rep. 1957; 72:151–7.
13. Drolet BC, Johnson KB. Categorizing the world of registries. J Biomed Inform. 2008;41: 1009–20.
14. Parkin DM. The evolution of the population-based cancer registry. Nat Rev Cancer. 2006;6: 603–12.
15. The Genetic Alliance. Discovering openness in health systems. In: The Genetic Alliance 2009 annual conference, Bethesda; 2009.
16. FDA. Guidance for industry and FDA staff. Procedures for handling post-approval studies imposed by PMA order. U.S. DHHS., FDA, Center for Devices and Radiological Health Rockville; 2007.
17. AHRQ. Registries for Evaluating Patient Outcomes: A User's Guide. Gliklich RE, editor. Rockville: Agency for Healthcare Research and Quality; 2007
18. Pedersen E. Some uses of the cancer registry in cancer control. Br J Prev Soc Med. 1962; 16:105–10.
19. Brooke EM. The current and future use of registers in health information systems. Geneva: World Health Organization; 1974.
20. USPHS. The automation of rheumatic fever registries; report of a seminar, May 21 and 22, 1968. Public Health Service, United States Department of Health, Education and Welfare: Washington, DC; 1969.
21. Sekar CC, Deming WE. On a method of estimating birth and death rates and extent of registration. J Am Stat Assoc. 1949;44:101–15.
22. Sekar CC, Deming WE. On a method of estimating birth and death rates and the extent of registration (excerpt). Am Stat. 2004;58:13–5.
23. Cochi SL, et al. Congenital rubella syndrome in the United States, 1970–1985. On the verge of elimination. Am J Epidemiol. 1989;129:349–61.
24. Tilling K. Capture-recapture methods – useful or misleading? Int J Epidemiol. 2001;30:12–4.
25. Schlesselman JJ. Case–control studies: design, conduct and analysis. New York: Oxford University Press; 1982.
26. Rothman K, Greenland S. Modern epidemiology. 2nd ed. Hagerstown: Lippincott Williams and Wilkins; 1998.
27. Weddell JM. Registers and registries: a review. Int J Epidemiol. 1973;2:221–8.
28. Green SB, Byar DP. Using observational data from registries to compare treatments: the fallacy of omnimetrics. Stat Med. 1984;3:361–73.
29. Rockette HE, Redmond CK, Fisher B. Impact of randomized clinical trials on therapy of primary breast cancer: the NSABP overview. Control Clin Trials. 1982;3:209–25.
30. Miettinen OS. The need for randomization in the study of intended effects. Stat Med. 1983;2: 267–71.
31. FDA. Guidance for industry. Establishing pregnancy exposure registries. U.S. Department of Health and Human Services, Food and Drug Administration, Center for Biologics Evaluation and Research (CBER); 2002.
32. Frost JH, et al. How the social web supports patient experimentation with a new therapy: the demand for patient-controlled and patient-centered informatics. AMIA Annu Symp Proc. 2008;2008:217–21.

33. European Medicines Agency (EMA). "International Conference on Harmonisation (ICH) Topic E 6 (R1). Guideline for good clinical practice; 2002." Guideline # CPMP/ICH/135/95. P. 59. Available at: http://www.emea.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500002874.pdf Accessed on 12 December 2011.
34. Murff HJ, Spigel DR, Syngal S. Does this patient have a family history of cancer? An evidence-based analysis of the accuracy of family cancer history. JAMA. 2004;292:1480–9.
35. Fridsma DB, et al. The BRIDG project: a technical report. J Am Med Inform Assoc. 2008;15:130–7.
36. Nadkarni PM, Brandt CA. The common data elements for cancer research: remarks on functions and structure. Methods Inf Med. 2006;45:594–601.
37. Richesson RL, Krischer JP. Data standards in clinical research: gaps, overlaps, challenges and future directions. J Am Med Inform Assoc. 2007;14:687–96.
38. Richesson RL, Andrew JE, Krischer JP. Use of SNOMED CT to represent clinical research data: a semantic characterization of data items on case report forms in vasculitis research. J Am Med Inform Assoc. 2006;13:536–46.
39. Andrews JE, Richesson RL, Krischer JP. Variation of SNOMED CT coding of clinical research concepts among coding experts. J Am Med Inform Assoc. 2007;14:497–506.
40. CHI. Consolidated Health Informatics. Standards adoption recommendation. Functioning and disability. U.S. DHHS, Consolidated Health Informatics; 2006.
41. White TM. Update on survey instruments and questions. In: Clinical LOINC® Meeting. Salt Lake City; 2004.
42. Carter J, et al. Making the "minimum data set" compliant with health information Technology standards. Executive Summary. U.S. Department of Health and Human Services; 2006.
43. Bakken S, et al. Evaluation of the clinical LOINC (Logical Observation Identifiers, Names, and Codes) semantic structure as a terminology model for standardized assessment measures. J Am Med Inform Assoc. 2000;7:529–38.
44. Feero WG, Bigley MB, Brinner KM. New standards and enhanced utility for family health history information in the electronic health record: an update from the American Health Information Community's Family Health History Multi-Stakeholder Workgroup. J Am Med Inform Assoc. 2008;15:723–8.
45. Godard B, et al. Data storage and DNA banking for biomedical research: informed consent, confidentiality, quality issues, ownership, return of benefits. A professional perspective. Eur J Hum Genet. 2003;11:S88–122.
46. DuchenneConnect. 2009. DuchenneConnect. Available from https://www.duchenneconnect.org. Cited 24 Aug 2009.
47. TREAT-NMD. TREAT-NMD neuromuscular network. 2009. Available from http://www.treat-nmd.eu/home.php. Cited 27 July 2009.
48. Warren JL, et al. Overview of the SEER-Medicare data: content, research applications, and generalizability to the United States elderly population. Med Care. 2002;40:IV-3–18.
49. Brooks JM, et al. Information gained from linking SEER Cancer Registry Data to state-level hospital discharge abstracts. Surveillance, epidemiology, and end results. Med Care. 2000;38:1131–40.
50. McNally RJ, et al. Geographical and ecological analyses of childhood acute leukaemias and lymphomas in north-west England. Br J Haematol. 2003;123:60–5.
51. Sweeney L. Weaving technology and policy together to maintain confidentiality. J Law Med Ethics. 1997;25:98–110.
52. Sweeney L. Privacy-preserving surveillance using databases from daily life. IEEE Intell Syst. 2005;20:83–4.

Part III
Knowledge Representation and Discovery

Chapter 14
Knowledge Representation and Ontologies

Kin Wah Fung and Olivier Bodenreider

Abstract Ontologies have become important tools in biomedicine, supporting critical aspects of both health care and biomedical research, including clinical research. Some even see ontologies as integral to science. 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, ontologies focus on what is always true of entities, i.e., definitional knowledge. 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. In this chapter, we focus on those ontologies of particular relevance to clinical research. After a brief introduction to ontology development and knowledge representation, we present the characteristics of some of these ontologies. We then show how ontologies are integrated in and made accessible through knowledge repositories and illustrate their role in clinical research.

Keywords Knowledge representation • Biomedical ontologies • Research metadata ontology • Data content ontology • Ontology-driven knowledge bases • Data integration • Computer reasoning

Ontologies have become important tools in biomedicine, supporting critical aspects of both health care 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, ontologies focus on what is always true of entities, i.e., definitional knowledge [3].

K.W. Fung, M.D., M.S., M.A. (✉) • O. Bodenreider, M.D., Ph.D.
Lister Hill National Center for Biomedical Communications,
US National Library of Medicine, National Institutes of Health,
8600 Rockville Pike, Bethesda, MD 20894, USA
e-mail: kfung@mail.nih.gov; obodenreider@mail.nih.gov

R.L. Richesson, J.E. Andrews (eds.), *Clinical Research Informatics*, Health Informatics, DOI 10.1007/978-1-84882-448-5_14, © Springer-Verlag London Limited 2012

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]. In this chapter, we focus on those ontologies of particular relevance to clinical research. After a brief introduction to ontology development and knowledge representation, we present the characteristics of some of these ontologies. We then show how ontologies are integrated in and made accessible through knowledge repositories and illustrate their role in clinical research.

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 1,234), 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 instances recorded in patient records and laboratory notebooks can be annotated.) 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 Basic Formal Ontology [7] (BFO) and Descriptive Ontology for Linguistic and Cognitive Engineering [8] (DOLCE). 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 Unified Medical Language System (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 [10] – 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 which form 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 [11]. 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 Epoch Clinical Trial Ontologies (CTO), Ontology of Clinical Research (OCRe), Ontology for Biomedical Investigations (OBI), Systematized Nomenclature of Medicine –

Clinical Terms (SNOMED – CT), National Drug File – Reference Terminology (NDF – RT), and the NCI (National Cancer Institute) Thesaurus, discussed later in this chapter, all rely on some sort of description logic (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 [12]. It is therefore not surprising that the main language for representing ontologies, the Web Ontology Language (OWL), 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 [13]. Other Semantic Web technologies, such as RDF/S (Resource Description Framework Schema) [14] and Simple Knowledge Organization System (SKOS) [15], 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 OBO family. The so-called OBO syntax [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 plug-ins extend the standalone 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. Other ontology editors related to some of the ontologies presented in this chapter include Apelon's proprietary Terminology Development Environment (TDE), based on the description logics KRSS and used for the development of NDF-RT, and the International Health Terminology Standards Development Organisation (IHTSDO) Workbench, an open-source, freely available editing environment created for the collaborative development of SNOMED CT.

Open Biomedical Ontologies Foundry and Other Harmonization Efforts

Two major issues with biomedical ontologies are their proliferation and their 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 has 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. The International Health Terminology Standard Development Organization (IHTSDO), 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), 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. Some examples of the two types of ontology will be described in more detail. Finally, examples of ontology-driven knowledge bases for translational research will be presented briefly.

Research Metadata Ontology

A survey of the public repository of ontologies in the Open Biomedical Ontologies (OBO) library hosted by the National Center of Biomedical Ontology (see below) yielded three ontologies that fit the description of research metadata ontology. These are the Epoch Clinical Trial Ontologies (CTO), Ontology of Clinical Research (OCRe), and Ontology for Biomedical Investigations (OBI).

Epoch Clinical Trial Ontologies

CTO is a suite of ontologies that encodes knowledge about clinical trials. The use of this ontology is demonstrated in the integration of software applications for the management of clinical trials under the Immune Tolerance Network [25]. By building an ontology-based architecture, the disparate clinical trial software applications can share essential information to achieve interoperability for efficient management of the trials and analysis of trial data. CTO is made up of the following component ontologies:

1. Clinical trial ontology – the overarching ontology that covers protocol specification and operational plan
2. Protocol ontology – the knowledge model of the clinical trial protocol
3. Organization ontology – supports the specification of study sites, laboratories, and repositories
4. Assay ontology – models characteristics of tests (e.g., specimen type, workflow of specimen processing)
5. Labware ontology – models the laboratory entities (e.g., specimen containers)
6. Virtual trial data ontology – models the study data being collected (e.g., participant clinical record, specimen workflow log)
7. Constraint expression ontology – models logical and temporal constraints
8. Measurement ontology – models physical measurements and units of measurement

There are three stated goals of CTO: to support tools which help acquire and maintain knowledge about protocol and assay designs, to drive data collection during a trial, and to facilitate implementation of querying methods to support trial management and *ad hoc* data analysis. A clinical trial protocol authoring tool has been developed based on CTO [26]. The ability to map from CTO to the Biomedical Research Integrated Domain Group (BRIDG) information model has been demonstrated [27].

Ontology of Clinical Research

While the main use case of CTO is in the automation of design and workflow management 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 [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 CTO and OCRe whose creations are rooted in clinical research, the origin of OBI is in the molecular biology research domain [29]. 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) [30]. 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 [31].

Another difference between OBI and the other two ontologies is the collaborative approach to its development. 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, centrifuge
3. Information content – e.g., electronic medical record, biomedical image
4. Design and execution of an investigation – e.g., study design, electrophoresis
5. Data transformation – e.g., principal components analysis, mean calculation

For domain-specific entities, OBI makes reference to other ontologies such as 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.

Data Content Ontology

While there are relatively few metadata ontologies, there is a myriad of ontologies that cover research data contents. Unlike metadata ontologies, in this group, the distinction between ontologies, terminologies, classifications, and code sets often gets blurred. Three ontologies are chosen for more detailed discussion here: the National Cancer Institute Thesaurus (NCIT), Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT), and National Drug File Reference Terminology (NDF-RT). These are chosen because they are arguably closer to the ontology end of the ontology-vocabulary continuum than most other artifacts in this category, and their content areas are most relevant to clinical research. All of them have concept-based organization with a rich network of interconcept relationships and use description logic formalism in content creation and maintenance. All three ontologies are available through the Unified Medical Language System (UMLS) and the BioPortal ontology repositories (see below).

National Cancer Institute Thesaurus

The National Cancer Institute Thesaurus (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 [32–34]. 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 is the reference terminology for the NCI's Cancer Biomedical Informatics Grid (caBIG) and other related projects. It is also one of the US federal standard terminologies designated by the Consolidated Health Informatics (CHI) initiative.

NCIT contains about 80,000 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.

Systematized Nomenclature of Medicine – Clinical Terms

SNOMED CT was originally developed by the College of American Pathologists. Its ownership was transferred to the International Health Terminology Standards Development Organisation (IHTSDO) in 2007 to enhance international governance and adoption [35]. There are currently 17 member countries including USA, United Kingdom, Canada, Australia, Netherlands, Sweden, and Spain. SNOMED CT is the most comprehensive clinical terminology available today, with almost 300,000 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 CHI-designated US Federal terminology standard. It is also one of the named terminology standards for the problem list in the “meaningful use” criteria for the Electronic Health Record published by the US Department of Health and Human Services [36, 37]. SNOMED CT is updated twice yearly. The use of SNOMED CT is free in all IHTSDO member countries, in low-income countries as defined by the World Bank, and for qualified research projects in any country. SNOMED CT is available in proprietary release format from the National Release Centers of the IHTSDO member countries.

National Drug File Reference Terminology

NDF-RT is developed by the US Veteran Health Administration (VA) as an extension to their National Drug File, which is the master list of drugs prescribed to VA patients. In addition to drug names, ingredients, dose forms, and strengths, NDF-RT contains hierarchies for the chemical structure, mechanism of action, physiologic effect, and therapeutic intent of drugs. There is also a disease hierarchy to which drugs may be linked through roles such as *may_treat*, *may_prevent*, and *may_diagnose*. NDF-RT contains about 4,000 drugs at the ingredient level. The coverage of NDF-RT has been evaluated using data outside of the VA system and found to be adequate [38, 39]. NDF-RT is in the public domain and is updated monthly [40]. It is available in XML and OWL formats. NDF-RT has recently been integrated with RxNorm and is now available through RxNav and its application programming interfaces (APIs) [41].

Ontology-Driven Knowledge Bases for Translational Research

Several ontology-driven knowledge bases have been developed in the past few years 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 [42], based on its own ontology, the Research Entity Dictionary, and STRIDE [43], 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 [44], other researchers have developed knowledge bases for cancer data (leveraging the NCI Thesaurus) [45] and in the domain of nicotine dependence (using an ontology developed specifically for the purpose of integrating publicly available datasets) [46]. 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 [47].

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

The US National Library of Medicine (NLM) started the Unified Medical Language System (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 [48–51]. 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 [52]. Currently, there are over 3,000 UMLS licensees in more than 50 countries. The UMLS is released twice a year.

Unified Medical Language System 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 2010AB release, the UMLS Metathesaurus incorporates 153 source vocabularies and includes terms in 20 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 133 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 450,000 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 helps users to abstract away from variations involving word inflection, case, and word order [53].

Unified Medical Language System 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, browsing the UMLS Metathesaurus, Semantic Network, and SPECIALIST Lexicon, and 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 56 ontologies (version 2010AB, as of April 2011), including SNOMED CT, the NDF-RT, and several translations of 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 SOAP (Simple Object Access Protocol) 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 credentials to be checked in order to gain access to UMLS data. Support for user authentication is provided through the UTS API itself.

Unified Medical Language System 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 interterminology 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 interterminology mapping by utilizing relationships and lexical resources available in the UMLS [54]. 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 [55, 56]. 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 [57].

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 [58].

BioPortal Ontologies

The current version of BioPortal integrates over 250 ontologies for biomedicine, biology, and life sciences and includes roughly five million terms. A number of ontologies integrated in the UMLS are also present in BioPortal (e.g., Gene Ontology, LOINC). 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 (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 and ChEBI). 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 African Traditional Medicine Ontology and the Electrocardiography 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 19 ontologies (as of April 2011, restricted to exact matches, including synonyms), 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

As the UMLS, BioPortal identifies equivalent concepts across ontologies in its repositories (e.g., between the term *listeriosis* in DermLex and in MedlinePlus 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 a list of textual resources, which have been preprocessed with the Annotator, including several gene expression data repositories, ClinicatTrials.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.

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 [59]. Users can also contribute equivalence maps between ontologies.

Ontology in Action: Uses of Ontologies in Clinical Research

To facilitate discussion, the use of ontologies and ontology-based technology in clinical research is classified into three major areas: workflow management, data integration, and computer reasoning [1]. 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 still stored and shared in textual documents and spreadsheets. The descriptors used are neither encoded nor standardized. Standalone 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 trials management applications is an attractive approach. One such effort of integration resulted in the creation of CTO (described above) which has been applied successfully in the Immune Tolerance Network, a large distributed research consortium engaged in the discovery of new therapy for immune-related disorders.

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 [60]. ACGT is a European Union cofunded project that aims at developing open-source, semantic, and grid-based technologies in support of postgenomic clinical trials in cancer research. One component of this project is the development of a tool called Trial Builder to create ontology-based case report forms (CRF). The Trial Builder allows the researcher to build CRFs based on a master ontology called ACGT Master Ontology (ACGT-MO) [61]. During this process, the metadata of the research is also captured which can be used in the automatic creation of the ontology-based data management system. The advantage of this approach is that the alignment of research semantics and data definition is achieved early in the research process, which guarantees easy downstream integration of data collected from disparate 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.

Data Integration

In the postgenomic 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 [62]. Ontologies can play important roles at different strategic steps of data integration [63].

For most existing data sources, data sharing and integration only occurs 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.

OCRe (described above) is specifically created to annotate and align clinical trials according to their design and data analysis methodology. Another effort is BIRNLex which is created to annotate the Biomedical Informatics Research Network (BIRN) data sources [64]. The BIRN sources currently 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, it also covers concepts such as experimental design, data types, and data provenance. BIRN employs a mediated 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 [65].

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 [66]. Another described a prototype system for ontology-driven indexing of public datasets for translational research [67].

One particular form of data integration supported by ontologies is represented by what has become known as *Linked Data* in the Semantic Web community [68]. 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 [44–47]. 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 [69], with particular interest in the domain of drug discovery through the Linking Open Drug Data initiative [70].

Computer Reasoning

To harness the reasoning power of computers is another important reason to use ontologies in clinical research. The use of ontologies to support reasoning is not new. The Foundational Model of Anatomy (FMA) has been used to predict the anatomic consequences of penetrating injuries and the physiological consequences of injury to the arteries supplying the heart [71–73].

The ready availability of enabling tools and utilities like Protégé, Web Ontology Language (OWL), and Semantic Web Rule Language (SWRL) makes it easier to implement computer reasoning through the use of ontologies. One example is the use of Protégé and the accompanying SWRL Temporal Built-In Library in a study of quality standards in the management of hypertension by family practitioners [74]. Clinical research often involves chronic patients with multiple comorbidities. Hierarchical and temporal types of queries are often necessary. Traditional data stored in relational databases cannot easily support queries involving hierarchical entities (e.g., all patients with codes related to hypertension) or temporal concepts (e.g., all patients with a lapse in antihypertension therapy during a certain period). These kinds of queries are often necessary in clinical trials (e.g., identifying subjects that are eligible for a particular study). As illustrated in this study, an ontology-based approach using readily available tools turned out to be a better solution.

Another area of the use of computer reasoning in clinical medicine is clinical decision support systems (CDSS). As CDSS become more widely used, it is not uncommon to find CDSS to be an important component in clinical research. CDSS often rely on ontologies to enable them to do logical reasoning. One example is ATHENA, which is an ontology-based inferencing system that encourages blood pressure control and recommends guideline-concordant choice of drug therapy in relation to comorbid diseases [75]. The ATHENA ontology specifies eligibility criteria, risk stratification, blood pressure targets, relevant comorbidities, and preferred drugs within each drug class. One special feature of ATHENA is that clinical experts themselves can customize the knowledge base to incorporate new evidence or to reflect local interpretation of guideline ambiguities.

The Way Ahead

Looking forward, it is encouraging that the value of ontologies in clinical research is becoming more recognized. This is evidenced by the increase in the number of research teams 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 is 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;47:67–79.
2. Smith B. Ontology (science). 2008. Nature precedings Available from Nature Precedings (http://hdl.handle.net/10101/npre.2008.2027.2). Accessed Aug 2011.
3. Bodenreider O, Stevens R. Bio-ontologies: current trends and future directions. Brief Bioinform. 2006;7:256–74.
4. Cimino JJ. Zhu X. The practical impact of ontologies on biomedical informatics. Yearb Med Inform. 2006;45:124–35.
5. Smith B, Ceusters W, Klagges B, Kohler J, Kumar A, Lomax J, Mungall C, Neuhaus F, Rector AL, Rosse C. Relations in biomedical ontologies. Genome Biol. 2005;6:R46.
6. Simmons P, Melia J. Continuants and occurrents. Proc Aristotelian Soc. 2000;74:59–75, 77–92.
7. BFO. http://www.ifomis.org/bfo/. Accessed Aug 2011.
8. DOLCE. http://www.loa-cnr.it/DOLCE.html. Accessed Aug 2011.
9. McCray AT. An upper-level ontology for the biomedical domain. Comp Funct Genomics. 2003;4:80–4.
10. Beißwanger E, Schulz S, Stenzhorn H, Hahn U. BioTop: an upper domain ontology for the life sciences - a description of its current structure, contents, and interfaces to OBO ontologies. Appl Ontol. 2008;3:205–12.

11. Baader F, Calvanese D, McGuinness D, Nardi D, Patel-Schneider P, editors. The description logic handbook: theory, implementation, and applications. Cambridge/New York: Cambridge University Press; 2007.
12. 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;284:34–43.
13. OWL 2 web ontology language document overview. http://www.w3.org/TR/owl2-overview/. Accessed Aug 2011.
14. RDF vocabulary description language 1.0: RDF schema. http://www.w3.org/TR/rdf-schema/. Accessed Aug 2011.
15. SKOS simple knowledge organization system reference. http://www.w3.org/TR/2009/REC-skos-reference-20090818/. Accessed Aug 2011.
16. The OBO flat file format specification. http://www.geneontology.org/GO.format.obo-1_2.shtml. Accessed Aug 2011.
17. Golbreich C, Horridge M, Horrocks I, Motik B, Shearer R. 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; 2007. p. 169–82.
18. Noy N, Tudorache T, Nyulas C, Musen M. The ontology life cycle: integrated tools for editing, publishing, peer review, and evolution of ontologies. AMIA Annu Symp Proc. 2010;2010: 552–6.
19. Protégé. http://protege.stanford.edu/. Accessed Aug 2011.
20. Day-Richter J, Harris MA, Haendel M, Lewis S. OBO-edit – an ontology editor for biologists. Bioinformatics. 2007;23:2198–200.
21. OBO-edit. http://oboedit.org/. Accessed Aug 2011.
22. Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, Goldberg LJ, Eilbeck K, Ireland A, Mungall CJ, Leontis N, Rocca-Serra P, Ruttenberg A, Sansone SA, Scheuermann RH, Shah N, Whetzel PL, Lewis S. The OBO foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol. 2007;25:1251–5.
23. Harmonization. http://www.ihtsdo.org/about-ihtsdo/harmonization/. Accessed Aug 2011.
24. Richesson RL, Krischer J. Data standards in clinical research: gaps, overlaps, challenges and future directions. J Am Med Inform Assoc. 2007;14:687–96.
25. Shankar RD, Martins SB, O'Connor M, Parrish DB, Das AK. An ontology-based architecture for integration of clinical trials management applications. AMIA Annu Symp Proc. 2007;2007:661–5.
26. Shankar R, Arkalgud S. Connor M, Boyce K, Parrish D, Das A. TrialWiz: an ontology-driven tool for authoring clinical trial protocols. AMIA Annu Symp Proc. 2008:1226.
27. Tu SW, Fridsma DB, Shankar R, Connor M, Das A, Parrish D. Bridging epoch: mapping two clinical trial ontologies. In: 10th international protege conference. Budapest; 2007.
28. Tu SW, Carini S, Rector A, Maccalum P, Toujilov I, Harris S, Sim I. OCRe: ontology of clinical research. In: 11th international protege conference. Amsterdam; 2009.
29. The OBI Consortium. http://obi-ontology.org/page/Consortium. Accessed Aug 2011.
30. Whetzel PL, Brinkman RR, Causton HC, Fan L, Field D, Fostel J, Fragoso G, Gray T, Heiskanen M, Hernandez-Boussard T, Morrison N, Parkinson H, Rocca-Serra P, Sansone SA, Schober D, Smith B, Stevens R, Stoeckert Jr CJ, Taylor C, White J, Wood A. Development of FuGO: an ontology for functional genomics investigations. OMICS. 2006;10:199–204.
31. Brinkman RR, Courtot M, Derom D, Fostel JM, He Y, Lord P, Malone J, Parkinson H, Peters B, Rocca-Serra P, Ruttenberg A, Sansone SA, Soldatova LN, Stoeckert Jr CJ, Turner JA, Zheng J. Modeling biomedical experimental processes with OBI. J Biomed Semantics. 2010;1 Suppl 1:S7.
32. de Coronado S, Haber MW, Sioutos N, Tuttle MS, Wright LW. NCI Thesaurus: using science-based terminology to integrate cancer research results. Medinfo. 2004;11:33–7.
33. Sioutos N, de Coronado S, Haber MW, Hartel FW, Shaiu WL, Wright LW. NCI Thesaurus: a semantic model integrating cancer-related clinical and molecular information. J Biomed Inform. 2007;40:30–43.

34. Fragoso G, de Coronado S, Haber M, Hartel F, Wright L. Overview and utilization of the NCI Thesaurus. Comp Funct Genomics. 2004;5:648–54.
35. SNOMED CT (systematized nomenclature of medicine-clinical terms). http://www.ihtsdo.org/our-standards/. Accessed Aug 2011.
36. Blumenthal D, Tavenner M. The "meaningful use" regulation for electronic health records. N Engl J Med. 2010;363:501–4.
37. 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;75:62686–90.
38. Brown SH, Elkin PL, Rosenbloom ST, Husser C, Bauer BA, Lincoln MJ, Carter J, Erlbaum M, Tuttle MS. VA National Drug File Reference Terminology: a cross-institutional content coverage study. Stud Health Technol Inform. 2004;107:477–81.
39. Rosenbloom ST, Awad J, Speroff T, Elkin PL, Rothman R, Spickard 3rd A, Peterson J, Bauer BA, Wahner-Roedler DL, Lee M, Gregg WM, Johnson KB, Jirjis J, Erlbaum MS, Carter JS, Lincoln MJ, Brown SH. Adequacy of representation of the National Drug File Reference Terminology physiologic effects reference hierarchy for commonly prescribed medications. AMIA Annu Symp Proc. 2003;2003:569–78.
40. National Drug File Reference Terminology. ftp://ftp1.nci.nih.gov/pub/cacore/EVS/NDF-RT/. Accessed on 13 Dec 2011
41. National Library of Medicine: (RxNav), http://rxnav.nlm.nih.gov/. Accessed on 13 Dec 2011
42. Cimino JJ, Ayres EJ. The clinical research data repository of the US National Institutes of Health. Stud Health Technol Inform. 2010;160:1299–303.
43. Lowe HJ, Ferris TA, Hernandez PM, Weber SC. STRIDE – an integrated standards-based translational research informatics platform. AMIA Annu Symp Proc. 2009;2009:391–5.
44. 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 KH. Methodology – advancing translational research with the Semantic Web. BMC Bioinformatics. 2007;8:S2.
45. McCusker JP, Phillips JA, Gonzalez Beltran A, Finkelstein A, Krauthammer M. Semantic web data warehousing for caGrid. BMC Bioinformatics. 2009;10 Suppl 10:S2.
46. Sahoo SS, Bodenreider O, Rutter JL, Skinner KJ, Sheth AP. An ontology-driven semantic mashup of gene and biological pathway information: application to the domain of nicotine dependence. J Biomed Inform. 2008;41:752–65.
47. Translational medicine ontology and knowledge base. www.w3.org/wiki/HCLSIG/PharmaOntology. Accessed Aug 2011.
48. 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.
49. Humphreys BL, Lindberg DA, Schoolman HM, Barnett GO. The unified medical language system: an informatics research collaboration. J Am Med Inform Assoc. 1998;5:1–11.
50. Lindberg DA, Humphreys BL, McCray AT. The unified medical language system. Methods Inf Med. 1993;32:281–91.
51. Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. Nucleic Acids Res. 2004;32(Database issue):D267–70.
52. Unified medical language system (UMLS), http://www.nlm.nih.gov/research/umls/. Accessed Aug 2011.
53. McCray AT, Srinivasan S, Browne AC. Lexical methods for managing variation in biomedical terminologies. Proc Annu Symp Comput Appl Med Care. 1994:235–9.
54. Fung KW, Bodenreider O. Utilizing the UMLS for semantic mapping between terminologies. AMIA Annu Symp Proc. 2005;2005:266–70.
55. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp. 2001:17–21.
56. Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. J Am Med Inform Assoc. 2010;17:229–36.

57. 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.
58. Noy NF, Shah NH, Whetzel PL, Dai B, Dorf M, Griffith N, Jonquet C, Rubin DL, Storey MA, Chute CG, Musen MA. BioPortal: ontologies and integrated data resources at the click of a mouse. Nucleic Acids Res. 2009;37:W170–3.
59. Ghazvinian A, Noy NF, Musen MA. Creating mappings for ontologies in biomedicine: simple methods work. AMIA Annu Symp Proc. 2009;2009:198–202.
60. Weiler G, Brochhausen M, Graf N, Schera F, Hoppe A, Kiefer S. 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.
61. ACGT master ontology. http://www.ifomis.org/wiki/ACGT_Master_Ontology_%28MO%29. Accessed Aug 2011.
62. Genome-wide association studies. http://grants.nih.gov/grants/gwas/. Accessed on 13 Dec 2011
63. 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), vol. LNBI 5109. Berlin/Heidelberg/New York: Springer; 2008. p. 1–4.
64. Biomedical informatics research network. http://www.birncommunity.org/tools-catalog/neurolex-formerly-birnlex/. Accessed on 13 Dec 2011
65. Rubin DL, Shah NH, Noy NF. Biomedical ontologies: a functional perspective. Brief Bioinform. 2008;9:75–90.
66. Cook C, Hannley M, Richardson JK, Michon J, Harker M, Pietrobon R. Real-time updates of meta-analyses of HIV treatments supported by a biomedical ontology. Account Res. 2007;14:1–18.
67. Shah NH, Jonquet C, Chiang AP, Butte AJ, Chen R, Musen MA. Ontology-driven indexing of public datasets for translational bioinformatics. BMC Bioinformatics. 2009;10 Suppl 2:S1.
68. Bizer C, Heath T, Berners-Lee T. Linked data – the story so far. Int J Semant Web Inf. 2009;5:1–22.
69. HCLS: Semantic Web Health Care and Life Sciences (HCLS) Interest Group. (HCLS), http://www.w3.org/wiki/HCLSIG/. Accessed on 13 Dec 2011
70. Linking open drug data. http://www.w3.org/wiki/HCLSIG/LODD. Accessed Aug 2011.
71. Rosse C, Shapiro LG, Brinkley JF. The digital anatomist foundational model: principles for defining and structuring its concept domain. Proc AMIA Symp. 1998:820–4.
72. Rubin DL, Dameron O, Bashir Y, Grossman D, Dev P, Musen MA. Using ontologies linked with geometric models to reason about penetrating injuries. Artif Intell Med. 2006;37:167–76.
73. Rubin DL, Dameron O, Musen MA. Use of description logic classification to reason about consequences of penetrating injuries. AMIA Annu Symp Proc.2005:649–53.
74. Mabotuwana T, Warren J. An ontology-based approach to enhance querying capabilities of general practice medicine for better management of hypertension. Artif Intell Med. 2009;47: 87–103.
75. Goldstein MK, Hoffman BB, Coleman RW, Tu SW, Shankar RD, O'Connor M, Martins S, Advani A, Musen MA. Patient safety in guideline-based decision support for hypertension management: ATHENA DSS. Proc AMIA Symp. 2001:214–8.

Chapter 15
Nonhypothesis-Driven Research: Data Mining and Knowledge Discovery

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

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

M.R. Cummins, Ph.D., APRN
College of Nursing, University of Utah,
10 S 2000 East, Salt Lake City, UT 84112-5880, USA
e-mail: mollie.cummins@utah.edu

R.L. Richesson, J.E. Andrews (eds.), *Clinical Research Informatics*, Health Informatics,
DOI 10.1007/978-1-84882-448-5_15, © Springer-Verlag London Limited 2012

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 [4]. 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

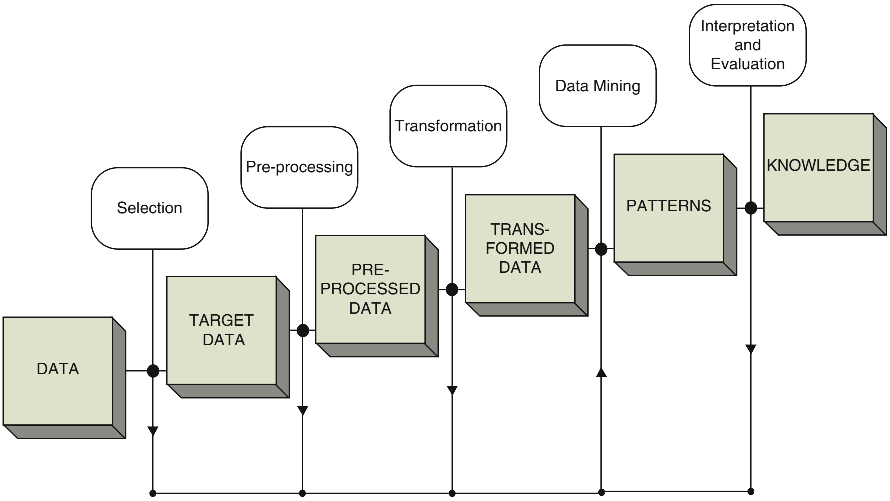


Fig. 15.1 Fayyad's knowledge discovery in databases process

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. 15.1. 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. 15.1 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 hypothetico-deductive scientific approach.

Data Selection

KDD projects are typically incepted 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 $= 2^n$. For a mere 16 inputs ($n = 16$), the number of possible combinations $= 65{,}536$. 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 [21] 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 incepted by McCulloch and Pitts at the University of Chicago during the 1930s and 1940s as a mathematical representation of frog neuron [7]. Contemporary artificial neural networks are multilayer networks composed of processing elements, variations of McCulloch and Pitt's original TLUs (Fig. 15.2). 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 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

Fig. 15.2 Multilayer artificial neural network

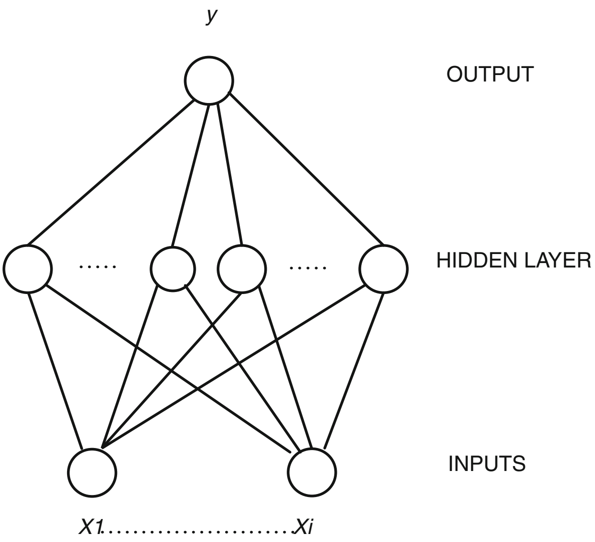
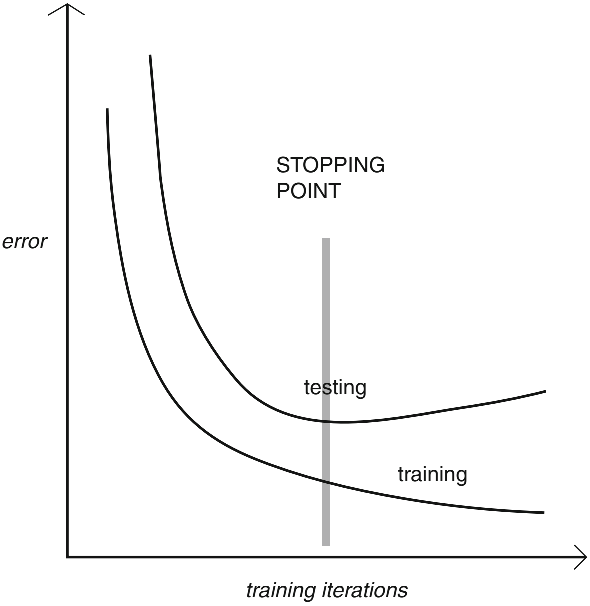


Fig. 15.3 Training/testing curves



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. 15.3).

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 [8]. 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.

Support Vector Machines

Support vector machine methods were developed by Vapnik and others in the 1970s through the 1990s [9–11]. 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) 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 [12]. 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 join 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 [13] 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 [13]. 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. [14].

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 [15].

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 noninstances 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 [16].

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 important infrastructure project in the United States is caBIG®, the cancer biomedical informatics grid. This project is addressing the barriers posed by lack of interoperability and siloed data by promoting fundamental change in the way clinical research is conducted. caBIG® collaborators are developing open-source tools and architecture that enable federated sharing of interoperable data, using an object-oriented data model and standard data definitions. These tools will facilitate data interoperability while allowing participants to retain control over the use of their own data. The project's use of object-oriented data definitions greatly facilitates the development of applications to support collection of caBIG® compatible data in clinical studies. In early 2009, the University of Edinburgh became the first European university to deploy a caBIG application, caTISSUE repository [17]. This project demonstrated caBIG's adaptability to European data.

Another major approach to facilitating biomedical knowledge discovery is that of the semantic web [18]. 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 [18].

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 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.

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 [19].

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 [20].

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 like caBIG.

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, et al. Combining decision support methodologies to diagnose pneumonia. Proc AMIA Symp 2001;12–6.
3. Lagor C, Aronsky D, et al. 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, PiatetskyShapiro G, et al. From data mining to knowledge discovery: an overview. In: Fayyad UM, Piatetsky-Shapiro G, Smyth P, Uthurusamy R, editors. Advances in knowledge discovery and data mining. Menlo Park: AAAI Press/MIT Press; 1996. p. 1–34.
5. Aronsky D, Haug PJ, et al. Accuracy of administrative data for identifying patients with pneumonia. Am J Med Qual. 2005;20(6):319–28.
6. Poynton MR, Frey L, et al. Representation of smoking-related concepts in an electronic health record. MEDINFO 2007: 12th world congress on health (Medical) informatics, International Medical Informatics Association, Brisbane; 2007.
7. McCulloch WS, Pitts WH. A logical calculus of the ideas immanent in nervous activity. Bull Math Biophys. 1943;5:115–33.
8. Quinlan J. C4.5: programs for machine learning. San Mateo, CA: Morgan Kaufmann; 1993.
9. Vapnik VN. The nature of statistical learning theory. New York: Springer; 1995.
10. Vapnik VN. Statistical learning theory. New York: Wiley; 1998.
11. Cristianini N, Shawe-Taylor J. An introduction to support vector machines: and other kernel-based learning methods. Cambridge/New York: Cambridge University Press; 2000.
12. Jonsson P, Wohlin C. Benchmarking k-nearest neighbour imputation with homogeneous Likert data. Empirical Softw Eng. 2006;11(3):1382–3256.
13. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982;143(1):29–36.
14. Lasko TA, Bhagwat JG, et al. The use of receiver operating characteristic curves in biomedical informatics. J Biomed Inform. 2005;38(5):404–15.
15. Cordero F, Botta M, et al. Microarray data analysis and mining approaches. Brief Funct Genomic Proteomic. 2007;6(4):265–81.
16. Pepe MS. The statistical evaluation of medical tests for classification and prediction. Oxford/New York: Oxford University Press; 2003.
17. Genomeweb. Persistent systems helps first european university deploy caBIG's caTissue repository. BioInform [serial on the Internet]. 2009; (May 22, 2009): Available from: http://www.genomeweb.com/persistent-systems-helps-first-european-university-deploy-cabigs-catissue-reposi
18. Ruttenberg A, Clark T, et al. Advancing translational research with the Semantic Web. BMC Bioinformatics. 2007;8 Suppl 3:S2.
19. Breiman L. Statistical modeling: the two cultures. Stat Sci. 2001;16(3):199–231.
20. Matheny ME, Ohno-Machado L, et al. Discrimination and calibration of mortality risk prediction models in interventional cardiology. J Biomed Inform. 2005;38(5):367–75.
21. Minsky ML. The society of mind. New York: Simon and Schuster; 1986.

Chapter 16
Natural Language Processing, Electronic Health Records, and Clinical Research

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 natural language processing and its common uses for extracting and transforming narrative information in EHR to support clinical research.

Keywords Electronic health records • Biomedical natural language processing • Sublanguage approach • Machine-learning approach • Decision tree • Rule-based approach

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 its common uses for extracting and transforming narrative information in EHR to support clinical research.

F. Liu, Ph.D. (✉) • H.Yu, Ph.D.
Department of Electric Engineering and Computer Science,
University of Wisconsin Milwaukee,
3200 N Cramer Street, Milwaukee, WI 53211, USA
e-mail: liuf@uwm.edu

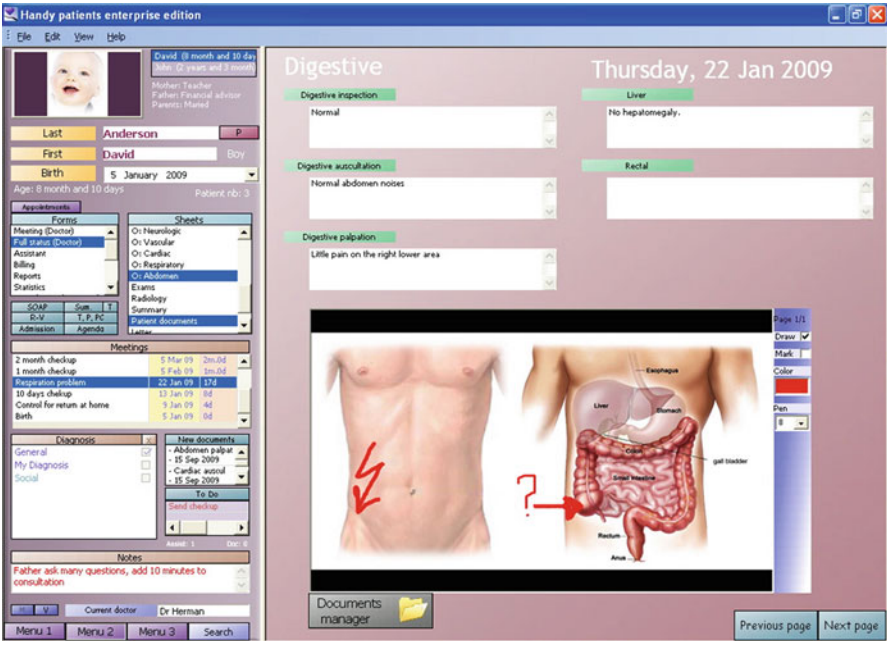
C. Weng, Ph.D.
Department of Biomedical Informatics, Columbia University,
622 W 168th Street, VC-5, New York, NY 10332, USA

R.L. Richesson, J.E. Andrews (eds.), *Clinical Research Informatics*, Health Informatics,
DOI 10.1007/978-1-84882-448-5_16, © Springer-Verlag London Limited 2012

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 National Institute of Health (NIH) has created the Clinical and Translational Science Awards (CTSA) program to reengineer the clinical research enterprise [6]. A potential powerful accelerator to clinical research is electronic health records.

An EHR is a legal computerized medical record for documenting patient information captured at every patient encounter [7, 8]. Figure 16.1 shows a sample EHR [9]. As of 2008, more than 40% of physicians in the USA were using EHRs, more than double the percentage at the start of the decade [10]. The resident population of the USA as of 2009 was 307 million [11]. During that same year, it was reported that 83% adults and 90% children had contact with a health-care professional, there were



David (8 month and 10 day
John (2 years and 3 month)

Mother: Teacher
Father: Financial advisor
Parents: Maried

Last: Anderson
First: David
Boy
Birth: 5 January 2009
Age: 8 month and 10 days
Patient nb: 3

Meetings

2 month checkup	5 Mar 09	2m.0d
1 month checkup	5 Feb 09	1m.0d
Respiration problem	22 Jan 09	17d
10 days chekup	13 Jan 09	8d
Control for return at home	9 Jan 09	4d
Birth	5 Jan 09	0d

Diagnosis
General ☑
My Diagnosis □
Social □

New documents
- Abdomen palpat
- 15 Sep 2009
- Cardiac auscul
- 15 Sep 2009

To Do: Send checkup

Assist: 1 Doc: 0

Notes: Father ask many questions, add 10 minutes to consultation

Current doctor: Dr Herman

Digestive

Thursday, 22 Jan 2009

Digestive inspection: Normal

Digestive auscultation: Normal abdomen noises

Digestive palpation: Little pain on the right lower area

Liver: No hepatomegaly.

Rectal:

Fig. 16.1 Illustration of a sample electronic health record (EHR)

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 EHRs 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 notes, discharge summaries, 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. 19). 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's 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 "prescreening" 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 preexisting 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 are 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 foundations of clinical NLP research in terms of sublanguage analysis and machine-learning models, including cutting-edge information extraction and retrieval techniques that can be applied to EHRs-based clinical research. Then, a few existing clinical NLP systems will be reviewed, 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 records (EHRs) 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 clinical narratives. Many NLP applications are developed by exploiting the sublanguage characteristics, that is, 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 nonbiological 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 characteristics 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 electronic 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 abovementioned tasks, such as entity (medications, diseases, doses). Extraction from EHRs can be realized using structured prediction models; 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.

Naive Bayes Classifier. The Naive Bayes classifier is based on Bayesian 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.

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 does not 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].

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 biosequence 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 cannot take into account the contextual dependency in the observation sequence. Another limitation is the Markov assumption itself, that is, the current state only depends on the immediate preceding state, which is also inappropriate for some problems.

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 constrains between random variables. In addition, MRF can represent certain dependencies that a Bayesian network cannot, such as cyclic dependencies, wherein it cannot 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 are 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.

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 do not generalize. Pruning can address this problem to some extent though.

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.

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 also construct a hyperplane or set of hyperplanes in a high dimension 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.

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.

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 neutrons (processing units) are connected together via unidirectional signal channels in different layers to mimic the biological neural network. Usually, only neutrons 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 overfitting, and lack a sound theoretical foundation.

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 would have better predictive power due to direct modeling on posterior, have flexible 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, augment small labeled data with larger amounts of cheap unlabelled 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 unlabelled 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 Mahalabobis, 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.

An Overview of Existing Clinical Natural Language Processing 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 [77, 78], 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 [79] 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. [80] 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. [81] 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, for example, 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) [82] 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 [83].

Agarwal and Yu developed two biomedical NLP systems named NegScope [84] and HedgeScope [85], 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 [86] corpus.

Lancet [87] 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 CRFs model [64] for tagging individual medication names and associated fields, and the AdaBoost model with decision stump algorithm [88] 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 health-care providers quickly and efficiently answer the questions that arise during their meetings with patients, Cao et al. [89] 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 [90], support vector machines (SVMs) learning algorithm [49] and CRFs model [64] were used for the question classification and keyword identification, respectively. The system is designed to enable health-care providers to efficiently seek information in clinical settings.

Liu et al. [91] 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 [92] using clinical questions.

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 USA, the Health Insurance Portability and Accountability Act of 1996 (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 are 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 [93] 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 subdomains, 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 become 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 feature 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 presents 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 are 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 exaggerated, 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 indispensible for clinical researchers and informatics professionals. 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 electronic health records (EHRs) to support important clinical research activities.

References

1. Sox HC, Greenfield S. Comparative effectiveness research: a report from the institute of medicine. Ann Intern Med. 2009;151:203–5.
2. National Center for Research Resources (U.S.). Widening the use of electronic health record data for research [electronic resource]. Bethesda, MD. : National Institutes of Health, 2009.

3. National Institute of Child Health & Human Development. Clinical Research and Clinical Trials. http://www.nichd.nih.gov/health/clinicalresearch/. Accessed Aug 2011.
4. Sung NS, Crowley WF, 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;289:1278–87.
5. Harris Interactive. Most physicians do not participate in clinical trials because of lack of opportunity, time, personnel support and resources. Rochester, NY, June 11, 2004. http://www.harrisinteractive.com/news/allnewsbydate.asp?NewsID=811. Accessed Aug 2011.
6. National Center for Research Resources. Clinical and Translational Science Awards. http://www.ncrr.nih.gov/clinical%5Fresearch%5Fresources/clinical%5Fand%5Ftranslational%5Fscience%5Fawards/. Accessed Aug 2011.
7. Garets D, Davis M. Electronic medical records vs. electronic health records: yes, there is a difference. A HIMSS Analytics White Paper, HIMSS Analytics, Chicago; 2005.
8. Garets D, Davis M. Electronic patient records, EMRs and EHRs: concepts as different as apples and oranges at least deserve separate names. Healthc Inform. 2005;22(10):53–4.
9. Wikipedia. File: Electronic medical record.jpg. http://en.wikipedia.org/wiki/File:Electronic_medical_record.jpg. Accessed Dec 2011.
10. Walker EP. More doctors are using electronic medical records 2010. http://www.medpagetoday.com/PracticeManagement/InformationTechnology/17862. Accessed Aug 2011.
11. U.S. Census Bureau. National Totals: Vintage 2009. http://www.census.gov/popest/data/national/totals/2009/index.html. Accessed Dec 2011.
12. Hazlehurst B, Mullooly J, Naleway A, Crane B. Detecting possible vaccination reactions in clinical notes. AMIA Annu Symp Proc. 2005; 2005:306–10.
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;13:281–8.
14. Liao KP, Cai T, Gainer V, Goryachev S, Zeng-treitler Q, Raychaudhuri S, Szolovits P, Churchill S, Murphy S, Kohane I, Karlson EW, Plenge RM. 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, Colecchi JA, Dubey A, Nathan DM, Glaser JP, Kohane IS. 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: Wiley; 1968.
24. Grishman R, Kittredge R. Analyzing language in restricted domains: sublanguage description and processing. New York: Routledge; 1986.
25. Johnson SB, Gottfried M. Sublanguage analysis as a basis for a controlled medical vocabulary. Proc Annu Symp Comput Appl Med Care. 1989:519–23.
26. Bronzino JD. The biomedical engineering handbook. New York: 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). New York: Springer; 2007.
30. Pearl J. Probabilistic reasoning in intelligent systems: networks of plausible inference. San Francisco: Morgan Kaufmann Publishers Inc; 1988.
31. Michalski RS, Carbonell JG, Mitchell TM. Machine learning: an artificial intelligence approach. Los Altos: Morgan Kaufmann Pub; 1986.
32. Manning CD, Schütze H. Foundations of statistical natural language processing. Cambridge: 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 Transact (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 cConference of the Cognitive Science Society, University of California, Irvine; 1985, p. 334, 329.
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. Ann 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, vol. 1. Providence: American Mathematical Society; 1950.
41. Komodakis N, Besbes A, Glocker B, Paragios N. Biomedical image analysis using Markov random fields & efficient linear programing. 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, Munich; 7–12 Sept 2009. p. 227–30.
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; 1995.
50. Brown MPS, Grundy WN, Lin D, Cristianini N, Sugnet CW, Furey TS, Ares M, Haussler D. 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. Proc IEEE Comput Syst Bioinform Conf. 2005:366–74.
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, Association for Computational Linguistics, Sapporo; 2003. p. 57–64.
53. Pan C, Yan X, Zheng C. Hard Margin SVM for biomedical image segmentation. In: Wang J, Liao X-F, Yi Z, editors. Advances in neural networks – ISNN 2005. Heidelberg: Springer; 2005. p. 754–9.
54. Fix E, Jr. Discriminatory analysis: nonparametric discrimination: consistency properties. Technical Report, No. Project 21-49-004, Report Number 4, 1951:261–279.
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. Shanmugasundaram 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. Pac Symp Biocomput. 2005:531–42.
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. IEEE Trans Biomed Eng. 1997;44:132–43.
61. Khan J, Wei JS, Ringner M, Saal LH, Ladanyi M, Westermann F, Berthold F, Schwab M, Antonescu CR, Peterson C, Meltzer PS. 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 Jr FEH, Marks JR, Winchester DP, Bostwick DG. 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: Proc of International Conference on Machine Learning (ICML), San Francisco, CA, 2001:282–89.
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. San Francisco: Morgan Kaufmann Publishers Inc.; 2000. p. 591–8.
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. Pac Symp Biocomput. 2008;13:652–63.
68. Bundschus M, Dejori M, Stetter M, Tresp V, Kriegel H-P. Extraction of semantic biomedical relations from text using conditional random fields. BMC Bioinformatics. 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, Association for Computational Linguistics, Colorado; 2009. p. 115–8.
70. Forgy E. Cluster analysis of multivariate data: efficiency vs. interpretability of classifications. Biometrics. 1965;21:768.
71. Jardine N, Sibson R. Mathematical taxonomy. New York: Wiley; 1971.
72. McLachlan GJ, Basford KE. Mixture models. Inference and applications to clustering. New York: Marcel Dekker; 1988.

73. Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander ES, Golub TR. 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 Suppl 2:ii196–205.
76. Schafer J, Strimmer K. An empirical Bayes approach to inferring large-scale gene association networks. Bioinformatics. 2005;21:754.
77. Sager N, Friedman C, Chi E. The analysis and processing of clinical narrative. Medinfo. 1986;1986:86.
78. Sager N, Friedman C, Lyman MS, others. Medical language processing: computer management of narrative data. Reading: Addison-Wesley; 1987.
79. 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.
80. 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.
81. 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.
82. Haug PJ, Koehler S, Lau LM, Wang P, Rocha R, Huff SM. Experience with a mixed semantic/syntactic parser. Proc Annu Symp Comput Appl Med Care. 1995;1995:284–8.
83. 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.
84. Agarwal S, Hong Yu. Biomedical negation scope detection with conditional random fields. J Am Med Inform Assoc. 2010;17(6):696–701.
85. 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. Epub 2010 Aug 13.
86. Vincze V, Szarvas G, Farkas R, Mora G, Csirik J. The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes. BMC Bioinformatics. 2008;9:S9.
87. 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(5):563–7.
88. Rennie J. Boosting with decision stumps and binary features. Relation. 2003;10:1666.
89. Cao Y, Liu F, Simpson P, Antieau L, Bennett A, Cimino JJ, Ely J, Yu H. AskHERMES: an online question answering system for complex clinical questions. J Biomed Inform. 2011;44(2):277–88. Epub 2011 Jan 21.
90. Cao Y-gang, Cimino JJ, Ely J, Yu H. Automatically extracting information needs from complex clinical questions. J Biomed Inform. 2010;43(6):962–71. Epub 2010 Jul 27.
91. Liu F, Kruse AM, Tur G, Hakkani-Tür D. Towards spoken clinical question answering: evaluating automatic speech recognition systems for clinical spoken questions. J Am Med Inform Assoc. 2011;18(5):625–30.
92. Stolcke A, Anguera X, Boakye K, Çetin Ö, Janin A, Peskin B, Wooters C, Zheng J. Further progress in meeting recognition: the ICSI-SRI Spring 2005 speech-to-text evaluation system. Vol. 3869, LNCS, MLMI workshop 2005;78:463–75.
93. 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.

Part IV
The Future of Clinical Research, Health, and Clinical Research Informatics

Chapter 17
Data Sharing: Electronic Health Records and Research Interoperability

Rebecca Daniels Kush

Abstract Data sharing is extremely important for a number of reasons, and its importance is increasing rapidly as we generate more and more information and data in new areas such as genomics. At the core, the value of data sharing is to allow different technology tools to work together to improve currently antiquated clinical research processes; however, data sharing can also serve to leverage the global uptake of electronic health records to improve workflow and enhance the link between research and healthcare. From the point of view of the patient, data sharing will allow for aggregation of sufficient data to support robust analyses and/or comparisons across studies that will increase the quality of research knowledge gained from healthcare. The concept of data sharing is critical to the advancement of healthcare, which relies on research information for informed clinical decisions.

Despite the recognized value of data sharing for the benefit of patients, which includes all of us, there are inherent challenges yet to be overcome. These include, but are not limited to regulations, trust and patient privacy, slow adoption of information technology and standards, as well as workflow, content and technical issues. This chapter focuses on efforts to address these challenges - in particular, collaborations among standards developing organizations and others to develop, harmonize and support interoperability among the standards and to improve workflows between clinical care and research processes. Currently available opportunities and initiatives with significant promise are identified, yet the importance of a stepwise, iterative approach is recognized.

Keywords Data sharing • Electronic health record • Clinical research • Data standards • Interoperability • Medical informatics • eSource • Workflow • Translational science

R.D. Kush, Ph.D., B.S.
Executive Operations, CDISC,
1500 Wild Basin Ledge, Austin, TX, USA
e-mail: rkush@cdisc.org

R.L. Richesson, J.E. Andrews (eds.), *Clinical Research Informatics*, Health Informatics,
DOI 10.1007/978-1-84882-448-5_17, © Springer-Verlag London Limited 2012

Benefits of Data Sharing

The appropriate use of valuable medical information can impact the feasibility and speed of clinical research and the safety of patients and the public. Aggregation of accurate health data can serve to improve individual healthcare experiences, expand collective knowledge about treatments and diseases, strengthen insights into healthcare systems in terms of effectiveness and efficiency, support public health, and offer other beneficial opportunities. Still, there is a limited understanding of these benefits and of the surrounding issues by the general public and a lack of coherent policy to guide the users [1]. Kahn et al. [2] explored current research policies and regulations in light of the benefits of using electronic health records (EHR) for research; however, the inconsistent requirements across multiple users based upon institutional interpretations and implementations of research policies and regulations remain a significant barrier to EHR providers in response to varied requests for system enhancements. These authors recommend well-conceived models based upon best practices to guide institutions in combining the clinical care and clinical research communities. Authors from the biopharmaceutical industry [3] organized 15 "use scenarios" for research use of EHR data: audit medication workflow, clinical trial data collection, clinical trial recruitment, document management for clinical trials, drug safety surveillance, epidemiology, outcomes research, remote site monitoring, study drug use postlaunch, support regulatory approval, trial subject compliance, understand disease progression, understanding disease mechanism, and virtual phase IV trials. In a study on the potential opportunities for using electronic health records to support biopharmaceutical industry needs, the most cited use cases were drug safety surveillance and clinical trial recruitment.

Clearly, sharing data between healthcare and research has the potential to increase the efficiency of current research and regulatory review and approval processes, to support new observational research, and to facilitate safety monitoring. In particular, there would be tremendous benefits of data reuse for safety reporting. The current paper-/fax-based process sharing of this information between the site and the sponsor or a regulatory authority is so cumbersome that most postmarketing adverse events go unreported. The process typically takes well over 30 minutes, which is unreasonable for a busy clinician who could be seeing another patient during that time.

The frustrations on the part of the clinical research community are similar to frustrations with the clinical care community without EHRs. Information that is collected on patients at many hospitals often ends up in paper archives in health information system databases that have been termed *graveyards*, with terabytes of data rarely looked at after the direct course of patient care [4]. Clinical research information that goes to a research sponsor is also frequently not integrated into useful databases and is difficult to access or analyze after each study is complete. When there are integrated databases at the sponsor site, they have proven to be extremely useful, even eliminating the need for subsequent studies that were thought to be necessary for additional analyses or new indications [5]. Moreover, in the case of safety reporting, the implementation of interchange standards and integration profiles in one study has reduced the reporting time from over 30 minutes to approximately 1 minute and has dramatically increased the reporting frequency [6]. Success stories such as these encourage further exploration into why standards and integration profiles such as these are not used more widely today.

The primary reason to share data is to benefit patients, and each of us is a patient. Regardless of our current health situation or the health of those close to us, we will all benefit from a more rapid translation of research information into knowledge that can be used for clinical decisions based upon safety and efficacy profiles of therapies and best practices. In other words, clinical research informs healthcare. It is likely that uniform adoption of data and content standards in EHRs and clinical research information systems will stimulate new ideas and opportunities for data sharing and subsequent process improvement and knowledge discovery.

Relevant Terms

Often the terms data sharing, data reuse, and secondary use of data are used interchangeably. All imply that routinely collected data in the context of medical care can be reused or shared to support clinical research or related activities. The clinical research activities, in a broad sense, include processes (e.g., eliminating data transcription or reentry, identification and recruitment of subjects, trial registration, regulatory reporting) and knowledge discovery (i.e., where the EHR data itself is used to support a study, analysis, or discovery in, e.g., public health or health services research, outcomes research, epidemiology). The American Medical Informatics Association (AMIA) has proposed a clarification of the primary/secondary use concept: "reuse of health data occurs when personal health data are used for purposes other than those for which they were originally collected" [1].

Another relevant term is eSource, which refers to "source data captured initially into a permanent electronic record used for the reconstruction and evaluation of a clinical study. *Permanent* in the context of these definitions implies that any changes made to the electronic data are recorded via an audit trial." [ICH, CDISC] [7]. The concept of eSource is important when seeking ways to facilitate data sharing without transcription, for example, from electronic health records or electronic diaries for collecting research data.

Data sharing obviously can include other *siloed* sources such as biological and genetic data. Other terms, such as interoperability and directionality, are important and contribute to our definition and conceptualization of data sharing; these are described later in this chapter, as are key concepts, general strategies, and challenges, which include workflow and standards and technology adoption. The nature and status of important initiatives and collaborations to encourage data sharing will be described. At the end of this chapter are explorations of development phenomena for the field of clinical research informatics.

Stakeholders

Historically, basic research, clinical research, and medical care have been viewed as distinct spheres, with different objectives, processes, actors, roles, rules and regulations, ontologies, terminologies, databases, and software. The importance of unifying these spheres has been heralded for many years and is taking upon new fervor most recently with the emergence of "translational science" and supporting informatics research and applications. The harmonization of these spheres is an enormous task and will likely continue for decades to come. However, great strides have been made, collaborations have been forged, communications on the topic are engaging, and standards development organizations are collaborating such that standards between the silos are beginning to converge through harmonization efforts. The infrastructure and technological tools are advancing, but players are recognizing this as not merely a technological problem but a social, political, and organizational problem as well.

The stakeholders are all-important. Those who need to be able to share data, albeit with appropriate privacy and security in place, include, but are not limited to, research partners (study sponsors of any type of study, investigators, institutional review boards, data safety monitoring boards, government funding agencies, regulators, biotechnology and pharmaceutical companies, clinical research organizations, technology providers), study registries and patient or disease registries, physicians referring patients to other physicians, biosurveillance centers or centers for disease control, quality forum and public health centers and patient advocacy groups, and obviously the patients themselves. Benefits of data sharing can be reaped by any of these numerous and varied stakeholders.

Challenges in Data Sharing

Despite general consensus that there are opportunities to improve the quality of healthcare based upon the collective and accumulating information in medical records that can inform clinical decision support, there are challenges that must be overcome. Because of these challenges, widespread interoperability of basic research, clinical research, and healthcare information has not yet been realized, apart from a handful of pilots or "point solutions" in limited geographic areas. Clearly, there is a need to agree on workflows to inform roles [2], and there are regulations and problem areas to be addressed [8]. Additional challenges include trust-related issues (privacy), a variety of disparate approaches to information technology, adoption of standards for data interchange, and content issues due to incompatible terminologies. These challenges are addressed in this section in light of the status and evolution of progress toward data sharing. Subsequent sections introduce key concepts and collaborations toward the goals of semantic interoperability and information linkages between healthcare and clinical research.

Regulations

In the current state, clinical research is still paper-focused, which does not allow rapid, real-time electronic data sharing. The time to develop a new therapy is on the order of a decade, and the translation of research results into clinical care decisions is even longer [9]. The clinical research side of the equation must be transformed in order to leverage this new electronic healthcare information at an acceptable pace, and guidelines and regulations must be reviewed and updated accordingly. Although the history of a paper-based process is still a strong influence for most regulations and procedures related to data privacy in research, we are seeing a change. Certain FDA regulations and guidance have been developed to address the retention of electronic records (21CFR11), and there is FDA guidance on the use of computerized systems in clinical research (CSUCI), but these are still not in sync with the policies for the use of electronic health records such as HIPAA. It is interesting that HIPAA itself was initially developed to encourage portability of data while 21CFR11 was developed to encourage the use of technology in clinical research; however, both of these regulations have frequently been “overinterpreted” such that they have the opposite effect due to fear on the part of the users. Data sharing technologies and implementation will improve when patients, providers, researchers, and organizations understand and trust the regulations, and procedures, for protecting patient data and privacy in a changing electronic environment.

Trust-Related Issues

While data sharing or secondary use or reuse of data can be a frightening concept, and appropriate processes and policies must be in place, patients with chronic diseases seeking improved therapies are advocating the use of their medical information to identify biomarkers and other means of obtaining better treatments for themselves and others who may suffer from the same chronic diseases in the future. The more patients and physicians agree to participate in ethical clinical research, the more information we can gain to improve healthcare—that is, assuming this information can be readily accessible for analysis and reporting.

The majority of data sharing use cases can be achieved in an ethical manner, protecting the security and privacy of the patient and for the ultimate benefit of the patient; however, there are cases of fraud, misuse, and unethical practices that undermine the value and benefit for all of us. For this reason, in addition to GCPs and regulations/guidance, measures and tools such as ClinicalTrials.gov and the WHO International Clinical Trials Registry Platform [10, 11] are being put in place around the globe for research sponsors to be transparent about the studies they are conducting and their results, whether the results are positive or negative. The ethical and regulatory issues are emerging and dynamic.

Slow Adoption of IT (and Standards) in Clinical Research

The adoption of new technology by the clinical research community has proven to be more challenging and slower than many would have anticipated two decades ago when the first "remote data entry (RDE)" tools became available. These tools are now typically referenced as electronic data capture (EDC) since the data capture occurs at the origin (not as these were initially viewed by sponsors—as remote to the data management activities). Granted, the EDC adoption trend has been on an incline, but even in 2006 (when some study sponsors declared that all of their studies use EDC), the overall percentage of clinical research studies conducted using electronic data capture technologies was estimated to be on the order of 40% (paper data collection on 3-part NCR forms accounted for the rest), while in 2009, the estimate was approximately 50–60%. The tools employed are varied, and, in 2004, nearly two-thirds of study sites had more than one EDC (eCase Report Form) application, and half of the sites had more than one ePRO (ePatient Reported Outcome or eDiary) application in use simultaneously. This is still, in 2011, a problem at research sites where an active site may well have between 6 and 12 different EDC tools, each with a different login, data collection requirements, and query resolution process. And, these tools require data re-entry/transcription from the medical chart, whether it is paper or an EHR. Obviously, this is not an ideal workflow process, and it increases the opportunity to introduce errors since the same data could potentially be entered and/or transcribed 4–7 times [12].

Workflow

The workflow and systems design issues that emerge from the reuse of data are critical. They are mentioned throughout this chapter and addressed more in depth in Chapter 3. The preponderance of disparate methods for collecting clinical research data at investigative sites needs to be addressed from a workflow perspective. With the recent incentives for increasing use and adoption of electronic health records, integrating clinical research and healthcare in one electronic setting such as the EHR is clearly part of a logical solution to improve workflow for a clinician involved in both research and healthcare. This has been a key objective of the Clinical Data Interchange Standards Consortium (CDISC) Healthcare Link Initiative since its inception [13]. The mission of the CDISC is "to develop and support global, platform-independent data standards that enable information system interoperability to improve medical research and related areas of healthcare" [14]. It has always been important to CDISC that the CDISC clinical research data interchange standards are harmonized with relevant standards for healthcare. In addition, the CDISC Healthcare Link Initiative has specific goals to: (a) make it easier for physicians to conduct clinical research, (b) collect data only once, preferably in an industry standard format for multiple downstream uses, and, thereby, (c) improve data quality and patient safety [15]. The overarching goal is to better leverage technology and information that emanates from healthcare for research purposes, ultimately to be able to better and more rapidly inform healthcare decisions based upon the latest research results. Specifically, the CDISC Healthcare Link Initiative encompasses a number of projects, including but not limited to:

- eSource Data Interchange (eSDI) initiative, the product of which is a document [16] that contains an extensive review and analysis of the relevant existing global regulations, 12 requirements for conducting regulated clinical research using eSource data collection (e.g., EHRs) in the context of existing regulations, five potential scenario (three based on use of EHRs in research), and other checklists for various responsible parties
- Biomedical Research Integrated Domain Group (BRIDG) Model [17], which is described later in this chapter
- Integration profiles developed through CDISC and Integrating the Healthcare Enterprise (IHE), also described later

Standards-based process improvements such as these address workflow by approaching it with a goal toward achieving "workflow integration" for concurrent clinical research and clinical care.

Content and Technical Issues

It may well be that in any clinical research data sharing scenario, there are certain differences in the original data sources and objectives between settings and purposes (i.e., healthcare vs. clinical research) that will never be reconciled. The purposes for EHR data collection are to support patient care (including longitudinal continuity of care and decision support) of almost endless scope, while research interests tend to be more prescribed, more focused, and more short term. For these reasons, there are different characterizations of "shared" data. Each domain (research and healthcare) has its own requirements and goals, and certain data may be valuable for one case or the other. That being said, there are clearly cases where data are valuable to both the documentation of care delivery and to secondary or enhanced uses such as clinical research; these data represent the "same" clinical constructs, but they may or may not be represented the same. In this sense, efforts at synergizing opportunities for data sharing have the potential to impact healthcare standards and data collection, and current EHR-clinical research harmonization efforts are focusing, obviously, on the areas with overlap. The ultimate goal is to have data standards for the overlapping information that are compatible and support data sharing and interchange, which is why many Standards Developing Organizations (SDOs) are working together for just this purpose. The Joint Initiative Council (JIC) is a group of SDOs addressing global health informatics standards harmonization, including clinical research [18]. In terms of standards harmonization and interoperability, there are "content" challenges and also "technical" challenges, which incidentally involve social and organizational issues. These are described by in more detail in Chapter 18.

Hence, key concepts we must address for effective data sharing include: (a) adoption and implementation of data interchange standards to enable the sharing of information across varied choices of technologies and applications (system interoperability), (b) terminology harmonization, (c) basing the standards and applications on an overarching information model and ensuring they are global, (d) understanding electronic health records with respect to clinical research regulations and guidelines to enable the use of electronic source data (e.g., EHR data) for research purposes, and (e) streamlining workflow and processes at research sites. The sharing of data across systems relies on data interchange standards that enable interoperability between computer systems. As we move into the new areas of translational and personalized medicine, the need to be able to share data, integrate and/or compare data across studies, and deal with large volumes of data will become increasingly critical. Currently, we struggle with the sharing of data among different applications used for an individual clinical research study—often within a single institution. Activities and progress toward addressing such content and technical challenges will be described and addressed in more depth, specifically in relation to data sharing, in subsequent sections of this chapter.

Key Concepts Around Data Sharing

Interoperability

It makes sense to use different information and computer technology applications for different purposes in an individual clinical research study, to be able to select the "best of breed" for activities such as data collection, data management, statistical analysis, and project management. However, this also means that the data need to be either re-entered or transcribed from one system to the next or, more appropriately, shared electronically among these tools. Sharing data electronically from one system or application to another efficiently, while retaining the meaning along with those data and avoiding mapping exercises, requires data interchange standards—both transport standards and content standards—including common terminology.

The IEEE defines *interoperability* as the ability of two or more systems or components to exchange information and to use the information that has been exchanged. The notion of interoperability involves both syntactic and semantic aspects.

If two or more systems are capable of communicating and exchanging data, they are exhibiting *syntactic interoperability*. Specified data formats, communication protocols, and the like are fundamental. In general, XML or SQL standards provide syntactic interoperability. Syntactical interoperability is required for any attempts of further interoperability.

Beyond the ability of two or more computer systems to exchange information, *semantic interoperability* is the ability to automatically interpret the information exchanged meaningfully and accurately in order to produce useful results as defined by the end users of both systems. To achieve semantic interoperability, both sides must defer to a common information exchange reference model. The content of the information exchange requests is unambiguously defined: what is sent is the same as what is understood [19].

Semantic interoperability is not usually an issue with the system itself but rather with the compatibility of the data or information. As an example, consider one study with Gender selections of Female = 1 and Male = 2; another study with Gender selections of Female = F, Male = M, and Unknown = U; another may have five different categories for gender vs. two; and yet another study has Sex instead of Gender in the data collection instructions or data collection form. When the research is completed, such data must be interpreted before it can be aggregated or compared across the four studies. In this common scenario, the data must be "mapped" to a common value set (sometimes referred to as 'normalization') before it can be meaningfully exchanged or aggregated. Not only is the mapping of data a time-consuming and costly activity (imagine mapping problem lists or laboratory test code values—thousands of values), but there is also a real risk of loss of meaning in the final database information [20]. According to this review and computer science literature/experience, the "mapping" of data from heterogeneous systems requires consideration of the data collection context and data model semantics in combination with terminological data values. An example of context would be the use of the word "epoch" in the standardization of study design as a part of a protocol. (See Chapter 9.) Most clinical research protocols include timeframes that are called cycles, periods, stages, or other such terms. In the CDISC Protocol Representation Model—Study Design [21], this has been named "epoch" to standardize on one word that subsumes the myriad of terms that are often used. Epoch has a very specific definition for a clinical research protocol: *Interval of time in the planned conduct of a study*. An epoch is associated with a purpose (e.g., screening, randomization, treatment, follow-up), which applies across all arms of a study. Note: Epoch is intended as a standardized term to replace period, cycle, phase, and stage.

Precise semantic definitions and concepts are critical to enabling computable semantic interoperability. Mead [22] writes of the *Four Pillars of Interoperability* (in the context of information messaging): a common information model, robust data types, a robust infrastructure for specifying and binding concept-based terminology values to specific message elements, and a formal top-down message development process. These four pillars are deemed necessary, but not necessarily sufficient, for computable semantic interoperability.

Directionality

The notion of *directionality* of data sharing is critical. The specifics of the tasks and objectives for data sharing are important in defining system and technical requirements. This chapter does not deal with directionality specifically, although the means to improve workflow to enhance data sharing is addressed in depth later in this chapter.

Harmonization

The pursuit of interoperability between heterogeneous systems, data representations, or other standards is what drives *harmonization*. Harmonization implies a consensus and agreement, in this context, on the specifications for the content and representation of health and research data, yet the term connotes slightly different meanings to different people and organizations. There are several broad approaches: [23] one is to map to terminologies or ontologies (e.g., ontology mappers); one is to translate between different standards, that is, define explicit ("mapping") relationships and equivalencies between overlapping concepts/data, so that they can be compared; one is to create common metadata and data elements; another is to create a use case, use scenario, or other such vignettes and identify standards to support these "information flows"; yet another is to create an overarching reference model and perhaps submodels (domain analysis models) that are less abstract and address a given area of concern (i.e., a domain). An example of a domain analysis model is the collaborative Biomedical Research Integrated Domain Group (BRIDG) model for "protocol-driven research" [17], described in the next section.

To support computerized semantic interoperability, harmonization is agreeably important; however, there is current confusion and varying approaches to achieve it, and it is not entirely clear how domain analysis models, detailed clinical models, clinical element models, templates, common data elements, concepts, and terminology or ontologies that support mapping/harmonization of the same information all compare. The collaborations described in the following section are examples of current efforts to develop and harmonize compatible standards and other activities necessary to support computerized semantic interoperability and data sharing that forge a better link between clinical research and relevant healthcare information.

Collaborations Supporting Data Sharing Between Healthcare and Research

In 2001, HL7 initiated discussions with the Clinical Data Interchange Standards Consortium (CDISC), and the two standards development organizations agreed that they should be working together to ensure that the clinical research standards (CDISC) and healthcare standards (HL7) could and would be harmonized to support the vision that clinical research should gain information from healthcare and, in turn, research should inform healthcare decisions. CDISC and HL7, therefore, signed an agreement to collaborate and initiated a special interest group to explore how this could best be executed. This collaboration has now flourished into numerous joint projects between CDISC and HL7.

Collaboration among SDOs and harmonization of standards are not always easy. While HL7 and CDISC have had many successful collaborative initiatives, there were a few that required significant effort. The benefits, however, do outweigh the challenges [24]. One particular area that remains to be resolved involves the distinction between content and transport standards; basically, CDISC and HL7 ended up having essentially competing transport standards, both in different "flavors" of XML; in fact, HL7 itself has competing transport standards. Another area involved the fact that clinical research domain experts did not comprehend the language of HL7. The resolution of these issues has involved the development of "choices" for transport of the same CDISC content (e.g., the CDISC LAB standard content can be carried by several different transport standards) and in the collaborative development of a domain analysis model, the BRIDG model, Release 3, which has a layer for domain experts to read and a layer directly mapped to the HL7 Reference Information Model (RIM) along with the originally designed BRIDG layer.

The CDISC and HL7 collaboration has now extended to the previously mentioned Joint Initiative Council (JIC). Additional global collaborations of potentially major importance are forming, such as CIMI (Clinical Information Modeling Initiative), an international effort to promote harmonization of detailed information models/specifications of content that are of common value to healthcare and research.

Clinical Research Standards

CDISC holds a niche in the formal development (as a recognized Standards Developing Organization) of standards in the clinical research domain. As noted previously, CDISC also makes a concerted effort to harmonize these standards with the relevant standards for healthcare; this is an ongoing challenge, but steady and significant progress is being made. For the most part, the CDISC standards are "content standards" in the domain of clinical research. The CDISC content standards encompass the information that is collected and exchanged to support any protocol-driven research (a subset of healthcare information). The initial set of CDISC standards addressed the safety information that is collected for essentially all research studies (e.g., demographics, medical history, physical exam, concomitant medications, adverse events—the 18 domains of the CDISC CDASH standard and analogous SDTM domains) [25]. More recently, certain therapeutic area standards have been developed, including those specific to tuberculosis (TB) and certain cardiovascular disorders, and others are in progress. The CDISC content standards began as metadata standards, specifically the data itself along with the information about that data that provides the means to understand the data. For example, if a number is used for weight, the metadata would indicate whether this number is in kilograms, grams, pounds, ounces, or whatever the units are as well as other information about that bit of data. Once these standards were developed, the complementary terminology was then developed for each element/data field, collaborating with HL7, NCI, FDA, EMA, and others. The term 'clinical data element' or 'concept' refers to the data and the metadata and more, such as a set of valid values (i.e., code list, pick list, or value set). The CDISC standards

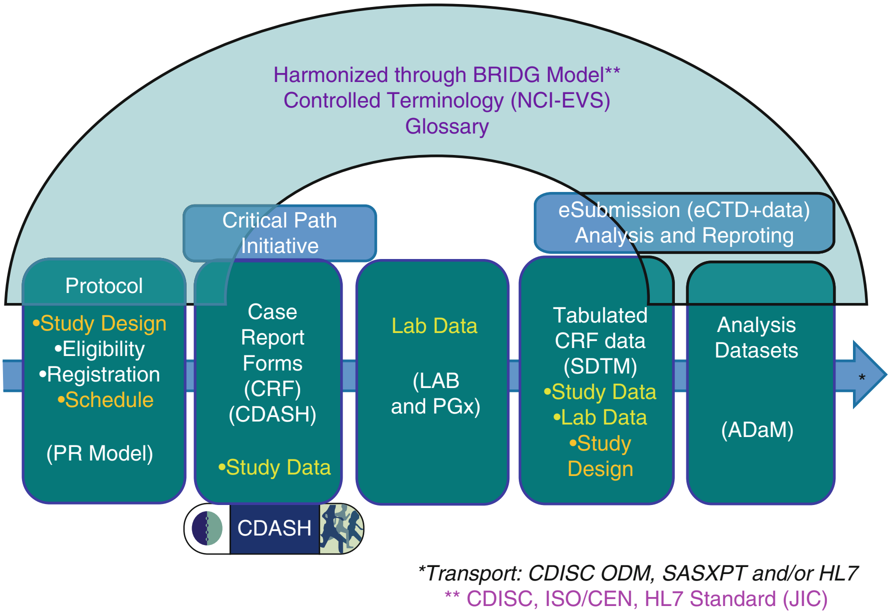

Harmonized through BRIDG Model**
Controlled Terminology (NCI-EVS)
Glossary
Critical Path Initiative
eSubmission (eCTD+data) Analysis and Reproting
Protocol
•Study Design
•Eligibility
•Registration
•Schedule
(PR Model)
Case Report Forms (CRF) (CDASH)
•Study Data
Lab Data
(LAB and PGx)
Tabulated CRF data (SDTM)
•Study Data
•Lab Data
•Study Design
Analysis Datasets
(ADaM)
*
CDASH
*Transport: CDISC ODM, SASXPT and/or HL7
** CDISC, ISO/CEN, HL7 Standard (JIC)

Fig. 17.1 Clinical research standards from protocol through analysis/reporting. The BRIDG model is shown here as a *harmonizing model* for all of the CR standards

now support clinical research end-to-end (from protocol through analysis and reporting), and to ensure that these standards are harmonized with each other to support streamlined data flow from protocol through analysis and reporting, a clinical research domain analysis model (BRIDG, described in next section) is used (Fig. 17.1).

Joint activities to agree on terminology now involve CDISC and HL7 as well as other organizations, including the National Cancer Institute (NCI), FDA, NIH, and the International Conference on Harmonization (ICH) which is comprised of the regulatory authorities and pharmaceutical manufacturing associations from Europe, Japan, and the USA. However, the development and adoption of a global concept-based terminology continues to elude the world and is now commanding attention from an increasing number of standards development organizations working together in order to solve this problem [26], not only for healthcare but also including clinical research terminology. In particular, to encourage global harmonization of standards (including terminology), the aforementioned JIC [18] was initiated with HL7, ISO TC 215 (Healthcare Informatics Standards), and CEN from Europe. CDISC and now the International Health Terminology SDO (IHTSDO) have also now joined this council. It is only through the collaboration of global SDOs and their stakeholders that semantic interoperability will be feasible.

The BRIDG Model

The initial exploration of how CDISC and HL7 could best ensure harmonization of the clinical research standards with the relevant healthcare standards to achieve semantic interoperability between healthcare and clinical research computer systems was inherently difficult since these two groups were speaking different languages in every sense of the word. The proposed solution came in the recommendation to develop a domain analysis model. This was initiated in 2004 by CDISC, and the scope was broadly defined as "protocol-driven research." The resulting model is a collaborative and open domain analysis model, the Biomedical Research Integrated Domain Group (BRIDG) model [17]. The two main reasons for initiating the BRIDG model were (a) to provide a domain analysis model for clinical research that harmonizes all of the CDISC standards and other relevant protocol-driven clinical research standardization efforts and (b) to link clinical research standards with healthcare standards toward the goal of semantic interoperability [25]. The BRIDG model now serves to bridge not only standards but also organizations and clinical research with healthcare.

The BRIDG model now includes the CDISC standards, along with harmonized adverse event reporting (safety reporting) standards from the FDA, NIH, NCI, ICH, and CDISC. The BRIDG model Release 3, which was the first "Production Release," has a layer that domain experts can readily understand and one that is directly mapped to the HL7 RIM, along with the original middle layer. The BRIDG is now being used to (a) ensure that any new clinical research standards that are developed are harmonized with the existing ones and (b) develop interoperable applications, to facilitate meaningful data sharing. Figure 17.2 illustrates the relationship of the BRIDG model to the HL7 Reference Information Model (RIM) and the domain-friendly models developed by the BRIDG Semantic Coordination Committee (SCC).

The first standard that emanated directly from the BRIDG model (as opposed to being harmonized into BRIDG) is the Protocol Representation Model [21]. The mission of the Protocol Representation Group, which developed this model, is "to develop a structured protocol representation that supports the entire lifecycle of clinical research protocols to achieve semantic interoperability (the exchange of content and meaning) among systems and stakeholders." The Protocol Representation Model V 1.0 includes subject eligibility criteria (not specific to therapeutic areas), study design (standard design information, including epochs, and time and events or study calendar), and clinical trial registry information (also supporting project/protocol management and reporting the study summary to FDA). The clinical trial registry elements in BRIDG and PRM have been harmonized to support EudraCT, WHO ICTRP and ClinicalTrials.gov.

The BRIDG model will become increasingly useful as the development and enhancements continue and as users better comprehend its structure and value. It is also important to ensure that, for all of the standards, enhancements and new releases are "backwardly compatible" so that early adopters are supported. As of this writing, HL7 and CDISC (through the BRIDG Board) are formalizing policies to ensure that emerging standards/artifacts, data models in the domain of clinical research are represented in the BRIDG model to ensure that these standards are compatible with each other.

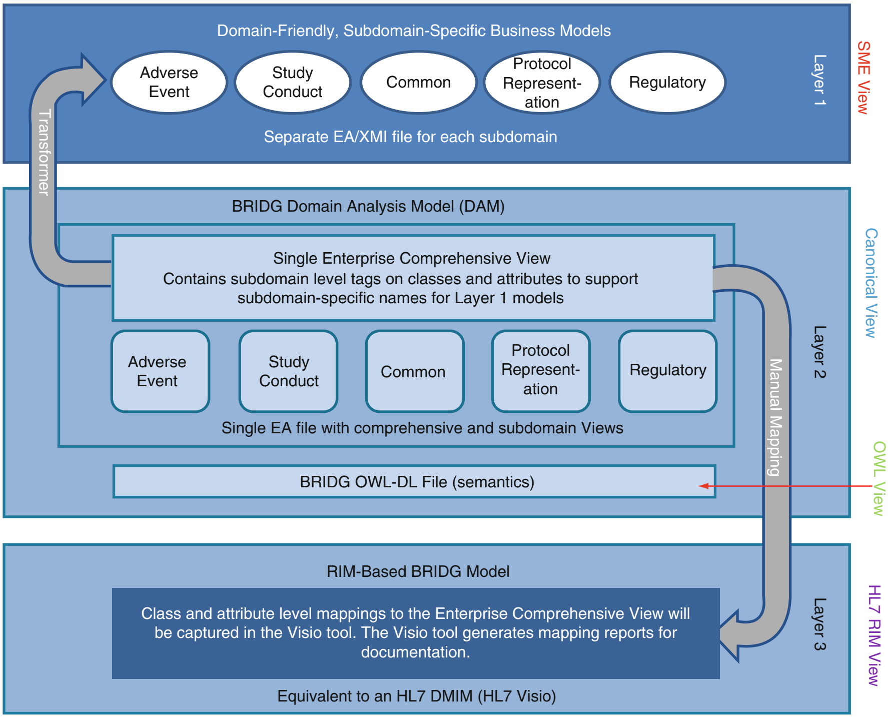

Domain-Friendly, Subdomain-Specific Business Models
Adverse Event
Study Conduct
Common
Protocol Represent-ation
Regulatory
Separate EA/XMI file for each subdomain
Layer 1
SME View
Transformer
BRIDG Domain Analysis Model (DAM)
Single Enterprise Comprehensive View
Contains subdomain level tags on classes and attributes to support subdomain-specific names for Layer 1 models
Adverse Event
Study Conduct
Common
Protocol Represent-ation
Regulatory
Single EA file with comprehensive and subdomain Views
BRIDG OWL-DL File (semantics)
Manual Mapping
Layer 2
Canonical View
OWL View
RIM-Based BRIDG Model
Class and attribute level mappings to the Enterprise Comprehensive View will be captured in the Visio tool. The Visio tool generates mapping reports for documentation.
Equivalent to an HL7 DMIM (HL7 Visio)
Layer 3
HL7 RIM View

Fig. 17.2 BRIDG as domain analysis model for protocol-driven research

BRIDG has also been adopted as a JIC Project, which means it will be not only a CDISC and HL7 standard but also an ISO/CEN standard as well.

For the BRIDG as a whole, as with data collection and reporting standards such as CDISC CDASH and SDTM, specific therapeutic area-specific standards would be very useful as would value sets bound to BRIDG. CDISC is working with other organizations in this area on a project called SHARE (Shared Health and Research Electronic Library). SHARE is defined as “a globally accessible electronic library built on a common information model, which (through advanced technology) enables precise and standardized data element definitions that can be used in studies and applications to improve biomedical research and its link with healthcare” [27].

Developments/Phenomena Driving Data Sharing

In addition to the collaborative BRIDG model, which serves not only to bridge standards and organizations but also to bridge clinical research and healthcare, there are several important initiatives that are taking place to link healthcare and clinical research, leveraging electronic health records (EHR). Three concurrent areas of development include: (a) translational biomedical research to integrate processes from biology and discovery of new therapies through to bedside treatment and the data standards and technologies needed to enable these processes, (b) the use of EHRs to support clinical research of various designs and personalized healthcare and the standards and technology to support these processes, and (c) aggregated databases to support research towards the development of new therapies.

Clinical and Translational Science

In the translational medicine arena, there are a number of technology efforts ongoing within and across academic institutions and by government health organizations and corporations around the globe. Without going into depth in this chapter, these initiatives include, but are not limited to, integrated data repositories [28], data warehousing [4], ontology-related projects, IT infrastructure considerations [29], and data integration initiatives [30]. The Clinical and Translational Science Awards (CTSA) in the USA encourage collaboration and data sharing among research organizations [31]. The Informatics for Integrating Biology and the Bedside (i2b2) project is indeed an exemplary initiative enabling data mining and producing integrated databases and ontology mappings that collaborators may apply to their research projects to augment cohorts and leverage larger aggregated sets of information for their research purposes [32].

The US National Cancer Institute (NCI) has sponsored a number of projects for the purpose of integrating discovery research and translating that to the bedside, including caGrid and caBIG [33]. The mission of caBIG® was to develop a truly collaborative information network that accelerates the discovery of new approaches for the detection, diagnosis, treatment, and prevention of cancer, ultimately improving patient outcomes. The NCI also collaborates with the European cancer communities to try to bring a more global perspective. At the very core of these initiatives, or any data sharing effort, is a set of standards that enable data sharing. The NCI uses established standards when available and has been a key contributor in standards development work by Health Level Seven (HL7) and the Clinical Data Interchange Standards Consortium (CDISC). The NCI is one of the key collaborators with the BRIDG model, along with CDISC, HL7, and FDA, and their applications are built using the BRIDG model to enable data sharing and aggregation for research purposes. Their core data collection standards are harmonized with those of CDISC (CDASH), and the controlled terminology for FDA and CDISC is housed and made available through the NCI Enterprise Vocabulary Services (EVS).

Healthcare Link

As mentioned earlier in the chapter, the eSDI concept is a powerful force changing workflows and research paradigms, thus promising improved efficiencies and knowledge discovery. In addition to the CDISC eSDI document, developed in collaboration with FDA, the European Medicines Agency (EMA) has also referenced the eSDI work in their “Reflection Paper on Expectations for Electronic Source Documents Used in Clinical Trials” for the GCP Inspectors Workgroup [34]. Through the CDISC Healthcare Link Initiative and building upon the eSDI work and the associated scenarios using EHRs for clinical research, Scenario 3, or “Single Source,” was piloted [35]. The concept was then brought by CDISC to Integrating the Healthcare Enterprise (IHE) to be developed into an integration profile. This integration profile, developed jointly by IHE and CDISC, is named Retrieve Form for Data Capture (RFD) [36]. Although relatively simple in concept, the RFD has proven to be extremely powerful in enabling the sharing of key information between EHRs and research-related systems. It is easily implemented by a variety of different EHR systems and has received support from EHR vendors for this reason [37]. The RFD allows a form (from another source) to be brought into the EHR environment, partially prepopulated via the EHR, and then the data entered into the remaining fields by investigative site personnel such that the resulting subset of data can then be sent, in a de-identified and private manner, to various reviewers (secondary users). The use cases supported thus far have been clinical research studies (eCRFs), safety reporting (MedWatch/ICH E2B), biosurveillance/outbreak reports, patient/disease registries, and postmarketing surveillance.

The value comes from RFD in a number of ways, including the following:

- Allowing clinicians to continue to use the healthcare technology/EHR system they are using while readily supporting other uses such as clinical studies (integrated workflow) thus creating efficiencies in research and healthcare
- Eliminating transcription or re-entry of data into multiple systems, thus improving quality (and reducing paper)
- Integrating the adherence to regulations for research and patient privacy into the process (RFD supports a standard archive process for the investigative site)
- Facilitating downstream use of the information from the EHR since it is in a standard format (e.g., the Continuity of Care Document (CCD) and CDASH have been used as standards to support clinical studies using the RFD for the process/workflow component)
- Providing the potential to improve patient safety by linking it into the clinical care session

Safety reporting (i.e., adverse events reporting) at one research center using RFD now takes on the order of 30 seconds instead of the original greater than 30 minutes timeframe [6]. The H1N1 flu threat in mid-2009 emphasized the need for the rapid sharing of information to support an adequate dataset from which to glean knowledge required to address the threat. The CDC was able to use the RFD and controlled vocabulary for their purposes. There are also implementations of the RFD in Europe and Japan, safety surveillance reporting to Japan’s regulatory agency, (PMDA) Pharmaceuticals and Medical Devices Agency.

The RFD is being augmented by additional integration profiles that are in development to streamline workflow from EHRs for research purposes and to ensure safety and privacy and use of the appropriate subset of data (redaction services). Specifically, there is now an integration profile and mapping to support the use of CCD to populate eCRFs using the CDISC CDASH standard and an integration profile to enable the EHR to "know" what data should be collected in the form based upon a protocol (Retrieve Protocol/Process for Execution or RPE).

The Interoperability Specification (IS) #158 for clinical research, developed through the US Health Information Technology Standards Panel (HITSP) initiative and ratified in January 2010, was specifically developed for the use of EHRs in clinical research, initially scoped as exchanging a core set of data between EHRs and research systems. The corresponding value case and use case were produced prior to the IS development steps. The standards/integration profiles identified for this IS are the HL7 CDA, the CDISC-IHE, and the CDISC CDASH as a core (minimal) research dataset. In fact, there are four HITSP IS that employ the RFD—clinical research, quality, biosurveillance, and public health, and, as mentioned previously, the RFD has been endorsed by the EHR Association for its ease of implementation and value.

Leveraging the standards, eSDI and RFD, an HL7 Functional Profile was developed to support the use of EHRs in clinical research. EuroREC, a certification agency in Europe, has also been involved in this functional specification development. These harmonization efforts are critically important in order to help ensure that healthcare standards and interoperability specifications take into account research needs such that there can be convergence of these efforts, rather than a detrimental divergence.

Also in Europe, the European Commission is requesting funding proposals for their Framework Programme 7: "Advanced environment for health professionals and researchers that enable seamless, secure and consistent integration or linking of clinical care information in electronic health records (EHR) with information in clinical research information systems, such as clinical trial systems." And, the Innovative Medicines Initiative (IMI) has funded a collaborative project with 33 partners to develop four use cases; it is called EHR4CR. The Healthcare Link is clearly of global interest.

Aggregated Databases to Improve Opportunities for New Therapies

An unprecedented collaborative initiative that demonstrates the value of data sharing is the Coalition Against Major Diseases (CAMD). A project of the Critical Path

Institute, CAMD has brought together collaborators from the biopharmaceutical industry, government, patient advocacy groups, and other key stakeholders to work together to develop a shared database with a goal to "bring greater speed, efficiency, safety and predictability to medical product development" [38]. The initial target is Alzheimer's disease, which continues to devastate patients, their caregivers, and their families. Aggregated data from failed clinical research studies conducted by multiple sponsors is being used to create a database with sufficient size to enable disease modeling and biomarker validation; a better understanding this disease will presumably lead to more effective therapies. A placebo database can also reduce the number of subjects required for studies and improve analysis of future research. To create the database required a common set of standards; in this case, the studies were all mapped into the CDISC SDTM format, which was augmented by a standard for the efficacy data to evaluate Alzheimer's treatments. In the future, studies can take advantage of the CDISC standards for data collection, thus eliminating the need for back-end mapping of data into a standard and improving the efficiency of the research. The CAMD collaborators are now looking at additional diseases such as Parkinson's disease, thus paving the way for more effectively conducting global research studies and evaluating the effectiveness of therapies in the future through data sharing. CDISC and the Critical Path Institute continue to collaborate in the development of therapeutic area standards to facilitate these invaluable data sharing initiatives.

Conclusion

The notion of information on paper centered around a single patient, or even in computers at an individual clinical care site, must be challenged. Advanced testing and genomic data require computerized systems and the aggregation of information for interpretation and/or comparison through research before they can be useful in a practice setting. The American Medical Informatics Association (AMIA) defines translational bioinformatics as "the development of storage, analytic, and interpretive methods to optimize the transformation of increasingly voluminous biomedical data into proactive, predictive, preventative, and participatory health" [39]. Inherent in this definition is the sharing and processing of data from multiple sites across a spectrum of clinical research. Indeed, as discussed in Chapter 19, even safety evaluations of therapies cannot be trustworthy or useful unless they are based upon sufficient amounts of aggregated data from multiple sources or sites.

The status quo at this writing is no doubt unacceptable to all the stakeholders mentioned in this chapter, among others. Should an insurance company have quicker and better access to greater healthcare information than the research community? Should the insurers be the first to identify an unsafe product as opposed to the manufacturer or regulatory authorities? Should individual states and countries have different requirements that inhibit sharing data across borders? Should government-funded research results be inaccessible beyond the walls of a given institution or investigator? And, finally, should physicians find research and the reporting of adverse events so cumbersome that they choose not to participate? The opportunities for data sharing are vast and could significantly benefit a multitude of stakeholders.

Data sharing is important for a number of reasons, primarily to allow different technology tools to work together to improve clinical research processes and to allow for aggregation of data from different studies and/or comparisons across studies. These steps are critical to the advancement of healthcare, which relies on research information for informed clinical decisions and is rapidly becoming encumbered with increasingly larger quantities of data, particularly in the area of genomics.

Clinical research processes are currently, unfortunately, antiquated and not changing quickly enough. Around half of studies are still done collecting data first on paper, and others are done using disparate systems where clinical data must be re-entered or transcribed, thus increasing the chances of errors and a negative impact on quality. Progress is being made, however. Standards developing organizations are ensuring that standards are harmonized for both research and healthcare by creating a domain analysis model (BRIDG). Academic research institutions are encouraged to collaborate on research, which requires sharing data; integration profiles and standards are being developed and employed to support new processes at investigative sites that will leverage electronic health records to streamline the research process and eliminate burdens on clinical research investigators; and, regulatory authorities and governments realize the importance of ensuring that research and healthcare are better linked. Indeed, the great potential of data sharing is a global concern for the future of healthcare and clinical research. All of the initiatives that rely upon data sharing require policies that protect patients while enabling research, a collaborative attitude among research and healthcare stakeholders (including patients and their advocates) and technology and service providers that implement and support standards development and adoption.

Challenges to informatics include tools, communication, and dialogue to understand current and changing processes, standards, objectives, and systems—for both healthcare delivery and clinical research. Creating a synergy between the two worlds will enable the exploitation of data sources for knowledge discovery and sharing and stand to positively impact all stakeholders. Also, imagination is needed to visualize and anticipate changes in the status quo for healthcare delivery and clinical research. Finally, we must understand ways to achieve interoperability across them all. Opportunities for education, coordination, improved collaborative communication and consensus tools, IT tools, and revised regulations that will encourage the link between healthcare and research (and thus data sharing) also abound.

The nirvana is to reach semantic interoperability, but we may have to take this step by step, all the while ensuring that the standards and technology that are implemented in healthcare will also support research. For these two related areas to develop, divergent or disparate systems and standards would only worsen the current tower of Babel [40] and inhibit the advancement of both research and healthcare. The ultimate benefit is for the patients, and we are all patients.

References

1. Bloomrosen M, Detmer D. Advancing the framework: use of health data – a report of a working conference of the American Medical Informatics Association. JAMIA. 2008;15:715–22.
2. Kahn MG, Kaplan D, Sokol RJ, DiLaura RP. Configuration challenges: implementing translational research policies in electronic medical records. Acad Med. 2007;82:661–9.
3. Kim D, Labkoff S, Holliday SH. Opportunities for electronic health record data to support business functions in the pharmaceutical industry – a case study from Pfizer, Inc. JAMIA. 2008;15:581–4.
4. Prokosch HU, Ganslandt T. Perspectives for medical informatics: reusing the electronic medical record for clinical research. Methods Inf Med. 2009;48:38–44.
5. Bishop S. Presentation at conference of the Drug Information Association presentation, Madrid; 2007.http://www.diahome.org/NR/rdonlyres/E1D7FCCF-6CF2-46E2-A676-817D2639BBE0/765/Program1.pdf.
6. Ibara M. ASTER Study (ADE Spontaneous Triggered Event Reporting), presentation at conference of the Drug Information Association. June 2009. http://www.asterstudy.com/index.php?option=com_content&view=article&id=10:aster-description. Accessed 13 July 2011.
7. CDISC. Glossary, applied clinical trials. 2008. http://www.cdisc.org/stuff/contentmgr/files/0/be650811feb46f381f0af41ca40ade2e/misc/cdisc_2008_glossary_version_7_0.pdf or www.cdisc.org/glossary. Accessed 13 July 2011.
8. Ohmann C, Kuchinke W. Future developments of medical informatics from the viewpoint of networked clinical research. Methods Inf Med. 2009;48:45–54.
9. Lamont J. How KM can help cure medical errors. 2005. KM World www.kmworld.com/Articles/Editorial/Feature/ow-KM-can-hel-cure-medical-errors-9606.aspx. Accessed 13 July 2011.
10. Clinical Trials.gov. Homepage. 2010. http://clinicaltrials.gov/. Accessed 13 July 2011.
11. WHO. International clinical trials registry platform. 2010. http://apps.who.int/trialsearch/. Accessed 19 Dec 2012.
12. Kush RD, Bleicher P, Kubick W, Kush ST, Marks R, Raymond S, Tardiff B. eClinical trials: planning and Implementation. Boston: Thompson CenterWatch; 2003.
13. CDISC. Healthcare link. 2010. http://www.cdisc.org/healthcare-link. Accessed 13 July 2011.
14. CDISC. Mission & principles. 2010. http://www.cdisc.org/mission-and-principles. Accessed Aug 2011.
15. de Montjoie AJ. Introducing the CDISC standards: new efficiencies for medical research. CDISC. 2009;2009:76.
16. CDISC. eSource data interchange. http://www.cdisc.org/esdi-document. Accessed 13 July 2011.
17. CDISC. Biomedical research integrated domain group (BRIDG) model. 2010. www.bridg-model.org or http://www.cdisc.org/bridg. Accessed 13 July 2011.
18. JIC (Joint Initiative Council). Global Health Informatics Standardization. http://www.global-e-health-standards.org/. Accessed 13 July 2011.
19. Wikipedia. Interoperability. 2010. http://en.wikipedia.org/wiki/Interoperability. Accessed 13 July 2011.
20. Sujansky W. Methodological review. Heterogeneous database integration in biomedicine. J Biomed Inform. 2001;34:285–98.
21. CDISC. Protocol representation model – study design – http://www.cdisc.org/protocol or Study/Trial Design Model – http://www.cdisc.org/study-trial-design. 2010. Accessed 13 July 2011.
22. Mead C. Data interchange standards in healthcare IT – computable semantic interoperability: now possible but still difficult, do we really need a better mousetrap? J Healthc Inf Manag. 2006;20:71–8.
23. Weng C, Gennari J, Fridsma DB. User-centered semantic harmonization: a case study. Biomed Inform. 2007;40:353–64.
24. Hammond WE, Jaffe C, Kush R. Healthcare standards development: the value of nurturing collaboration. AHIMA. 2009;80:44–50.

25. CDISC. Standards. 2010. http://www.cdisc.org/standards. Accessed 13 July 2011.
26. Kwak YS, Dickerson A. Just what the doctor ordered-the benefits of health informatics. ISO Focus. 2009;6:35–7.
27. SHARE. CDISC shared health and clinical research electronic library. 2010. http://www.cdisc.org/cdisc-share. Accessed 13 July 2011.
28. Mobed K, et al. AMIA summit posters on ontologies and ontomappers. 2009. http://ctsi.ucsf.edu/informatics. Accessed 13 July 2011.
29. CDISC. Life sciences industry architecture white paper. 2007. http://www.cdisc.org/stuff/contentmgr/files/0/33af249eae6b3d649052bf232d34e6d9/misc/benefits_lifesciencesindustryarchitecture.pdf. Accessed 13 July 2011.
30. Poh Alisa. Integrating lead discovery. BioIT World. 2009;8(4):26. http://www.tectrends.com/tectrends/article/00175337.html. Accessed 13 July 2011.
31. National Center for Research Resources. Clinical Translational Science Awards. 2010. http://www.ncrr.nih.gov/clinical_research_resources/clinical_and_translational_science_awards/. Accessed 13 July 2011.
32. Partners Healthcare. Informatics for integrating biology and the bedside. 2010. https://www.i2b2.org/. Accessed 13 July 2011.
33. National Cancer Institute. caBIG. 2010. https://cabig.nci.nih.gov. Accessed 13 July 2011.
34. EMEA. Reflection paper on expectations for electronic source. 2010. Link to the page: http://www.ema.europa.eu/ema/index.jsp?curl=pages/regulation/document_listing/document_listing_000136.jsp&jsenabled=true. Link to the document: http://www.ema.europa.eu/docs/en_GB/document_library/Regulatory_and_procedural_guideline/2010/08/WC500095754.pdf. Accessed 13 July 2011.
35. 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. JAMIA. 2007;14:662–73.
36. IHE. Retrieve form for data capture. 2009. http://wiki.ihe.net/index.php?title=Retrieve_Form_for_Data_Capture. Accessed 13 July 2011.
37. EHRA. Letter of endorsement for RFD. 2008. http://www.cdisc.org/stuff/contentmgr/files/0/f5a0121d251a348a87466028e156d3c3/miscdocs/ehra_cdisc_endorsement_letter_100908.pdf. Accessed 13 July 2011.
38. CAMD. 2010. http://www.c-path.org/CAMD.cfm. Accessed Aug 2011.
39. Butte AJ. Translational bioinformatics: coming of age. J Am Med Inform Assoc. 2008;15(6):709–14.
40. Kush RD, Helton E, Rockhold FW, Hardison CD. Electronic health records, medical research and the tower of Babel. N Engl J Med. 2008;358:1738–40.

Chapter 18
Standards Development and the Future of Research Data Sources, Interoperability, and Exchange

W. Ed Hammond and Rachel L. Richesson

Abstract In this chapter, the case is made for the use of standards to support clinical research. In particular, the focus is on the standards development process, identification of relevant standards, and selection and implementation issues. Types of standards discussed include physical connectivity, data modeling, information modeling, terminology, organizational process, and documentation of technical specifications. The notion of "certification" in standards activities and the implications for research are introduced. Specifically, an argument for research standards (in the form of common data elements) that are complementary to healthcare standards is introduced. Finally, a review of key advocates for clinical research standards is presented.

Keywords Research data sources • Healthcare informatics • Clinical research informatics • Clinical research data standards • Certification • Conformance • Common data elements

The importance of standards to healthcare informatics has long been recognized and is taking an increasingly prominent place in clinical research informatics. However, the

W. Ed Hammond, Ph.D., FACMI, FAIMBE, FIMIA, FHL7 (✉)
Director, Duke Center for Health Informatics, Duke Translational Medicine Institute

Director, Applied Informatics Research, Duke Health Technology Solutions

Associate Director, Biomedical Informatics Core, Duke Translational Medicine Institute

Professor, Department of Community and Family Medicine, School of Medicine

Professor Emeritus, Department of Biomedical Engineering, Pratt School of Engineering

Adjunct Professor, Fuqua School of Business, 2424 Erwin Road, Durham, NC 27705, USA

Duke University, Durham, NC, USA
e-mail: william.hammond@duke.edu

R.L. Richesson, Ph.D., MPH
Division of Informatics, Duke University School of Nursing,
311 Trent Drive, Durham, NC 27710, USA
e-mail: rachel.richesson@dm.duke.edu

R.L. Richesson, J.E. Andrews (eds.), *Clinical Research Informatics*, Health Informatics,
DOI 10.1007/978-1-84882-448-5_18, © Springer-Verlag London Limited 2012

lack of consensual data representation standards and compliant implementations are still seen as the primary obstacles to achieving widespread use of patient healthcare records as part of a coordinated and efficient national healthcare infrastructure. Internationally, countries with less complicated healthcare models, such as single payer, have made considerably more progress in ubiquitous implementation of electronic health record systems. Even so, there are still significant unsolved problems. In many of these countries, there is still a disconnect between inpatient and outpatient care. Secondary use of data is a universal and cogent goal that is not widely achieved. Effective use of clinical decision support remains a goal for the future. In short, the ability to coordinate patient care or support research across national boundaries is essentially unrealized.

Clinical research has been even slower than care delivery to adopt data standards, and the ability to compare or use data across clinical research controlled trials is almost nonexistent. Paper-based data collection still dominates clinical research and is being replaced more slowly than in healthcare. The motivations for adopting data standards in clinical research have been directly proportionate to the low motivation to share data among researchers, given the highly competitive financial and academic stakes. Despite the obstacles to data standards in clinical research, progress has been made just in the past 5 years. In the previous chapter, Data Sharing: Electronic Health Records and Research Interoperability, Dr. Kush alludes to the sociopolitical and organizational challenges of developing standards necessary for interoperability. The systems and workflow issues related to clinical research data collection are addressed in Chapters 3 and 8. This chapter focuses more deeply on the heterogeneity and features of standards themselves and generic aspects of their development. Informatics has a role to support processes of both standards development and standards implementation. As we discuss the various types and features of standards, it is important – especially in the rapidly growing use of electronic data collection for clinical research – to recognize that there are both *content* and *technical* aspects to data standards. We describe both within the context of a future vision of interoperable health information systems. We propose the need for standards that bring the content and technical requirements together in reusable and transparent ways in order to ensure interoperability between and across clinical research and a spectrum of healthcare and population health activities.

Future Milieu for Standards (The Vision)

For the moment, consider what the value of coordinated health information and research infrastructure, with truly interoperable information systems, would mean. Current and important problems of the clinical research community – such as identifying candidates for clinical trials, obtaining well-informed consent efficiently, locating existing biological samples – would become relatively effortless. The data collected in the documentation of patient care could be used seamlessly to support research planning, observational research (e.g., clinical and comparative effectiveness research), and knowledge discovery (e.g., data mining). Computer-directed guidance could enhance the quality and consistency of data collection, by providing actionable guidance so that the data are collected by the most appropriate person (or device), at the proper time, according to protocol specifications. Issues of *provenance* (i.e., who recorded, changed, and authenticated each piece of data) could be reliably assured. Patient data would be available to facilitate the continuity, safety, and quality of patient care, as well as support clinical trials, data mining, reimbursement, audit and performance measures, and other purposes with little additional effort [1]. Data for research protocols could be conveniently collected at the time of clinical visits, or research visits could be managed separately. The data specific to research could be collected as specified in the research protocol, and the protocol events could be directed without significant additional cost. Adverse events for protocols could be managed within the context of comprehensive patient care records, so that health providers would be aware of research participations and interventions in their patients, and researchers (and regulatory and patient safety organizations, see Chap. 19) could be aware of emergent health issues (i.e., possible adverse events, safety issues, or confounding factors) that present to the healthcare providers. Multinational research studies could utilize clinical data or be supported. In a world with truly interoperable health information systems, public health reporting would be timely, accurate, and complete. Population experience would influence the nature and funding of new research, and subsequent discoveries could be put into practice and evaluated quickly and continuously, as described in the final chapter of this text (Chap. 21). Quality of care could be explored – near real time, within and across organizations and national boundaries, and process and interventional features could be performed. Informatics, enabled with interoperable systems, could indeed create a Brave New World. All of these scenarios require the sharing of patient data, which is dependent upon a consistent – or standard – representation of the data and its context.

Features of Standards

The word “standard” is used frequently in our conversations to refer to many different things. Sometimes we mean a formal document that was developed using prescribed procedures, balloted through an open, consensus process, and thus has received a stamp of authority. Sometimes we merely mean an informal set of rules by which we create expectations of behavior. Sometimes we mean a specified list of codes that reference concepts. We often confuse standards with operating instructions or legal regulations. This variety of definitions for “standard” is part of the confusion – since user communities cannot easily assess which standards are overlapping or are appropriate for their domain or purview, nor can users quickly plug-and-play standards into their own information systems and applications. In this chapter, we use the term standard broadly to mean consensual specifications for collection or exchange of data. Readers should be cognizant that these standards can take many forms.

A few generic features of a standard are worth pointing out to better explore the heterogeneity in standards and the challenges of assimilating them for meaningful data exchange. In the next paragraph, we will describe some examples of various layouts of standards but only with the intent to emphasize that there is variety, which contributes to the difficulty of achieving the vision of integrated healthcare and research systems. Secondly, all standards imply some level of consensus and shared use, both of which imply the notion of multiple individuals or groups. Basically, standards are accredited or endorsed by some process or authorizing body. Standards can be produced by several methods: ad hoc, de facto, government mandate, and consensus, as described in [2]. Other important aspects of standards include conformance and certification (of implementation), and although these issues are not addressed here, they are an important aspect of standards compliance, expense, and popularity. There are also some specific categories of standards such as informative standard (information only) or normative standard (must be explicitly followed to be compliant), and to further complicate matters, the coexistence of draft standards (for trial use) and multiple versions illustrates the fact that standards are ever changing, as needs and knowledge continue to evolve. Individuals or organizations actually *create* standards using domain expertise, a defined process, and mutual understanding of the exchange or cooperation problem that the standard is intended to address. Consensus standards become official through some type of industry-balanced balloting process. Critical to the ability to use any standard are implementation guides, which may be developed by the organization developing the standard, by another related group, or by a single user (individual or an organization) for its own purposes. In other words, regardless of the source, there can be implementation guidance or standards, and this should include where, when, how, and by whom a standard is used.

In all of these standards development processes and steps, there are conceptual, methodological, procedural, and logistical issues related to distributed collaboration, communication, and consensus. There are also conceptual, methodological, procedural, and logistical issues related to group identity, membership, governance, funding, uptake, and "public relations." Solutions to standards problems will no doubt require contributions from many disciplines, including cognitive science, organizational psychology, sociology, education, communications, political science, and, of course, information science and technology. These issues, though seemingly mundane, all impact the quality of the standard and also the perceptions (of stakeholders and potential users) about the utility, quality, completeness, and equitability (fairness) of that standard, and all of these features – we presume – have a direct impact on the use. Undoubtedly, there are gratifying, exciting, endless opportunities for informaticists within clinical research domain and other healthcare areas to explore ideal development formats and features for standards. Regardless of domain or training, those engaged in clinical research in the contemporary era should be familiar with standards and standards developing organizations. Accordingly, we encourage readers to inform and engage in relevant standards activities – to ensure that standards remain relevant, useful, and usable as they evolve.

International Landscape and Coordination

Often, people refer to standards with the 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 enterprise flow together. This has not been the case with healthcare information systems to date, though global and national standards efforts need to embrace this coordination approach. As we have mentioned, the interaction between EHR and clinical research systems is critical and will have strong influence in redesigning the clinical research enterprise. As an increasingly mobile society, the need for EHR and patient data to flow across international borders is clear. Similarly, as noted throughout this text, research knows no national borders. The need for multinational coordination and cooperation in research is clear. Because of this, the challenges for meaningful standards are tremendous, but facing them is inevitable. Seeing the number of organizations involved in the standards landscape, as shown in Fig. 18.1, can give readers an appreciation for the complexities of identifying and harmonizing standards.

Figure 18.1 illustrates a number of Standards Developing Organizations (SDOs) that exist and are creating standards. [Note that the figure is a comprehensive, but not exhaustive, list of organizations.] Appendix 18.1 identifies each of the acronyms and organizations depicted on this figure. The figure illustrates several different kinds of organizations. The standards developing organizations are international (CDISC,

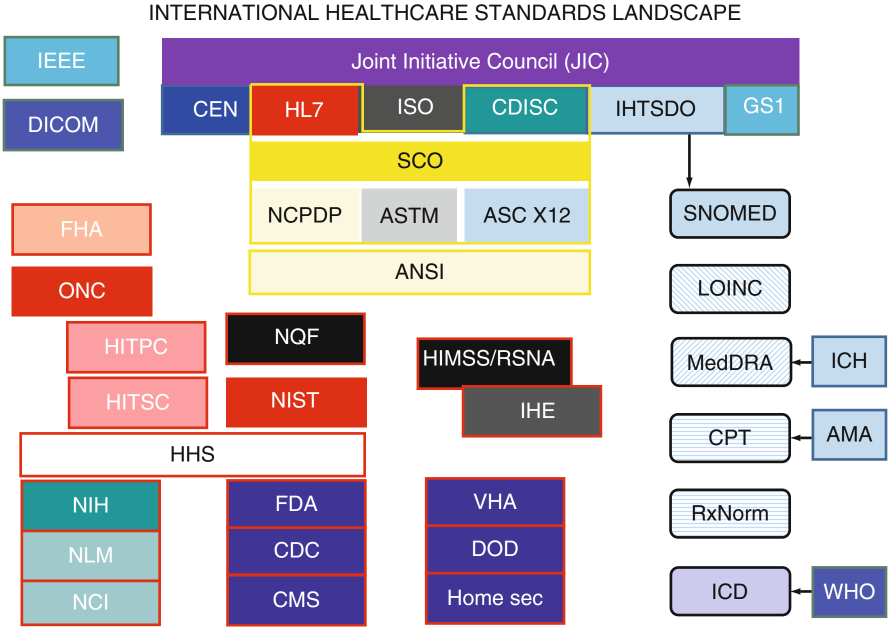


Fig. 18.1 The International Standards Landscape (acronyms defined in Appendix 18.1)

CEN TC 251, DICOM, GS1, HL7, IEEE, IHTSDO, and ISO TC 215) 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 do not formally participate 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 are controlled terminologies that are both international (SNOMED, MedDRA, ICD) and domestic (LOINC, RxNorm, CPT) in scope. The other boxes represent US federal activities as part of the Office of the National Coordinator for the Department of Health and Human Services, which is driving the efforts at 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 an increased role is identifying and testing standards. Even though there are many players and many standards, there is movement toward harmonization and cooperation among the different groups.

Mapping

Despite emerging and promising cooperative efforts, there are still, unfortunately, many overlapping and competing standards addressing all aspects of healthcare systems. The most common approach has been to allow the coexistence of overlapping standards by supporting mapping efforts between the standards. The Unified Medical Language System (UMLS) in the USA (globally available) has facilitated the mapping of various terminologies and coding systems.

Mapping is the process of finding a concept in a target terminology that “best matches” a particular concept in the source terminology, although what “best matching” 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 often are not an ideal or easy solution to a lack of a uniform standard. Mapping between two standards must always result in a 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 workaround and not a solution.

The organizations represented in Fig. 18.1 and defined in Appendix 18.1 represent many, 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.

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 in increasing use worldwide. In a very separate organizational effort and distinctively different modeling approach, HL7 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 – using XML 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 for imaging diagnostic reports. ASTM has created the document standard Continuity of Care Record (CCR) for the exchange of patient summary data. HL7 has the Clinical Document Architecture (CDA) standard. As a harmonization effort between two SDOs, HL7 took the content of the ASTM CCR and implemented it in the CDA standard. This product, called the Continuity of Care Document (CCD), is essentially an implementation guide using the HL7 CDA standard. Within HL7, there is a new and dedicated effort (called "Fresh Look") that is attempting to bring all of these activities and specifications together. As this effort's name implies, the effort will lead an unbiased and critical look at what has been done after 25 years of standards and see what new approaches might be effective. This initiative is striving to break the cycle of backward standards compatibility and eliminate interoperability barriers that are intrinsic to fundamental designs of early standards models. If successful, the Fresh Look effort might drive tangible achievements toward interoperability over the next year.

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 electronic health record vendors to define profiles using suites of standards to achieve, or at least contribute to, end-to-end interoperability. In many countries, the government identifies, and in some cases mandates, which standards are used for what purposes. In the international scene, such profiles will require global governance and vision across all healthcare domains and businesses. The European Union has projects underway that will enable a multinational coordination process.

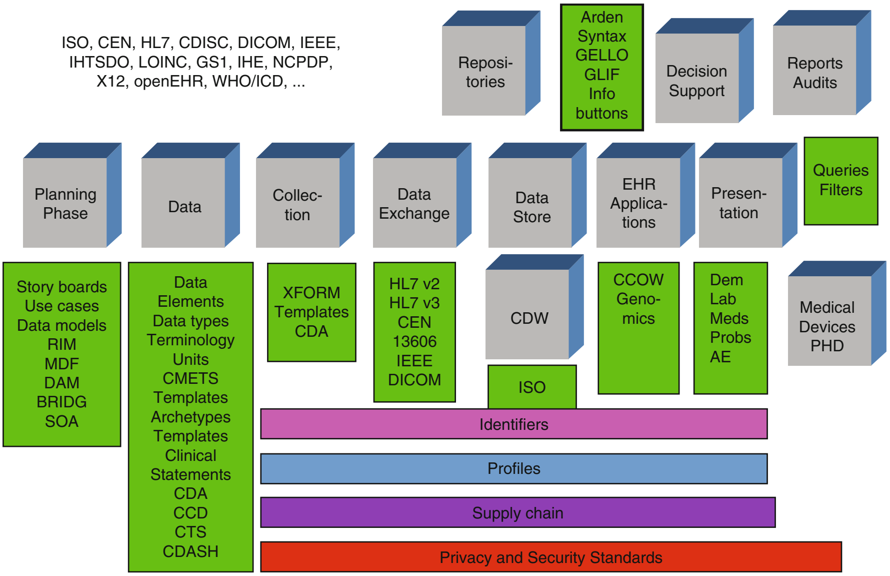

ISO, CEN, HL7, CDISC, DICOM, IEEE,
IHTSDO, LOINC, GS1, IHE, NCPDP,
X12, openEHR, WHO/ICD, ...
Reposi-tories
Arden Syntax GELLO GLIF Info buttons
Decision Support
Reports Audits
Planning Phase
Data
Collec-tion
Data Exchange
Data Store
EHR Applica-tions
Presen-tation
Queries Filters
Story boards Use cases Data models RIM MDF DAM BRIDG SOA
Data Elements Data types Terminology Units CMETS Templates Archetypes Templates Clinical Statements CDA CCD CTS CDASH
XFORM Templates CDA
HL7 v2 HL7 v3 CEN 13606 IEEE DICOM
CDW
ISO
CCOW Geno-mics
Dem Lab Meds Probs AE
Medical Devices PHD
Identifiers
Profiles
Supply chain
Privacy and Security Standards

Fig. 18.2 Standards specifications (*green*) by function (*gray*) (acronyms defined in Appendix 18.1)

Standards by Function

The challenges with achieving collaboration, consensus, and coordination in standards go beyond the obvious geographic, cultural, and language issues in trying to unite international communities. Often, these challenges also involve different scientific and professional communities, who all speak different “languages.” Most IT standards are sophisticated and complex and require technical experts to evaluate, develop, and implement. Additionally, the stakeholders are either domain (clinical) experts or research implementers. The technical and content experts differ more in their intended use. Study design and authoring, implementation and monitoring, data collection, transfer, and storage all require different players, who approach problems and discussions differently.

Part of the reason there are so many standards to be coordinated is because there are many different processes related to the collection, storage, transfer, and use of data in healthcare and research work processes. Figure 18.2 presents a high-level view of the key processes underlying all of the data interchanges: planning, data (representation), (data) collection, data transfer and storage, and applications that address the use and presentation of data. Standards exist at each of these levels, designed to perform a specific function. These processes 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.

The purpose of Fig. 18.2 is to provide an overview of the components and of the large number of standards that exists in each of the areas. These standards have been created by a number of US and international standards bodies – sometimes working independently, sometimes working together; sometimes working competitively, sometimes working harmoniously. Other standards exist that are also important to healthcare applications. There are now a number of regulatory standards developed by HL7, ISO, and others in this area and adopted by the US Food and Drug Administration [5]. 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 presence of multiple transport standards (HL7 versions 2 and 3; CEN 13606, DICOM) for the exchange of data can also be seen on Fig. 18.2. The existing standards have been developed by many different SDOs and are mostly focused on specific applications. These standards are redundant, overlapping, and competing. The most common form of a data interchange 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. This standard, the first standard developed by HL7, starting in 1987, had, as its purpose, the exchange of data for building a "best of breed" hospital information system. Created at a time of limited bandwidth and computing power, the 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. Generalized data elements such as lab tests with results are defined by a name-value pair. HL7 also has introduced a more robust and sophisticated model-based exchange standard, version 3, which enables interoperability through the use of a Reference Information Model (RIM). The HL7 CDA standard and its specific domain implementations (harmonized with the RIM) can also be used for data interchange.

Other data transport standards include the Digital Imaging and Communications in Medicine (DICOM), which is used universally for echanging images. DICOM evolved out of the American College of Radiology and the National Electrical Manufacturers Association (ACR/NEMA) and is now a global SDO. As a US SDO, the National Council for Pharmacy Drug Program (NCPDP) has created a set of standards for e-prescribing and reimbursement for drug prescriptions. Another US SDO, the Accredited Standards Committee X12N, has created a set of data exchange standards to support the reimbursement process. ASTM created the CCR standard as previously noted that may also be used for data interchange. IEEE, an international SDO working with ISO, CEN, and HL7, has created a family of standards to support moving data from medical devices to electronic patient database. ISO TC 215 – health informatics – supports messaging standards primarily through harmonization with other SDOs.

Standards for the storage of data are still an open issue. For the most part, certainly as part of an electronic health record system, the EHR architecture is a proprietary issue. CEN, in its EN 13606 standard, suggests an EHR architecture; others have suggested that the EHR is a set of CDAs or CCR documents. Data is most frequently used and presented independent of its collection. Consequently, data must be stored in its most finely grained form to enable efficiency and maximum utility. However, there will appropriately exist data directly stored with its modifiers. An example is a heart murmur will have sets of modifiers that include location, timing, and other attributes.

Selecting and Evaluating Standards

For every data standard, there is a need that motivates the use of standard. A use case is a narrative scenario of a real data exchange situation, and can be developed to support the development of meaningful standards. A use case must include all relevant parties or systems that generate data or use the systems. The use case forms the basis of requirements, which can be functional or representational requirements. These requirements provide the criteria to select from existing standards or to develop new ones and also form the criteria with which to evaluate the standards and systems selected. The appropriate scope and content of the use case are critical to the success of a standard: if the use case is too constrained, the standards will not accommodate important uses. For this reason, a variety of *multiple* use cases and the engagement of *all* stakeholders are important. Although this does take time, the lack of complete use cases and stakeholder requirements is often the reason that standards are continuously being expanded and changed.

If using an existing standard, it must be accessible – requiring a physical connection and access procedures. Additionally, issues of licensing and managing versions, as standards are always changing, must be addressed. In many situations, the physical structure, format, costs, and update schedules for the standards are important criteria for the standard in the first place. The standard must be incorporated into existing systems and workflows. With the use of terminological data standards in particular, there are several approaches to implementation. Healthcare or research staff can be trained to code at point of care or research observation [7], or the data can be coded centrally in batches after the patient visit, as the FDA does with MedDRA coding of adverse events, which are reported as free text and coded later. (Many readers might be surprised to know how much data is actually collected in an unstructured fashion and coded later in this manner by an external party.) Regardless of the approach for implementing the standard, there are technical development aspects and user training aspects to consider. Ultimately, the standard is evaluated. This evaluation can then produce updates or improvements to the standard or to the systems and workflow processes for using that standard or could motivate the rejection of one standard and the use or development of a different standard.

Functional requirements for data modeling and exchange standards can be evaluated by successful demonstration of use-case activities. By far the most common evaluation metric for a controlled vocabulary or content standard is coverage – in other words: content, content, content [8, 9]. Other criteria include ease of use, precision, and recall [10–13]. Characteristics of useful vocabularies are certainly fundamental reading for anyone new to data standards [10]. The complexities and hierarchies of the representation of medical data first described by Blois are also interesting reading [14].

Specific Standards Relevant to Clinical Research

Standards are not homogeneous and are not plug-and-play. 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 have a history of accepting emerging domains if they are relevant to health data exchange or EHR systems). There is not always coordination – or even communication – among standards on scope or content, so 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. [Note that there is a formal collaboration and coordination between IHTSDO and LOINC. However, they are still different standards, represent different curation and user communities, and will continually be developed separately and independently for the foreseeable future.] 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 is crossing the fine line between "standardized" assessments to data elements. The list of overlaps and expanding scope goes on and on, but essentially the success and growth of standards are driven by familiarity, marketing, and social issues. Each of these standard terminologies has (and will have) a specific user community, various stakeholder groups, business considerations, and a desire to accommodate the evolving needs of users.

In clinical research, the notion of standards includes standard case report forms, which are collections of data elements. In study implementation and development, these data elements are the items or questions on forms. They can be reused (including the formatting and layout work and inclusion in the database) and thereby increase efficiencies in protocol implementation and data capture system design. A recent review of CRF (Case Report Form) data standards organizes existing CRF-related standards into three types: structural features of forms and data items, content standards, and specifications for using terminologies [15]. In response to the Food and Drug Administration (FDA)'s 2004 report, "Innovation/Stagnation: Challenge and Opportunity on the Critical Path to New Medical Products," a CDISC effort, Clinical Data Standards Acquisition Standards Harmonization (CDASH), addresses data collection standards through standardized CRFs [16]. Initial CDASH standards focused on cross-specialty areas such as clinical trial safety. Disease- or therapeutic-specific standards are now being considered, along with tools and process development to facilitate data element reuse across diseases. Because clinical research is highly protocol specific, forms-development *processes* are more easily standardized than is CRF content, and current CDASH standards include guidance on forms development, including multidisciplinary expertise in their development. The CDASH standards document, "Recommended Methodologies for Creating Data Collection Instruments," presents important and necessary features of the CRF development process. The techniques described include: adequate and "cross-functional" team review, version control, and documented procedures for design, training, and form updates. The FDA also requires rigor in the development, validation, and use of data elements related to patient-reported outcomes as study endpoints in investigational new drug studies [17].

Standard Data Elements

Significant progress has been made toward data transport and terminology standards in medicine; however, standard data elements and definitions for information generated and used in care are lacking, resulting in our inability to leverage computers to exchange and use information between and across healthcare and research settings, and ultimately has hindered achieving the vision of integrated health described at the start of this chapter. Data may be exchanged between care providers, but variations in meaning, measurement, recording, formatting, and coding systems limit their effectual use.

A data element is a discrete unit of data collection that is clearly defined and reusable. Data elements can be viewed as the fundamental unit of data exchange across integrated health information systems [18–20]. Data element–based information exchange standards in other industries (e.g., commerce, library sciences, environmental sciences) support the growing consensus that the data element is the level at which specification must occur to support semantic interoperability in healthcare and research industry. Data elements do indeed have a conceptual model as we describe below, but the attributes of this model are designed to capture data collection features of the data element. Data elements can include concepts from specific controlled terminologies, but often data elements have narrative clinical definitions that go beyond clinical concepts to capture time, role (e.g., observer, reporter), and clinical definition.

The development of reference sets of data elements is the most important means to standardize data. Certainly, the development of standard data elements is the logical research approach, given that research deals with variables and values. CDISC is developing standards using a variable (or data element)-based approach. Their first standards identified generic, pan-disease data elements, and now data specific. NCI has had the caDSR data element repository for many years [21]. In research, the idea of shared data elements is intuitive. Probably because the currency of research has been – and will always be – variables that have standard and structured value sets (codes) and definitions. The notion of a data dictionary is accepted best practice in statistics and clinical research practice. The quality of that data dictionary might need some work, and the use of shared common data elements can support consistent high-quality definitions that promote consistent use of data elements, leading to more comparable data. The notion of data elements, shared definitions, and metadata are a new paradigm for healthcare but a necessary shift in thinking and standards, especially in the new age. In the past, EHRs would have to just produce predictable standard codes – like report ICD-9 codes. They have never yet been challenged for semantic data exchange. This is a new challenge and an opportunity for research and care delivery to synch together.

We provide some guidelines and important features for data elements. The data element first must remove any ambiguity in using a rich set of attributes to not only provide structure and definition but also include related knowledge representation, linkages, and operational characteristics. Language in itself is not an absolute science. Words are frequently used with different meanings between the speaker and hearer. A study of the literature, done in 2011, identified 67 different meanings and use of the phrase 'unstable angina'. Yet, on a data-gathering form for a clinical trial, there is one box for unstable angina. Which of the 67 different meanings did the author mean? In healthcare, this ambiguity can be dangerous. Furthermore, we believe the definitions must be made by a group of domain experts acting as judges – not by a consensus process.

The data element is represented by a single code that in itself has no meaning. It is absolute and functions as an index into a metadictionary that contains the data element and all of its attributes. A core set of essential attributes should accompany data elements to ensure that their semantics and context are adequately represented to support meaningful exchange and reuse (i.e., interoperability). These attributes include: numerical code, definition, long and short names, synonyms, units, data type, value set, class and categories and domain (for indexing and classifying the data element), purpose, and language. Note that data element definitions should be addressed using multiple attributes for both human readability and computer action using description logic. Human-readable definitions should be precise and unique and clinically meaningful. Additional attributes can provide more definitional information, such as how the value is measured (e.g., removal of shoes for weight or height measure). Other attributes can be used to connect data elements to other standards. For example, a “relationship links” attribute can identify specific causes and manifestations of a clinical problem using external ontologies and controlled vocabularies, and a “triggers for services” attribute could be used to connect clinical data with decision support algorithms to display relevant alerts or prompts for additional research data collection.

Fundamental to each of the stakeholder groups across the informatics-using community are a set of data elements that define their universe; in other cases, these data elements may be shared with other groups. We propose, however, that practically, there can be only one steward of a data element, and that steward should have complete authority over the definition of that element. A second level association is the affiliate steward, and there may be several of these, would be able to strongly influence the set of attributes but would not have ultimate authority. The remaining community could offer comments and opinions, but the authority rests with the judges. Therefore, additional "administrative attributes" that should be associated with a data element will include steward, affiliated stewards, version, and date of last activity, all of which will be critical in coordinating content and data elements across many practice and research domains.

It is important that all related stakeholder groups participate in this process, and the groups must be international. Many of these groups would be defined by the clinical specialty organizations and other organizations. These groups would be required to register and declare their domain of interests. They would agree to a common process and identify the judges. We fully recognize that many data elements that support clinical care have little or no secondary use or research value purpose. We also recognize that many research elements would not be appropriate for an EHR. Therefore, we can envision that professional and scientific communities might develop certain content standards and maintain their own research data element repositories, such as the caDSR, with highly specific elements. But we also recognize that, over time, these specific elements would be increasingly integrated into EHRs, as new and more frequent successes in translational research will change the standard of care, and hence add new requirements to the primary healthcare data documentation. Certainly, as EHRs become more sophisticated (driven in the USA by legislated "meaningful use" requirements and ever-increasing attention on usability and cognitive science research) and widely adopted in all specialty practice domains and care settings, we can envision the need for many, many highly specific data elements. Clinical research – though not new to the concept of data elements – would nevertheless need to be represented as a major stakeholder in these healthcare standards discussions. The important role of clinical researchers acting as motivated stakeholders in the development of clinical documentation standards sufficient to support secondary research use is an area to watch and engage in [1, 22]. Undoubtedly, this challenge will require coordination, collaboration, and strategic planning on an unprecedented scale.

There is no consensus process for developing collections of domain-specific data elements (also called clinical content standards), although there have been successful demonstrations with documented efforts [20]. This is an open research area and one that the US FDA is trying to understand. The process should include the following: open submission process, transparent governance and operations, vetting (for representational and content features; implications for other communities), stewardship, and management (duplications, updates, status, and versioning). The HL7 standards development methodology and process (www.hl7.org) is a likely candidate, as it is international in scope and designed to address all health-related information flows, including those of clinical research. HL7 will also be well poised to ensure data capture and exchange standards that can address multiple needs, including research. The American College of Cardiology (ACC) has with much success collected research data elements as part of registries, in the context of healthcare provider organizations, and these registries have been used as evidence to effect patient care changes [23]. The registries are based upon many standards developed by ACC and others. These professional groups remain engaged in the development of standard data elements for cardiovascular care and are leaders in developing *content-based* standards. The process and formatted standards (for distribution) are being developed as pilot projects within HL7 CIC–Clinical Interoperability Council. The development of content-specific data standards is described in depth in [20] for tuberculosis and cardiovascular. These groups have fully described their experience and results from their work developing a process for (1) engaging stakeholders from primary and secondary data uses, (2) identifying data used and generated in patient care, and (3) providing authoritative natural language definitions of those data elements. These efforts use the ISO 11179 data element as defined by ISO 11179 [24] and UML class and activity diagrams [25] to document the data and their definition, and refer to this documentation as a clinical domain analysis model (DAM). The goal of these pilots was to develop an effective, sustainable, and generalizable process for developing data elements in therapeutic practice areas.

Important Principles of Standards

This chapter has conveyed the message that data standards are highly heterogeneous and have many different formats, scope, organization, and objectives. They emerge from many types of organizations and entities, with varying levels of cost and access, and different governance and curation operations (i.e., the politics related to what gets in). Understanding this diversity and lack of shared vision across standards developers, designators, and stakeholders/users is important to appreciate so that clinical research can effectively engage, and effect constructive change, in data standards.

Because harmonious data standards (with shared semantics) between healthcare and research are essential for the Brave New World we described at the start of this chapter, we believe that successful standards for the future will be at this data element level. This represents a major paradigm shift – especially for healthcare – but it is a strategy that will allow health information exchange and reuse for research purposes. The approach will require data element *registries* that can support the structured collection of data in healthcare and in research using standard questions and answer sets. In the world of clinical research, this is akin to the questions on case report forms that translate to variables used in analysis. There can certainly be some autonomy and efficiency for the clinical research world, but there also should be pathways, mechanisms, and cultural forethought about connections between research data collection and future healthcare delivery. Some data elements will be relevant only for patient care contexts, some will be useful only to research, but most data elements should and will exist in both domains. If we are to enable translational informatics, clinical data elements and research data elements must coexist, and informatics professionals can develop applications and systems that facilitate coordination and synergies between clinical and research communities.

While technologies are changing continuously, standards are evolving, and legislation and multinational, multiorganizational collaborations changing, we provide some general observations and advice regarding standards. We put forth some basic principles that we feel apply to all standards efforts and will be helpful to keep in mind for new standards endeavors across the variety of project applications and standards needs for CRI for years to come. Essentially, these are significant lessons learned. We doubt that these are controversial, though we will say that these are the opinions of authors. Undoubtedly, this list is incomplete, but it is a start. We have already established that standards are not standard, and that standards organizations are not standard. We also assert that:

1. *Standards are multifaceted* and must address several key areas. Standards include artifacts that specify the information *content* of a clinical domain, as well as a *representation* for that content. Standards also represent the *process* by which they are developed and maintained. The criteria for evaluating standards address all of these areas.
2. *The purpose and context of the standard must clearly be defined*. This is a fundamental theorem for informatics and vocabulary development in general. The notion that a terminology can only be assessed or evaluated in the context of its intended use is critical [26]. Many standards are developed for a given context but used in research only because they are the only representation widely used in available electronic data resources. (The persistent use of International Classification for Diseases (ICD) codes for various research purposes, despite the well-known shortcomings of ICD for research or clinical documentation, is a classic example.) But what we hope the reader will take from this is that the purpose of the standard must accommodate for the ideal vision of data sharing and interoperability that we present here and throughout the text. This can be challenging given the lack of shared vision of EHR and research interoperability. This is also particularly challenging given the current transition of paper-based workflows to electronic workflows that we are now seeing in clinical research informatics.
3. *The development of a standard must be tied to both functional requirements and content requirements*. Historically, many data standards were being developed in isolation. They were silos and in their free-standing silo status were including dimensions that overlap with other processes. Now that data standards are being developed with the exchange and interoperability in mind, we are seeing new aspects. Lenza and colleagues suggest that the modeling of domain concepts should be separated from IT system implementation (i.e., IT systems should be implemented by IT experts and medical knowledge should be modeled and maintained by domain experts) [27]. This separation is not easy. Moving forward, it is clear that these communities must communicate but also recognize their unique expertise and skills.
4. *Existing efforts can be leveraged*. The focus on data elements provided here does not negate existing work. Existing controlled vocabularies play a vital role in this vision. They represent domain-specific knowledge and important concepts. They can be used as both value sets (e.g., LOINC can be the value set for a data

element about "tests"), or they can be used to organize data elements, such as the use of SNOMED CT to index data entry questions for patient registries [28]. Regardless, we will say that:

(a) *It is always easier to repurpose and extend than to build anew*. Often, individuals reinvent a standard and underestimate the amount of work involved – often by years or decades. Re-creation of a standard requires as much work – and trial and error – as the original process.

(b) *It takes effort to find existing relevant standards* for repurposing (above), but there are many efforts historically or actively working on many problems, and they do not always advertise or market themselves well. In this sense, and because there are so many standards of so many different types that are potentially relevant, developers must invest time to identify them all.

5. *Standards for electronic data require both technical and content experts*. All relevant stakeholders should be included, and informatics persons are central to coordinating dialog about technical features, requirements, and content. Medical information is complex and has many levels of specificity [14]. An understanding of this and a master strategy are essential in order to allow specific data to be aggregated (or collected) more generally in a useful manner.
6. *Standards are dynamic and need to be maintained*. The maintenance process needs to be well documented and thoughtfully designed to allow the standard to evolve with the field and stay relevant and useful [29]. The process also needs to recognize the potential for duplication of effort with other standards or combinatorial explosion. There is always a tension between interface terminologies and reference terminologies. These lessons should not be relearned in CRI contexts, but rather CRI should be adding to the scientific literature regarding specific data representation and standards maintenance requirements and approaches. Commercial developers incorporate standards in products and must be permitted to receive the return on the investment before changes are introduced. Further, if the currently implemented standard meets the users need, it is unlikely that user will spend more money just to be up-to-date, hence the reason for so many versions of a standard.
7. *People generally want fast and simple* and focus on an immediate problem without a bigger vision. Fast efforts have emerged to date – they can successfully focus on a single area and get a standard. But they have not considered bigger implications and have added to the fragmentation we have now. This is very understandable because to address the bigger vision simply takes time and money, and no one can expect a sponsor with a particular business need to take on greater challenges that offer less immediate value. But the lack of complete stakeholder involvement and use cases often leads to incomplete requirements, an insufficient standard, the need for more work, and scope creep.
8. *Scope creep. The human nature is the fact that humans want to expand.*On the other hand, what is the value of half-solving a problem? As SDOs get into creating a standard, more people/stakeholders become involved, technology changes, and the standard must evolve. The temptation is always to expand scope, and

the tendency is to want to address needs and requests of stakeholders – one to keep the standard dynamic and exciting but also to keep it alive. As a result, standards can get complex and unwieldy, and then a backlash happens and people reject the standard for one that is more simple. The key is to be able to change but still keep it simple.

The creation of standards, by its very nature, is a slow process. The fact is that there is a tension between speed and sophistication. Unfortunately, the new age of research (as part of this Brave New World) requires standards to support interoperability, which implies more stakeholders, more cross-disciplinary collaboration, and more time. The more information exchange functions that are required will necessitate more stakeholders. The more stakeholders involved, the more complex the standard will become. However, this also presents new opportunities for informatics solutions and new needs to pull in contributing disciplines such as information technology, communications, sociology, cognitive science, computer science, and others. Tools for standard development, visualization, and distributed collaboration are critical. These tools must be easy to access, install, and use. By default, many standards are being developed using insufficient tools (like Excel spreadsheets), and there are many needs for tools that enable distributed collaboration and standards development.

9. *Use cases and exchange scenarios can guide efforts and are a preferred approach to standards*. Since the greatest benefits of standards involve the ability to exchange and share, these uses must be considered. The scope of a standard, therefore, can go beyond what the user group envisioned or is scoped to do. So, they are multiorganizational and transdisciplinary by definition – a major change from previous standards scoping efforts that tend to get one community. Narrative use cases are often a means to elicit the specific requirements in the context of a practical and needed use. These use case stories can take many forms, but human-readable versions support the engagement of many different individuals, organizations, and cultures. These use cases must describe that data exchange specifics as well as the static content (that in reality is never quite static). Nahm et al. propose that the data elements, workflow activities, and data movement specifications collectively be called clinical domain analysis models (CDAMs) [20]. As many use cases should be developed as possible – more than a very few and enough to cover full scope of the needs and the domain. After the standard is created, other use cases should be played against the standard to make sure their requirements are met.
10. *All standards/data collection/information system efforts are tied to workflows.* As our information changes, global cooperation in research changes; the workflows are becoming formally developed and including redesign and reengineering, and then our terminology needs will change and evolve. These changes will be consensus, and both bottom-up and top-down approaches. Workflows are dynamically and radically changing in clinical research, based upon changing technology, reactive legislation, and newer motivations for multinational research collaboration. These have implications for standards, which, at the end

of the day, are only used and adopted when they fit within the job or interests of stakeholder organizations or individuals (care providers, researchers, and patients themselves). New and creative models for incentivizing the development, implementation, coordination, and evaluation of standards are badly needed and represent a wide open area for CRI experts.

The Future of Clinical Research

If a standard process and representations for data elements were to be achieved, then one could conceive of our vision of an integrated health information system. A global master set (or coordinated sets) of data elements could be available throughout the world at no cost, and every healthcare site and related business would commit to the use of these data elements. No one would use everything, but no one would use any data element not included in the master set. The commitment would be to use the data element from beginning to end. The data element would be defined at the point of creation and used without a change in any and all use. For example, the name and characteristic of the data element in a chemical laboratory machine would use the same name as the clinician or researcher using the data element.

A data set of data element codes would become the language of business transactions. For example, a medical center can post a public data set that defines the data element collected by the institution. A researcher who is setting up a clinical trial can use that database to set up a search algorithm to identify possible patients for the clinical trial. If the researcher needs additional data elements, she can negotiate with the data collection process to add that element to the collection package. Reimbursement requirements specify the data elements required by code. The exchange of data elements between sites of care, for example, between a hospital and a nursing home, can be specified by defining a set of codes. Presentation and exchange of data can be controlled by just indicating the codes along with some other parameters such as date-time. That vision of tomorrow's health environment explodes if such a system of standard and accessible data elements were in place throughout the world. (The emerging concept of *provenance* is becoming increasingly recognized as critical to understanding the integrity of health and research data, and standards issues related to representing the origin, changes, and integrity of medical and research data will be a significant informatics challenge in the near future.)

Data mining is evolving into one if not the most important application of HIT. By having a rich set of unambiguous and clearly identified data elements for a large number of patients, knowing patient data value sets, problems, treatments, and outcomes, data mining can and must support the creation of knowledge and provide the evidence for evidence-based medicine. This knowledge coupled with the clinical research community conduct of clinical trials largely from the same data sets can more easily and quickly support clinical effectiveness research. Further, since the data elements are mostly present in the patient care system, the translation from research to routine patient care is quickly accelerated. Instant learning can immediately influence and enhance patient care. The created knowledge for evidence-based medicine can be used to drive clinical guidelines and decision support algorithms.

A few final points about data standards – whether they be standards from newly collected research data or standards for healthcare data that can be leveraged as a source of research data – are worth mentioning. The collection and generation of research data sets are expensive and time-consuming, and there is a need for researchers, regulators, and the public to understand the data, reproduce the analysis to validate or contradict the conclusions, and aggregate data or compare findings across multiple studies. The value of data standards can grow (and be appreciated) over time as new research combines data sets to make new discoveries. In this regard, the scientific community has a responsibility to share data in a meaningful way. Data standards should not override scientific judgment of investigators or protocol-specific procedures and data collection. In other words, data standards should not standardize what is done experimentally but should describe what was actually done so that other systems and individuals can sensibly use that data. In addition, data standards can support information flows and actions supporting clinical research activities, enabling technology to impact protocol conduct and patient safety.

The clinical research community is actually relatively new to shared standards (CDISC is only 10 years old). The prospect of widespread electronic healthcare data collection is only just on the horizon in developed countries and a bit beyond in developing nations. There is room for standards participation on many levels. There is a need for incredibly detailed and specific content standards. Domain experts are needed as knowledge engineers who can participate in terminology development. There is a need for architects of a master system to determine how interaction between EHR and CRI should occur, what are best mechanisms for data sharing, and so forth. Finally, there is a need to coordinate conceptual models, functional requirements, and clinical content standards across domains. The development of resources and tools to support the development and access, and hence sharing, of data standards will all be required.

The data standards vision from the US National Library of Medicine has been for a long time an interlocking set of standards that allow detailed representation and general representation that can be folded into each other or used (and extended) as needed [12]. This is difficult in practice given the variety of domains and motivations; but the vision of interoperable health IT systems will undoubtedly require standards that facilitate data collection and semantic representation at all levels of granularity and for all aspects of health and medicine and research. Informatics, health data, and thus health data standards must be part of the process of bringing previous separate interests or domains together. Standards, specifically ontologies (Chap. 14), must be implemented throughout the healthcare spectrum – from biomolecular and 'omics informatics, to clinical informatics, to patient care informatics, to community and public health informatics, to population informatics. Of course, the common thread throughout all of this is the individual.

Like the Great Wall of China, the achievement of standardized data and interoperable health and research information systems will take a shared and big vision and lots of workers and coordination. The bigger the scope, the more stakeholders will be involved, and the longer the process will be. As we move forward with our vision, we have many stakeholders. But at the same time we have many beneficiaries. We are all beneficiaries. So, as we witness the slowness of standards, we hope that this chapter helps to explain it and suggest sound approaches toward achieving standards. Of course, this goal and the incremental steps to get there will be the subject of debate for many years to come and an exciting aspect of clinical research informatics.

Appendix 18.1: 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 3,000+ 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.

The 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 health care. Founded in 1928. http://www.ahima.org.

The 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 health care. 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 Single European 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/website/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 health care*. 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 health care. Its mission is to lead health care transformation through the effective use of health information technology. It was founded in 1961. http://www.himmss.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 2,300+ members include approximately 500 corporate members who represent more than 90% of the information system vendors serving health care. 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 health care professionals and industry to improve the way computer systems in health care 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 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 health care terminology in the world. Formed in 2006. http://www.ihtsdo.org/.

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.

The National Quality Forum (*NQF*) – A nonprofit organization with a mission to improve the quality of American health care 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 health care stakeholders, including consumer organizations, public and private purchasers, physicians, nurses, hospitals, accrediting and certifying bodies, supporting industries, and health care 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 health care 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 health care 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 health care. 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 health care 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.

World Health Organization (*WHO*) – WHO is the directing and coordinating authority for health within the United Nations 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.

United States Government Organizations Developing and Naming Standards

Centers for Disease Control and Preservation (*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 health care, 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 United States 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 health care insurance programs. HHS programs are directed by the Office of the Secretary and administered by 11 operating divisions, including eight agencies in the US Public Health Service and three 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 health care 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.

The National Institutes of Health (*NIH*) – A division of the US Department of Health and Human Services and the primary agency of the United States 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 United States 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 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 health care. The ONC makes recommendations on health IT 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 provides 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 provides for the 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. 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 United States 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 United States. It is currently in its 10th 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. 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 the International Health Terminology Standards Development Organisation (IHTSDO), a not-for-profit association in Denmark. The CAP continues to support SNOMED CT operations under contract to the IHTSDO and provides SNOMED-related products and services as a licensee of the terminology. http://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html.

Resources

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. Kallem C, Richesson R, DuLong D, et al. Advancing secondary data uses through data standards. J AHIMA. 2011;82(4):38–9.
2. Hammond WE, Cimino JJ. Standards in biomedical informatics. In: Shortliffe EH, Cimino JJ, editors. Biomedical informatics. Computer applications in health care and biomedicine. New York: Springer; 2006. p. 265–311.
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, Hales JW, Lobach DF, et al. Integration of a computer-based patient record system into the primary care setting. Comput Nurs. 1997;15(2 Suppl):S61–8.
5. FDA. FDA Resources for data standards. 2010 [cited 2010 Dec 6]. Available from: http://www.fda.gov/forindustry/datastandards/default.htm. Accessed 14 Dec 2011.
6. ICH. Data elements for transmission of individual case safety reports E2B(R3). International conferenceonharmonisation;2005.http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm274966.htm. Accessed 26-Jan-2012.
7. Richesson RL, Smith SB, Malloy J, et al. Achieving standardized medication data in clinical research studies: two approaches and applications for implementing RxNorm. J Med Syst. 2010;34(4):651–7.
8. Wasserman H, Wang J. An applied evaluation of SNOMED CT as a clinical vocabulary for the computerized diagnosis and problem list. AMIA Annu Symp Proc. 2003;2003:699–703.

9. Elkin PL, Brown SH, Husser CS, et al. Evaluation of the content coverage of SNOMED CT: ability of SNOMED clinical terms to represent clinical problem lists. Mayo Clin Proc. 2006; 81(6):741–8.
10. Cimino JJ. Desiderata for controlled medical vocabularies in the twenty-first century. Methods Inf Med. 1998;37(4–5):394–403.
11. Cleverdon CW, Keen EM. Factors determining the performance of indexing systems. Vol.1: Design, Vol. 2: Test results, Aslib Cranfield research project. Cranfield, 1966; 1967.
12. Chute GC, Cohn SP, Campbell JR. A framework for comprehensive health terminology systems in the united states. J Am Med Inform Assoc. 1998;5(6):503–10.
13. Chute CG, Elkin PL, Sheretz DD, et al. Desiderata for a clinical terminology server. 2004. Proc AMIA Symp. 1999:42–46.
14. Blois MS. Information and medicine. Berkeley: The University of California Press; 1984.
15. Richesson RL, Nadkarni P. Data standards for clinical research data collection forms: current status and challenges. J Am Med Inform Assoc. 2011;18(3):341–6.
16. CDISC. Clinical data acquisition standards harmonization: basic data collection fields for case report forms. Draft version 1.0. 2008 Oct [cited 2010 Sept 1]. Available from: http://www.cdisc.org/cdash. Accessed 14 Dec 2011.
17. FDA. Guidance for industry. Qualification process for drug development tools. In: Draft guidance. Rockville: FDA; 2010.
18. ISO/IEC. 11179–3:2003 information technology metedata registries (MDR) part 3 registry metamodel and basic attributes. 2003. Accessed 14 Dec 2011.
19. Lee YW, Pipino LL, Funk JD, et al. Journey to data quality. Cambridge: MIT Press; 2006.
20. Nahm M, Walden A, McCourt B, et al. Standardising clinical data elements. IJFIPM. 2011;3(4):2010.
21. Warzel DB, Andonaydis C, McCurry B, et al. Common data element (CDE) management and deployment in clinical trials. AMIA Annu Symp Proc. 2003:1048. Accessed 14 December 2011.
22. Kush R, Alschuler L, Ruggeri R, et al. Implementing single source: the STARBRITE proof-of-concept study. J Am Med Inform Assoc. 2007;14(5):662–73.
23. Bhatt DL. Advancing the care of cardiac patients using registry data: going where randomized clinical trials dare not. JAMA. 2010;303(21):2188–9.
24. ISO/IEC. 11179–1:2004 information technology – metadata registries (MDR) – part 1: framework. 2004 http://metadata-stds.org/11179/. Accessed 14 Dec 2011.
25. Object Management Group (ed.) Unified modeling language superstructure and infrastructure. Vol. 21, Issue: 2; p. 226. Available at: http://www.mendeley.com/research/omg-unified-modeling-language-omg-uml-infrastructure/ DOI: Accessed 14 Dec 2011.
26. Campbell KE, Oliver DE, Shortliffe EH. The unified medical language system: toward a collaborative approach for solving terminologic problems. J Am Med Inform Assoc. 1998;5(1):12–6.
27. Lenza R, Beyera M, Kuhnb KA. Semantic integration in healthcare networks. In: Connecting medical informatics and bio-informatics: proceedings of MIE2005 – the XIXth international congress of the European federation for medical informatics. Geneva: IOS Press; 2005.
28. Richesson R, Shereff D, Andrews J. [RD] PRISM library: patient registry item specifications and metadata for rare diseases. J Libr Metadata. 2010;10(2–3):119–35.
29. Oliver DE, Shahar Y, Shortliffe EH, et al. Representation of change in controlled medical terminologies. Artif Intell Med. 1999;15(1):53–76.

Chapter 19
Pharmacovigilance

A.C. (Kees) van Grootheest and Rachel L. Richesson

Abstract Pharmacovigilance is a scientific discipline concerned with the safety of drugs as used in clinical practice and whose main purpose is to balance the risk–benefit ratio to the public. This chapter discusses the progression of pharmacovigilance as a discipline from its start in the early 1960s during the thalidomide tragedy to its current status as a visible and pervasive part of the health care delivery and research. It provides an overview of the recent activities and science supporting pharmacovigilance and their informatics dependencies and implications, as well as the settings in which pharmacovigilance activities are undertaken. Major informatics themes related to pharmacovigilance include the design and support of drug safety data collection systems, identification of alternative data sources, methodological development to support new analyses and discovery, methods to support the use of patient-specific genetic profile data as mitigating factors, and communication of complex information to physicians, policy makers, and patients. This chapter also reviews the various methods of detecting new adverse drug reactions, including pre- and postmarketing studies, spontaneous reporting, intensive monitoring, and database studies, as well as the pros and cons of each. Both pharmacovigilance and drug safety monitoring impact the activities and workflows of clinical research, and the knowledge generated from these activities will lead to safer use of drugs in the future.

A.C. (Kees) van Grootheest, M.D., Ph.D. (✉)
Mathematics and Natural Sciences, Netherlands Pharmacovigilance Centre,
Goudsbloemvallei 7, 5237 MH, s-Hertogenbosch Lareb, The Netherlands

Department of Pharmacy, University of Groningen,
s-Hertogenbosch, The Netherlands
e-mail: ac.vangrootheest@lareb.nl

R.L. Richesson, Ph.D., MPH
Division of Informatics, Duke University School of Nursing,
311 Trent Drive, Durham, NC 27710, USA
e-mail: rachel.richesson@dm.duke.edu

R.L. Richesson, J.E. Andrews (eds.), *Clinical Research Informatics*, Health Informatics,
DOI 10.1007/978-1-84882-448-5_19, © Springer-Verlag London Limited 2012

Keywords Pharmacovigilance • Drug safety monitoring • Database of adverse drug events • Postmarketing surveillance • Pharmacoepidemiology • Quantitative signal detection • Risk management plans • Informatics

Pharmacovigilance is defined by the World Health Organization (WHO) as "the science and activities relating to the detection, assessment, understanding and prevention of adverse effects or any other drug-related problems" [1]. The activities of pharmacovigilance are information intensive and include the collection, exchange, aggregation, analysis, interpretation, and communication of data related to patient experiences and use of therapeutic agents. Pharmacovigilance plays a key role in ensuring that patients receive safe drugs, and its activities are becoming increasingly visible with growing public scrutiny of the drug industry. New regulatory, political, and scientific requirements are changing the practice of pharmacovigilance and are creating many demands and opportunities for the application of information science and technology. On a regulatory level, the proviso of conditional approval for marketed drug products, and related requirements for risk management plans (including phase IV clinical studies and patient registries), reinforce the need for efficient event detection and reporting systems. Politically, increased patient involvement and the push for "transparency" in the business of research and medicine is affecting the nature of pharmacovigilance as well – strengthening demands for systems to monitor and act upon possible threats to patient safety as soon as possible.

Pharmacovigilance and drug safety monitoring activities have a definite impact on clinical research practice and related information management activities. Many pharmacovigilance and population monitoring activities are clinical research activities by definition and can also impact research and development. The results of pharmacovigilance and drug safety monitoring impact decisions on which agents are developed or manufactured and for which populations and indications and can impact the practice, and therapeutic interventions, for future human trials.

As our scientific understanding and cumulative knowledge about existing products continually increases, so do opportunities for medical discovery, risk communication, and prevention of adverse side effects. Pharmacovigilance systems and methods have undoubtedly evolved with the ever-increasing number of electronic data sources and will continue to do so as information technology permeates health care and research domains. The aim of this chapter is to provide an overview of the activities and science that support pharmacovigilance and their informatics implications. New paradigms and recent developments that affect the practice, information requirements, and impact of pharmacovigilance will also be discussed. The organization and themes of this chapter are inspired by a recent review of pharmacovigilance activities and outcomes [2].

Historical Perspectives

Without a systematic detection and assessment process for adverse drug effects, thousands of individuals might experience an event before an astute clinician might bring his suspicions forward for public investigation and action. In 1961, the Australian physician McBride had a letter published in The Lancet in which he suggested a connection between congenital malformations in newly born infants and the hypnotic thalidomide, which had been marketed under various names in many countries, both as a prescription and as an over-the-counter (OTC) drug, as well as in many compositions of simple analgesics [3]. Later that same year, the manufacturer Chemie Grünenthal withdrew thalidomide from the market but only after an accrual of very perceptible and debilitating congenital defects. The number of children born with serious congenital malformations as a result of maternal use of thalidomide is estimated between 6,000 and 12,000, the majority of which were born in Germany [4, 5]. Had the biological sequelae of fetal thalidomide exposure been less visible or severe, perhaps tens or hundreds of thousands more would have been affected. Since then, the use of deliberate data collection and analysis activities in the interest of public health has detected and removed several dangerous products from public markets.

An adverse event (AE) is broadly defined as any clinical event, sign, or symptom that goes in an unwanted direction [6]. There is no assertion of causality implied with adverse events – they are merely events. The notion of an adverse drug reaction (ADR) includes the suggestion of a causal relationship (e.g., probable, possible, etc.) between the event and a therapeutic agent or device. After an ADR is suspected (i.e., adverse consequences are speculated to be caused from a drug), then careful and systematic data collection is required to evaluate that suspicion for further action. The decision to remove thalidomide from the market was made in response to *active* data collection of adverse events related to the drug. The thalidomide experience soon led the FDA (the Food and Drug Administration, USA) to initiate a systematic collection of reports on all types of adverse drug reactions, chiefly through the Hospital Reporting Program. In various countries, the thalidomide tragedy had prompted the immediate formulation of criteria for safety and efficacy that new drugs would need to meet in order to receive marketing authorization. In addition, the marketing authorization holders were commissioned to establish a postmarketing surveillance system to facilitate the early detection of adverse reactions to prevent a similar tragedy from occurring in the future. In 1968, ten countries that supported a spontaneous reporting system for adverse drug reactions formed a collaboration under WHO Pilot Research Project for International Drug Monitoring [7]. In 1971, a resolution of the Twentieth World Health Assembly laid the foundations for WHO International Drug Monitoring Programme [8]. In 1972, a report was published which formed the basis of the current international system of national centers collaborating in WHO Programme [9, 10]. Although sophisticated methods have evolved, the motivation (public safety) and strategy (population monitoring) have not changed.

The overarching purpose of pharmacovigilance has been to balance the risk–benefit ratio to the public. Pharmacovigilance is based upon the premise that there are risks associated with any therapeutic agent, and these risks are not evenly distributed across a population. Biological variations make some individuals more vulnerable to side effects or adverse drug reactions. Individual variations also impact the course of a disease and preferences for treatment options and tolerability of side effect.

Global Perspectives

Various national authorities and the pharmaceutical industry have played a role in the development of pharmacovigilance practice, which in turn has resulted in new legislation and qualitative requirements for the drug, the industry, and their products. The Council for International Organizations of Medical Sciences (CIOMS) and the International Conference on Harmonisation (ICH) have been instrumental in developing pharmacovigilance standards and practice and continue to operate as forums for discussion and standardization of drug safety methods and requirements. Through such venues, it has become clear that a globally integrated and deliberate pharmacovigilance system is required to protect the public. The WHO International Drug Monitoring Programme is supported and coordinated by the WHO Collaborating Centre for International Drug Monitoring ("the Uppsala Monitoring Centre"), which maintains and implements the international database of adverse drug events.

It is necessary that countries maintain some autonomy in terms of regulatory organizations and legal authority because they each make their own decisions about drugs marketed in their countries and their populations have different risks and benefits for various products. Despite the need for national autonomy, however, standardized drug safety reporting and global communication is in the best interest of all countries because it can enable countries to share data and learn from drug exposure experiences from other populations. An awareness of products available in other countries is critical to fully protect the safety of patients. Not only is it helpful for countries to be familiar with products, product information, and regulatory actions of other countries, sharing of data between countries could increase the monitored population size and potentially identify a signal (i.e., a potential relationship between product and untoward medical consequence) that might not be detected otherwise. Additionally, the multinational character of pharmaceutical companies has influenced the need for shared data and standards in the area of drug safety. The idea of a single drug company having to report information in multiple formats for multiple counties could prove to be a disincentive for much needed multinational trials. Pharmacovigilance use cases (related to standardized reporting formats for AE information), in fact, are one of the highest impact use cases that drive international standardization efforts such as CDISC and efforts toward standardization and reuse of data collected in electronic health records (see Chap. 17).

Criticisms of Current Pharmacovigilance Systems

In the aftermath of the withdrawal of rofecoxib in 2004, the FDA and the current system of postmarketing surveillance was heavily criticized [11–15]. Points of criticism were that the FDA is using only a limited number of data sources (clinical trials, spontaneous reporting) for information about the safety of a drug. Furthermore, the FDA has no legal control over the conduct and completion of postmarketing safety studies. The majority of postmarketing study commitments are never initiated, and the completion of postmarketing safety studies (i.e., phase IV studies) declined from 62% between 1970 and 1984 to 24% between 1998 and 2003. The FDA has no authority to take direct legal action against companies that do not fulfill their postmarketing commitments [16]. Some critics also claim that the FDA has become too close to the industry that they are supposed to regulate; consequently, a separation between regulatory duties and the postmarketing surveillance activities has been advocated [17]. In response to the criticism, the Centre for Drug Administration (CDER) at the FDA asked the Institute of Medicine (IOM) to assess the US drug safety system. In September 2006, the IOM released the committee's findings and recommendations in a report "The future of drug safety: promoting and protecting the health of the public" [18]. The main message in this report is that the FDA needs to follow the safety of a drug during its whole life cycle. This life cycle approach includes identifying safety signals, designing studies to confirm them, evaluating benefits as well as risks, using risk–benefit assessments to integrate study results and communicating key findings to patients and physicians [7, 19].

Similarly, in Europe, the withdrawal of rofecoxib lead to an assessment of the pharmacovigilance system in the different European Union member states, which was published in March 2006. The report "Assessment of the European Community System of Pharmacovigilance" highlighted the strengths and weaknesses of the European pharmacovigilance system. The report's recommendations focused on the breadth and variety of data sources, the proactive use of registration, the speed of decision making, the impact of regulatory action and communication, the compliance by marketing authorization holders, and the general principles of quality management and continuous quality improvement [20, 21].

Different Aspects of Pharmacovigilance

The activities undertaken in the name of pharmacovigilance fall within three settings: regulatory, industry, and academia. Each of these settings reflects a different perspective on the scope and practice of pharmacovigilance. Consequently, these three perspectives have influenced the nature of pharmacovigilance and will continue to be a source of influence in the future.

Regulatory Pharmacovigilance

Regulatory authorities, both at the national and increasingly at the international level, initially defined and fostered the field of pharmacovigilance. Because of the crucial role that regulatory authorities have played in the development of pharmacovigilance, Waller et al. actually used the label "regulatory pharmacovigilance" which they define as "the process of evaluating and improving the safety of marketed medicines" [22]. They demarcated the responsibilities that national governments have in the monitoring of drug safety, which was well accepted by many nations, sobered by the outcome of the thalidomide tragedy [7, 10]. It is undeniable that in several countries, most notably in the USA and UK, legislation has significantly contributed to the advance of pharmacovigilance as a specialized field of knowledge. The role of WHO stands out here. The collaborative program launched under the auspices of WHO by ten countries in 1968 was the start of a historic international cooperative effort, resulting in WHO International Drug Monitoring Programme [7]. The technical report, entitled "International Drug Monitoring: The Role of National Centres" and published as the proceedings of one of WHO meetings in 1972, laid the theoretical and practical foundation for the further development of pharmacovigilance [9]. The program has also resulted in the WHO Collaborating Centre for Drug Monitoring (the Uppsala Monitoring Centre) which maintains the international ADR database and fulfills an important role particularly by the support it offers to the pharmacovigilance centers in low-income countries.

The Role of the Pharmaceutical Industry

The second great influence on the development of pharmacovigilance is the pharmaceutical industry. This is not surprising since it is their product, a product they themselves have both developed and manufactured, that is the object of study. From their circles, great influence has been exerted to come to international agreements, many of which have since been formalized in the various reports the Council for International Organizations of Medical Sciences (CIOMS) and the International Conference on Harmonisation (ICH) have issued. Initially, the sector's main interest lay in the epidemiological approach and causality assessment, but nowadays, aspects of risk management are also given due attention.

Pharmacovigilance as a Science

Pharmacovigilance is also a scientific discipline, dedicated to the safety of drugs as used in the clinical practice, based on experiences from the clinical practice, thus generating knowledge on the harmful effects of drugs, both at the individual and the population level. The knowledge generated by pharmacovigilance activities will

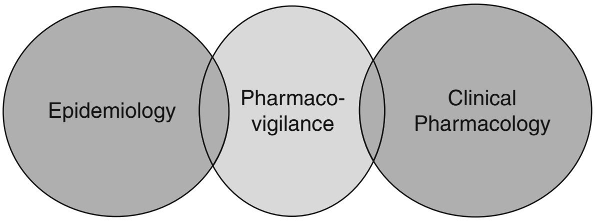


Fig. 19.1 Pharmacovigilance and related disciplines

eventually be applied to clinical practice and thus lead to a safer use of drugs. Many of the statistical methodologies of pharmacovigilance are conceptualized, applied, and evaluated from university and academic settings [23, 24].

As in most applied sciences, the field of pharmacovigilance is an amalgam of numerous other scientific domains, each contributing their own expertise to the field. The combined knowledge fosters drug safety reasoning. Pharmacovigilance is essentially a clinical science [25]. To allow a sound judgment of any adverse effects of drugs, we need clinical knowledge at the level of the individual patient. It takes extensive general medical knowledge, preferably supported by direct experience with patient care, to be able to make an accurate assessment of the impact pharmacotherapy is likely to have, which becomes even more urgent when unintended adverse events occur. However, as the availability of robust electronic clinical data sets increases, pharmacovigilance activity is becoming even more interdisciplinary. Academics are now publishing pharmacovigilance methods in statistical journals, clinical specialty journals, informatics journals, database and computer science journals, and health policy journals. We expect this trend to continue. Pharmacoepidemiology, the science concerned with the effects of drugs in large populations, has been a key contributor and, among other contributions, has helped establish the basis for the statistical analysis techniques and risk assessments in pharmacovigilance (Fig. 19.1).

Lastly, it has been the international scientific societies that have been vital in furthering pharmacovigilance as a discipline in its own right. The International Society of Pharmacoepidemiology (ISPE), founded in 1984, has helped formulate the epidemiological underpinning of the safety aspects of drugs. The International Society of Pharmacovigilance (ISoP), founded as the European Society of Pharmacovigilance in 1992, has promoted the field's clinical and communication aspects. Both organizations organize international courses in their field, in addition to their annual conferences.

Methods in Pharmacovigilance

Regardless of who performs a pharmacovigilance activity (e.g., regulatory, industry, or academia), similar methods are used in the detection of new adverse drug reactions (Table 19.1).

Table 19.1 Activities of pharmacovigilance

Suspected ADR signal generation and formation of hypothesis
Analysis of all issues around the signal, particularly confirmation (of refutation) of hypotheses, estimation of the size of the risk, and whether susceptible patients exist
Consideration of possible changed benefit-to-risk issues in therapy
Communication of information to health professionals and patients in a useful way and possible regulatory action
Consequence evaluation

From Edwards [26], with permission of Wolters Kluwer

We describe the basic methodological strategies, and the pros and cons of each, below.

Premarketing Studies

The main method of gathering information about a drug in the premarketing phase is to conduct a clinical trial. As mentioned in Chap. 4, premarketing clinical trials can be divided into three phases. Phase III studies are often double-blind randomized controlled trials, which are considered to be the most rigorous way of determining whether a cause–effect relationship exists between a treatment and an outcome. However, when it comes to monitoring the safety of a drug, this study design is not optimal. Due to the limited number of patients participating, it is not possible to identify rare ADRs. The relatively short duration of clinical trials makes it difficult to detect ADRs with a long latency. Another limitation with clinical trials is the population in which a drug is tested. The characteristics of these persons do not always correspond to the characteristics of the population in which it will later be used; therefore, it might be difficult to extrapolate the results obtained from clinical trials to the population at large [27]. This is especially true for the elderly, for women, or for people belonging to a minority [28, 29]. In order to study rare ADRs, ADRs with a long latency, and ADRs in specific populations, careful monitoring of the drug in the postmarketing phase is essential.

Postmarketing Studies

Postmarketing studies can be descriptive or analytical. Descriptive studies are hypothesis generating and try to describe the occurrence of events related to drug toxicity and efficacy. Analytical studies are hypothesis testing and seek to determine associations or causal connections between observed effects and particular drugs and to measure the size of these effects. Descriptive studies are widely used in postmarketing surveillance because they are able to generate hypotheses that will become starting points for analytical studies [30]. Two forms of descriptive studies, spontaneous reporting and intensive monitoring, will be discussed in two subsections below. Analytical studies can be conducted via a variety of methods including case–control studies, cohort studies, and clinical trials. In order to be able to conduct retrospective cohort and case–control studies, data which have been collected in a reliable and routine fashion need to be available.

Spontaneous Reporting

Spontaneous reporting systems (SRS) have been and still are the chief method of collecting postmarketing information about the safety of drugs. The primary function of SRS is to facilitate the early detection of "signals" of new, rare, and serious ADRs. Via a spontaneous reporting system, physicians, and increasingly pharmacists and patients, are able to report suspected adverse drug reactions to a pharmacovigilance center [31–33]. The task of the pharmacovigilance center is to collect and analyze the reports and to inform stakeholders of the potential risk when "signals" of new ADRs arise. Spontaneous reporting is also used by the pharmaceutical industry to collect information about their drugs. Via a spontaneous reporting system, it is possible to monitor all drugs on the market throughout their whole life cycle at a relatively low cost. The main criticism against spontaneous reporting is selective and underreporting [34]. In a review article, the magnitude of underreporting was investigated, and it was shown that more than 94% of all ADRs remain unreported [35]. Underreporting can lead to the false conclusion that a real risk is absent, while selected reporting of suspected risks may give a false impression of a risk that does not exist. However, underreporting and selective reporting can also been seen as an advantage. Because only the most severe and unexpected cases are reported, it is easier to detect new signals because the person reporting the issue have already pinpointed what might be a new safety issue. Against this background, perhaps the system should be called *concerned reporting* instead of spontaneous reporting, seeing as those reporting the issues are highly selective of what they are reporting [36].

Although critics say that spontaneous reporting is not the ideal method for monitoring the safety of drugs, it has proven its value throughout the years. Between 1999 and 2001, 11 products were withdrawn from the UK and US markets. Randomized trial evidence was cited for two products (18%) and comparative observational studies for two products (18%). Evidence from spontaneous reports supported the withdrawal of eight products (73%), with four products (36%) apparently withdrawn on the basis of spontaneous reports only. For two products, the evidence used to support their withdrawal could not be found in any of the identified documents [37]. Out of nine recent significant drug safety issues handled in the European Union since 1995, six were detected by spontaneous reports. See Table 19.1, showing the strength of spontaneous reporting in detecting new safety issues [21].

Intensive Monitoring

In the late 1970s and early 1980s, a new form of active surveillance was developed in New Zealand (Intensive Medicines Monitoring Programme) and the UK (Prescription Event Monitoring). These intensive monitoring systems are using prescription data to identify users of a certain drug. The prescriber of the drug is asked about any adverse event occurring during the use of the drug being monitored. These data are collected and analyzed for new signals. The methodology of these intensive monitoring systems has been described in depth elsewhere [38–41].

Intensive monitoring is a noninterventional observational cohort, differentiating itself from spontaneous reporting because it only monitors selected drugs during a certain period of time. Through its noninterventional character, intensive monitoring provides real-world clinical data involving neither inclusion nor exclusion criteria throughout the collection period. It is unaffected by the kind of selection and exclusion criteria that characterize clinical trials, thereby eliminating selection bias. Another strength of the methodology is that it is based upon event monitoring and is therefore capable of identifying signals for events that were not necessarily suspected as being ADRs of the drug studied. Intensive monitoring also allows estimation of the incidence of adverse events which makes it possible to quantify the risk of certain ADRs. Intensive monitoring also has recognized limitations. The proportion of adverse effects that go unreported to doctors is unknown.

The studies also produce reported event rates rather than true incident rates. This is the same for all studies based on medical record data including computer databases and record linkage. There is no control group in standard intensive monitoring studies, and the true background incidence for events is therefore not known [42].

Although the intensive monitoring methodology was developed more than 20 years ago, this methodology has received renewed interest in the last years. In the European Commission consultation "Strategy to better protect public health by strengthening and rationalising EU Pharmacovigilance," intensive monitoring is mentioned as one tool in improving the pharmacovigilance system [43].

Database Studies

In order to test a hypothesis, a study has to be performed. This kind of study can be conducted using a variety of methods including case–control studies and cohort studies. Limitations of these methods include power considerations and study design. In order to be able to conduct retrospective cohort and case–control studies, data which have been collected in a reliable and routine fashion need to be available. As an example, the General Practice Research Database is described. Other database- and record-linkage systems are available for research purposes both in Europe and in North America [44].

In the UK, virtually all patient care is coordinated by the general practitioner, and data from this source give an almost complete picture of a patient, his illnesses, and treatment. In any given year, GPs, who are members of the General Practice Research Database, collect data from about three million patients (about 5% of the UK population). These patients are broadly representative of the general UK population in terms of age, sex, and geographic distribution. The data collected include demographics (age and sex), medical diagnoses that are part of routine care or resulting from hospitalizations, consultations or emergency care, along with the date and location of the event. There is also an option of adding free text, referral to hospitals and specialists, all prescriptions including date of prescription, formulation strength, quantity and dosing instructions, indication for treatment for all new prescriptions, and events leading to withdrawal of a drug or a treatment. Data on vaccinations and miscellaneous information such as smoking, height, weight, immunizations, pregnancy, birth, death, date entering the practice, date leaving the practice, and laboratory results are also collected. Similar databases of prescriptions and drug-specific registries can afford other opportunities. As electronic health data collection becomes standardized and adopted in the United States and other countries, then similar electronic resources will exist.

Quantitative Signal Detection

Signal detection can be defined as the search for information on a possible causal relationship between an adverse event and a drug, of which the relationship is unknown or incompletely documented previously. The data analysis/exploration methods that comprise signal detection activities can be applied on various data sources, including case reports or other data streams. This is an active research area and a ripe area for clinical research informatics professionals. Signal detection has, in the past, mainly been done on the basis of case by case analyses of reports, but in recent years, data mining techniques have become more important. The term *data mining* refers to the principle of analyzing data from different perspectives and sorting out relevant information. Often, algorithms are used to discover hidden patterns of associations or unexpected occurrences (i.e., “signals”) in large databases. Although the methodology of the various data mining methods applied in pharmacovigilance differ, they all share the characteristic that they express to what extent the number of observed cases differs from the number of expected cases [23].

Several analytic approaches are currently in use. Proportional reporting ratios, PRRs, compare the proportion of reports for a specific ADR reported for a drug with the proportion for that ADR in all other drugs. The calculation is analogous to that of relative risk. Using the same information, it is also possible to calculate a reporting odds ratio [45]. The method, Bayesian confidence propagation neural network (BCPNN) is used to highlight dependencies in a data set. The method uses Bayesian statistics implemented in a neural network architecture to analyze all reported adverse drug reaction combinations. Quantitatively unexpectedly strong relationships in the data are highlighted relative to general reporting of suspected adverse effects. WHO Collaborating Centre for International Drug Monitoring uses this method for data mining [46]. A related approach is used by the FDA, which uses the Multi-Item Gamma Poisson Shrinker (MGPS) for data mining of their spontaneous report's database. The MGPS algorithm computes signal scores for pairs and for higher-order (e.g., triplet, quadruplet) combinations of drugs and events that are significantly more frequent than their pairwise associations would predict [47]. All data mining approaches currently cannot distinguish between already-known associations and new associations. Moreover, clinical information described in the case reports is not taken into account. There is still a need for a reviewer to analyze these events.

In contrast to hypothesis testing where quantitative estimates are used to express the frequency of a signal, in spontaneous reporting systems, quantitative estimates are used to determine the probability of a combination being a signal or not, based on disproportionate reporting [38]. Various quantitative procedures can be used to focus attention of human reviewers, who ultimately are required to review and evaluate any potential signal. A recent comparison of common pharmacovigilance measures found that the various measures that are used in quantitative signal detection in various national centers are comparable when four or more reports constitute the drug–ADR combination [48].

System Developments in Pharmacovigilance

Transparency

The Erice declaration, as well as *Waller and Evans*, stated that transparency is important for the future of pharmacovigilance. In the last few years, transparency around adverse drug reactions has increased. Clinical trials registration will allow the necessary tracking of trials to ensure full and unbiased reporting for public benefit [49]. A number of countries, including Canada (www.hc-sc.gc.ca), the Netherlands (www.lareb.nl), and the UK (www.mhra.gov.uk), have put their databases containing the data from the spontaneous reporting system online, available to the public.

Conditional Approval

The FDA report, as well as the report from the European Union described earlier, emphasizes that compliance by marketing authorization holders, when it comes to additional postmarketing studies, needs to be improved. A possible solution to this problem would be a time-limited conditional approval, which would place pressure on the manufacturers to conduct and report additional safety studies [50].

Within the European Union, the European Medicines Agency EMA (formerly EMEA) has introduced a conditional marketing authorization. The CHMP delivers a conditional marketing authorization for products where there is a specific patient need. Examples include products for seriously debilitating or life-threatening diseases, medicinal products to be used in emergency situations in response to public threats, and products designated as orphan medicinal products. A conditional marketing authorization is granted in the absence of comprehensive clinical data referring to the safety and efficacy of the medicinal product. However, a number of criteria have to be met including a positive risk–benefit balance of the product, a likelihood that the applicant will eventually be able to provide comprehensive clinical data, and an unmet medical need. Essentially, it must be demonstrated that the public health benefits of the immediate availability of the medicinal product outweigh the risks inherent in the absence of additional data.

Conditional marketing authorizations are valid for 1 year, on a renewable basis. The holder is required to complete ongoing studies or to conduct new studies with a view to confirming that the risk–benefit balance is positive. In addition, specific obligations may be imposed in relation to the collection of pharmacovigilance data. The authorization is not intended to remain conditional indefinitely. Rather, once the missing data are provided, it should be possible to replace it with a formal marketing authorization. The granting of a conditional marketing authorization will allow medicines to reach patients with unmet medical needs earlier than might otherwise be the case and will ensure that additional data on a product are generated, submitted, assessed, and acted upon. Methods for the collection, communication, and interpretation of these data will require informatics and integration of health care, clinical research, and regulatory information.

Risk Management Plans

Another step in a more proactive postmarketing surveillance is the introduction of risk management plans, RMPs [51]. RMPs are being set up in order to identify, characterize, prevent, or minimize risk relating to medicinal products, including the assessment of the effectiveness of those interventions. An RMP may need to be submitted at any time in a product's life cycle, for example, during both the preauthorization and postauthorization phases. An RMP is required for all new active substances, significant changes in established products (e.g., new form/route of administration), established products introduced to new populations, significant new indications, or when an unexpected hazard is identified.

The EU Risk Management Plan contains two parts, the first part containing a Safety Specification and a Pharmacovigilance Plan and the second part containing an evaluation of the need for risk minimization activities and, if necessary, a risk minimization plan. The safety specification contains a summary of what is known and what is not known about the safety of the product. This encompasses the important identified risk and any information and outstanding safety questions which warrant further investigation in order to refine understanding of benefit–risk during the postauthorization period.

A risk minimization plan is only required in circumstances where standard information provision, via a medicine's summary of product characteristics, is considered inadequate. Insufficient patient information leaflets or inadequate labeling of the medicine are additional reasons for drawing up a risk minimalization plan. Where a risk minimization plan is considered necessary, both routine and additional activities are to be included. Some safety concerns may have more than one risk minimization activity, each of which should be evaluated for effectiveness.

Many risk management plans have already been established; however, no quantitative or qualitative reports have been released by the EMA. Information to the public about RMPs has also been scarce. If RMPs are to take an important place in pharmacovigilance they need to be made public and easily accessible for scientists, professionals, and patients.

Patients as Reporters

Another important development is the recognition of the patient as an important player in pharmacovigilance. Patients are the users of drugs. Their use of a drug in a safe manner is the ultimate goal of pharmacovigilance activities. In an increasing number of countries, patients are now allowed to report adverse drug reactions to the spontaneous reporting system. The European Commission acknowledges the role of the patient in spontaneous reporting [43]. Patients and patient organizations are getting more involved in pharmacovigilance, especially when it comes to risk communication [52, 53].

After introducing patient reporting in the spontaneous reporting scheme in 2004 [54], the Netherlands Pharmacovigilance Centre Lareb took patient reporting one step further, and in 2006, an intensive monitoring program using patients as a source of information was introduced. Lareb Intensive Monitoring, (LIM), follows the PEM methodology in the way of identifying patients via prescriptions. Eligible patients are identified in their pharmacies when they come and pick up the drug under study for the first time. Patients can register at the LIM website, and during a certain period of time, they will receive questionnaires asking them about adverse events. The system is totally web-based, meaning that questionnaires can be sent via e-mail to participating patients at different points, allowing the collection of longitudinal data. The high grade of automation also allows quick data collection and analysis [55]. Further, the notion of patients entering their data directly in personal health records [56] and exercising control over their data entered in provider electronic health record (EHR) systems will likely increase the involvement of patients as reporters or gatekeepers to ADR information and drug experience.

International Developments

The United States

In 2007, in response to the IOM report, the FDA announced several initiatives designed to improve the safety of prescription drugs [19]. These initiatives fall into four main categories. The first is increasing the resources for drug safety activities. Perceptions of the agency as being overly dependent on industry funding have lead to proposals of eliminating user fees. The second category of proposed reform is new authority for the FDA; the agency needs regulatory tools to help assure drug safety. This authority would be exercised through a required risk evaluation and mitigation strategy including measures such as prescribing restrictions, limits on direct to consumer marketing, and requirements for postmarketing studies. The FDA could impose monetary penalties for noncompliance. A third aspect of the reform is improvement of postmarketing surveillance. A routine systematic approach to active population-based drug surveillance that could identify potential safety problems is needed. Finally, changes in the FDA management practices and safety supervision are necessary [57]. The latter two changes clearly will require informatics expertise in developing systems for data collection and communication that enhance opportunities for collaboration and integrated response.

In May 2007, the US Senate passed its version of reform for the Food and Drug Administration. The senate proposed that the Prescription Drug User Fee Act, which allows the pharmaceutical industry to pay money directly to the FDA, should increase their payments to the FDA with close to 400 million US dollars. Furthermore, this reform would give the FDA new authority to order companies to undertake formal safety studies of drugs that are being marketed and to fine those who do not honor their postmarketing commitments. However, when it came to changing the structure of the FDA, the proposal to create an independent office for the monitoring of the safety of drugs was rejected by one vote [58–60].

Europe

In 2005, a document was drafted by the Heads of Medicines Agencies called "*Implementation of the Action Plan to Further Progress the European Risk Management Strategy*." In July 2007, the EMEA published a document wherein the achievements thus far were discussed. Achievements included implementation of legal tools for monitoring the safety of medicines and for regulatory actions. Particular emphasis was placed on [2]:

1. Systematic implementation of risk management plans
2. Strengthening the spontaneous reporting scheme through improvements of the EudraVigilance database

3. Launching the European Network of Centres for Pharmacoepidemiology and Pharmacovigilance (ENCePP) project to strengthen the monitoring of medicinal products
4. The conduct of multicenter postauthorization safety studies
5. Strengthening the organization and the operation of the EU pharmacovigilance system.

In the course of the next 2 years, two main areas will be covered by the European Risk Management Strategy: further improving of the operation of the EU Pharmacovigilance System and strengthening the science that underpins the safety monitoring for medicines for human use [61, 62]. In December 2007, a public consultation "Strategy to Better Protect Public Health by Strengthening and Rationalising EU Pharmacovigilance" was published on behalf of the European Commission. This document contains legislative strategy and key proposals for legislative changes within the European Union. Areas where legislative changes are necessary include: fast and robust decision making on safety issues; clarification of roles and responsibilities for industry and regulators; strengthening the role of risk management planning; improving the quality of noninterventional safety studies; simplification of ADR reporting including introducing patient reporting; strengthening medicines safety, transparency, and communication; and, including clearer safety warnings in the product information to improve the safe use of medicines [43].

Role of Informatics

Pharmacovigilance is by definition an information-intensive discipline. As electronic data sources become available, pharmacovigilance will become even more interdisciplinary, requiring information and computational science professionals as well as statisticians and clinical experts. The essential activities of pharmacovigilance defined at the start of this chapter will continue to be the foundation of pharmacovigilance, although new data sources, analytic methods, and communication platforms will be added.

Informatics will play an important role in pharmacovigilance. Key informatics areas include the design and support of safety data collection systems, identifying alternative data sources for pharmacovigilance, methodological development to support new analyses and discovery (signals detection, probabilistic methods, etc.), methods to support the use of patient-specific genetic profile data as mitigating factors, and communication of information to physicians, policy makers, and patients. As discussed in Chap. 8, thoughtful design and implementation strategy serve as the foundation of any of these application areas. The need for global communication and collaboration will drive all of these systems to support multiple languages and regulations. Increasing volumes of electronic health data and the changing role of the patient in health care, clinical research, and pharmacovigilance mean that the design and implementation of data collection, information, and communication systems must consider many different user groups.

In order to receive information about the safety of a drug at the earliest point, active surveillance is necessary. Real-time data collection and information gathering is vital for active postmarketing surveillance. Spontaneous reporting has indeed been shown to be a useful tool in generating signals, but the relatively low number of reports received for a specific association makes it less useful in identifying patient characteristics and risk factors that will contribute to the occurrence of an ADR for a given person, which is the essential information for providers making treatment decisions. Furthermore, when facing an ADR, both patients and providers want to know: Will this ADR disappear? How long will it take before it does? What treatment is needed? Current pharmacovigilance resources and methods cannot address these questions. The development of infrastructure and process that can follow cohorts of patients using various drugs can support this type of population monitoring. These systems will likely develop in parallel with electronic health record (EHR) and personal health record (PHR) systems that enable continuity of health care. To support longitudinal follow-up of patients and their adverse drug events, future data streams and sources should include not only information linked to the individual (identifiers, demographics, etc.) but also information regarding treatment, concomitant medications, medical history, exposures, setting, and other clinical data. This will require the identification and linkage of multiple data sources over time, which has been a fundamental challenge for clinical and public health informatics for decades. Informatics professionals in the clinical research domain should collaborate and advocate for these solutions to improve synergy between health care and clinical research information systems.

The availability of electronic health data resources is driving requirements for the representation of adverse events themselves. Historically, ADRs are identified by astute clinicians and communicated via human language for subsequent actions. As distinct pharmacovigilance systems have evolved, formal representations (i.e., controlled terminologies) for the adverse events themselves have become necessary. Because these controlled terminologies were developed explicitly for pharmacovigilance and regulated reporting, they were created anew (e.g., MedDRA) rather than by adapting existing systems. Years ago, access to clinical terminologies such as SNOMED CT was very limited, and the prospect of utilizing clinical data for pharmacovigilance purposes was not immediate. However, as the prospect of reusing existing clinical data streams is now within reach (see Chap. 17), regulatory bodies are becoming pressured to require clinical data standards and terminologies that are endorsed by their governments.

Additionally, because drug safety is now considered not only as a part of the "drug life cycle" but also as an important component of individual and population health, postmarket surveillance and population monitoring are increasingly being viewed as clinical activities that need to be supported by electronic medical records. We will likely see an increased emphasis in clinical data standards discussions on the information needs to support pharmacovigilance activities.

Because of the needs and precedent of global communication and cooperation in the area of drug safety, internationally accepted data standards will be particularly critical. From a terminological standpoint, standard names for drugs and medical devices themselves will be required, as concepts related to patient experience using that drug. This will include at a minimum clinical observations and findings, comorbidities and preexisting conditions, environmental and behavioral exposures, and concurrent treatments and procedures. Some of these terminologies have been developed specifically for this purpose and are de facto standards (e.g., WHO Drug, MedDRA). As pharmacovigilance moves to support developing countries, particularly those that are playing an increasing role in clinical research (e.g., India), then licensing models for access to these terminologies and tools for their appropriate use will become essential feature for national clinical research and population health infrastructures. Identification and use of controlled terminology and standard data models are discussed in depth in other chapters, but they are clearly very important to any comprehensive pharmacovigilance program.

Adverse events can encompass physical findings, complaints, and laboratory results [6]. Per FDA and ICH definitions, adverse events also include worsening of preexisting conditions, so comprehensive AE coding systems must encompass diseases, disorders, and conditions as well [63]. There are several competing adverse event data standards (including terminologies and classifications), each with very different structures. Having multiple standards does complicate data sharing and requires resources to achieve interoperability. An ultimate goal would be to have the same standard, or harmonious standards, across the different user communities. The need for informatics professionals to identify strategies to harmonize these standards or share data between them will increase over time. The importance of workflows – both for care delivery and pharmacovigilance – will become more visible as multiple organizations, stakeholders, and nations work toward global cooperation in pharmacovigilance.

Given the industry investment and FDA/ICH endorsement at this time, it seems likely that MedDRA may continue to be a standard for FDA reporting, and if so, there will be future needs for mappings between SNOMED CT and MedDRA. However, only clear specifications for workflows and information exchanges can inform us as to which directions and which content these mappings should be addressed and how their creation should be prioritized. Since SNOMED CT is the likely standard for the collection of clinical data in EHR, the "collect once, use many" paradigm would dictate mapping SNOMED CT (as the source clinical data collection) to specialized terminologies and classifications that would support adverse event reporting as a secondary data use. The importance of connecting the health care and research domains for pharmacovigilance will affect the development agendas for tools that allow data to be interoperable in the appropriate direction and maintain the level of precision needed for all users.

Conclusion

Pharmacovigilance is an important clinical activity with strong implications for population health and for clinical research portfolios and conduct. Pharmacovigilance has made a long journey since it started in the early 1960s after the thalidomide disaster. Recent events and current popular media coverage show that it is very much a subject that lies close to people's hearts. In the past few years, there has been a major push in trying to change the existing pharmacovigilance systems in order to meet the demands of the future. While drug safety is very much a patient-centered issue and the origins of pharmacovigilance are essential clinically, the future of pharmacovigilance will reflect systems and political issues that will drive policy and impact patient safety and human health. The core of the future pharmacovigilance is the systematic collection of valid and representative data that can be rigorously analyzed, interpreted, and acted upon as part of medical care. Scientific underpinning of pharmacovigilance is needed to make sure that it will develop as a scientific discipline and thereby contribute to the innovation needed in this field. Pharmacogenetics will play a role in identifying individual risk factors for the occurrence of certain ADRs, and these data will be critical to address in future systems [64]. The pharmacovigilance of tomorrow must be able to identify new safety issues without delay and be able to communicate that information to patients and providers in languages and representation that facilitate decision making and protective action for patients everywhere.

References

1. WHO: The importance of pharmacovigilance. The WHO collaborating centre for international drug monitoring. 2002. http://www.who-umc.org. Accessed 15 May 2010.
2. Harmark L, van Grootheest AC. Pharmacovigilance: methods, recent developments and future perspectives. Eur J Clin Pharmacol. 2008;64:743–52.
3. McBride W. Thalidomide and congenital malformations. Lancet. 1961;2:1358.
4. Wade OL. The dawn of concern. In: Adverse reactions to drugs: 1–10. Chichester: Acford Ltd; 1970.
5. Randell T. Thalidomide's back in the news, but in more favorable circumstances. JAMA. 1990;263:467–8.
6. Friedman LM, Furberg CD, DeMets DL. Assessing and reporting adverse events. In: Fundamentals of clinical trials. New York: Springer; 1998. p. 170–84.
7. Lindquist AM. Seeing and observing in international pharmacovigilance. Academic thesis. Nijmegan: Katholieke Universiteit Nijmegan; 2003.
8. WHO. Handbook of resolutions and decisions of the World Health Assembly and Executive Board. 11th ed. Geneva: World Health Organization; 1972. WHA 20.51.
9. WHO. International drug monitoring: the role of national centres, Technical Report Series. Geneva: WHO; 1972.
10. Olsson S. The role of the WHO programme on international drug monitoring in coordinating worldwide drug safety efforts. Drug Saf. 1998;19:1–10.
11. Mitka M. Report criticizes lack of FDA oversight. JAMA. 2006;296:920.
12. Lenzer J. FDA is incapable of protecting US against another Vioxx. Br Med J. 2004;329:1253.
13. Furberg CD, et al. The FDA and drug safety: a proposal for sweeping changes. Arch Intern Med. 2006;166:1938–42.
14. Strom B. How the US drug safety system should be changed. JAMA. 2006;295:2072–5.
15. Ray WA, Stein CM. Reform of drug regulation – beyond an independent drug-safety board. N Engl J Med. 2006;354:194–201.
16. Avorn J. Paying for drug approvals – who's using whom? N Engl J Med. 2007;356:1697.

17. Psaty BM, Charo RA. FDA responds to institute of medicine drug safety recommendations – in part. JAMA. 2007;297:1917–20.
18. Baciu A, Stratton AK, Burk S. The future of drug safety: promoting and protecting the health of the public. Washington D.C.: National Academy Press; 2007.
19. Psaty B, Burke S. Protecting the health of the public – Institute of Medicine recommendations on drug safety. N Engl J Med. 2006;355:1753.
20. EMEA: Assessment of the European community system of pharmacovigilance. 2006. http://publica.fraunhofer.de/eprints/urn:nbn:de:0011-n-499398.pdf. Accessed 17 Aug 2011.
21. Raine J. Risk management – a European regulatory view. In: Andrews E, editor. Pharmacovigilance. Chichester: Wiley; 2007.
22. Waller PC, Coulson RA, Wood SM. Regulatory pharmacovigilance in the United Kingdom: current principles and practices. Pharmacoepidemiol Drug Saf. 1996;5:363–75.
23. Hauben M, Madigan D, Gerrits CM, et al. The role of data mining in pharmacovigilance. Expert Opin Drug Saf. 2005;4:929–48.
24. Egberts AC, Meyboom RH, van Puijenbroek EP. Use of measures of disproportionality in pharmacovigilance: three Dutch examples. Drug Saf. 2002;25:453–8.
25. Moore N. The role of the clinical pharmacologist in the management of adverse drug reactions. Drug Saf. 2001;24:1–7.
26. Edwards IR. Pharmacovigilance – beyond 2000. Opinion & evidence. Auckland: Adis; 2000.
27. Gross CP, Mallory R, Heiat A, et al. Reporting the recruitment process in clinical trials: who are these patients and how did they get there? Ann Intern Med. 2002;137:10–6.
28. Zarin DA, Young JL, West JC. Challenges to evidence-based medicine: a comparison of patients and treatments in randomized controlled trials with patients and treatments in a practice research network. Soc Psychiatry Psychiatr Epidemiol. 2005;40:27–35.
29. Heiat A, Gross CP, Krumholz HM. Representation of the elderly, women, and minorities in heart failure clinical trials. Arch Intern Med. 2002;162:1682–8.
30. Wardell WM, Tsianco MC, Anavekar SN, et al. Postmarketing surveillance of new drugs: II. Case studies. J Clin Pharmacol. 1979;19:169–84.
31. van Grootheest K, Olsson S, Couper M, et al. Pharmacists' role in reporting adverse drug reactions in an international perspective no conflict of interest was declared. Pharmacoepidemiol Drug Saf. 2004. doi:10.1002/pds.897.
32. van Grootheest K, De Jong-vanden Berg L. Patients' role in reporting adverse drug reactions. Expert Opin Drug Saf. 2004;3:363–8.
33. van Grootheest AC, Passier JLM, van Puijenbroek EP. Direct reporting of side effects by the patient: favourable experience in the first year. Ned Tijdschr Geneeskd. 2005;149:529–33.
34. Eland IA, Belton KJ, van Grootheest AC, et al. Attitudinal survey of voluntary reporting of adverse drug reactions. Br J Clin Pharmacol. 1999;48:623–7.
35. Hazell L, Shakir S. Under-reporting of adverse drug reactions: a systematic review. Drug Saf. 2006;29:385.
36. Edwards I. Spontaneous reporting – of what? Clinical concerns about drugs. Br J Clin Pharmacol. 1999;48:138.
37. Clarke A, Deeks JJ, Shakir SAW. An assessment of the publicly disseminated evidence of safety used in decisions to withdraw medicinal products from the UK and US markets. Drug Saf. 2006;29:175–81.
38. Mann R. Prescription-event monitoring – recent progress and future horizons. Br J Clin Pharmacol. 1998;46:195.
39. Mackay F. Post-marketing studies: the work of the drug safety research unit. Drug Saf. 1998;19:343.
40. Coulter D. The New Zealand intensive medicines monitoring programme in pro-active safety surveillance. Pharmacoepidemiol Drug Saf. 2000;9:273–80.
41. Coulter D. The New Zealand intensive medicines monitoring programme. Pharmacoepidemiol Drug Saf. 1998;7:79–90.
42. Shakir S. PEM in the UK. In: Pharmacovigilance. Chichester: Wiley; 2002. p. 333–44.

43. European Commission, European Commission Enterprise Industry Directorate. General strategy to better protect public health by strengthening and rationalising EU pharmacovigilance. Brussels: European Commission; 2007.
44. Strom B, editor. Pharmacoepidemiology. 4th ed. Chichester: Wiley; 2005.
45. van Puijenbroek E, Diemont W, van Grootheest K. Application of quantitative signal detection in the Dutch spontaneous reporting system for adverse drug reactions. Drug Saf. 2003;26:293–301.
46. Bate A, Lindquist M, Edwards IR, et al. A data mining approach for signal detection and analysis. Drug Saf. 2002;25:393.
47. Szarfman A, Machado SG, O'Neill RT. Use of screening algorithms and computer systems to efficiently signal higher-than-expected combinations of drugs and events in the US FDA's spontaneous reports database. Drug Saf. 2002;25:381–92.
48. van Puijenbroek EP, Bate A, Leufkens HG, et al. A comparison of measures of disproportionality for signal detection in spontaneous reporting systems for adverse drug reactions. Pharmacoepidemiol Drug Saf. 2002;11:3–10.
49. Sim I, Chan AW, Gülmezoglu AM, et al. Clinical trial registration: transparency is the watchword. Lancet. 2006;367:1631–3.
50. EMEA: Human medicines – EMEA pre-submission guidance. 2005. http://www.emea.europa.eu/htms/human/presub/list.htm. Accessed 15 May 2010.
51. EMEA: Guideline on risk management systems for medicinal products for human use. 2005. http://www.emea.europa.eu/pdfs/human/euleg/9626805en.pdf. Accessed 15 May 2010.
52. Hugman B. The Erice declaration: the critical role of communication in drug safety. Drug Saf. 2006;29:91.
53. van Grootheest K, de Graaf L. Consumer adverse drug reaction reporting: a new step in pharmacovigilance? Drug Saf. 2003;26:211.
54. de Langen J, et al. Adverse drug reaction reporting by patients in the Netherlands: three years of experience. Drug Saf. 2008;31:515.
55. Harmark L, et al. Web-based intensive monitoring, a new patient based tool for early signal detection. Drug Saf. 2006;29:911.
56. Tang PC, et al. Personal health records: definitions, benefits, and strategies for overcoming barriers to adoption. J Am Med Inform Assoc. 2006;13:121–6.
57. McClellan M. Drug safety reform at the FDA–pendulum swing or systematic improvement? N Engl J Med. 2007;356:1700.
58. Zwillich T. US lawmakers tackle safety reforms at the FDA. Lancet. 2007;369:1989–90.
59. Hennessy S, Strom BL. PDUFA reauthorization – drug safety's golden moment of opportunity? N Engl J Med. 2007;356:703–1704.
60. Coombes R. FDA tightens its grip on drug regulation. Br Med J. 2007;334:290.
61. EMEA: European risk management strategy: achievements to date. 2007. http://www.emea.europa.eu/pdfs/human/phv/30816707en.pdf. Accessed 15 May 2010.
62. EMEA: Public status report on the implementation of the European risk management strategy. 2007. http://www.emea.europa.eu/pdfs/human/phv/16895407en.pdf. Accessed 15 May 2010.
63. 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;29:635–45.
64. Pirmohamed M, Park BK. Genetic susceptibility to adverse drug reactions. Trends Pharmacol Sci. 2001;22:298–305.

Chapter 20
Clinical Trials Registries and Results Databases

Karmela Krleža-Jerić

Abstract 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. In this new paradigm, knowledge will be generated from the *culmination* of all existing knowledge – not just from parts and bits 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, advance research more efficiently, provide more reliable evidence for diagnostic and therapeutic prescriptions, and regain public trust. 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. In the case of clinical trials, the public disclosure of data is realized by posting them in well-defined, freely accessible clinical trial registries and results databases.

Keywords Transparency in clinical research • Trials registries • Registry design • Registry quality • International standards • Results databases • Primary registry

K. Krleža-Jerić, M.D., M.Sc., D.Sc.
Knowledge Translation, Canadian Institutes of Health Research,
Ottawa, ON, Canada,

Department of Epidemiology and Community Health, Medical Faculty,
University of Ottawa, Croatian Medical Journal, Ottawa, ON, Canada
e-mail: karmela.krleza-jeric@cihr-irsc.gc.ca, krlezajk@hotmail.com

R.L. Richesson, J.E. Andrews (eds.), *Clinical Research Informatics*, Health Informatics,
DOI 10.1007/978-1-84882-448-5_20, © Springer-Verlag London Limited 2012

Background

Rationale for Trial Registration

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. In this new paradigm, knowledge will be generated from the *culmination* of all existing knowledge – not just from parts and bits of previous knowledge, as is largely the case now.

We are in the era of evidence-informed decision-making in health care 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. Registration of protocol items, publication of the complete protocol, and public disclosure of trial findings in peer-reviewed journals – complemented with public (internet-based) disclosure of results that include microlevel data collectively – 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 all of these personal and policy decisions. Randomized clinical trials and their systematic reviews are considered the gold standards for evidence creation, as illustrated by positioning them at the top of the pyramid of evidence (Fig. 20.1). This position of trials on the evidence pyramid implies that the reliability of data that they collect is very important. Because they may be directly implemented in decision-making, their quality should be constantly scrutinized. Unfortunately, the reliability of trial-based evidence is questionable due to the publication and outcome reporting bias of trials. Consequently, incomplete evidence can lead to biased clinical decisions, with often harmful consequences, and damages the public trust in research and medical interventions. Following medical deontology, doctors' prescription habits are supposed to be judiciary, which would require the complete and total knowledge of benefits and potential harms that a given medication or simultaneously prescribed medications might have. This is difficult at best, and impossible if the information about the given diagnostic tools, medications, or devices is not available.

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 [1], provide more reliable evidence for diagnostic and therapeutic prescriptions, and regain public trust. 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. In the case of clinical trials, the public disclosure of data is realized by posting them in well-defined, freely accessible clinical trial registries and results databases.

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

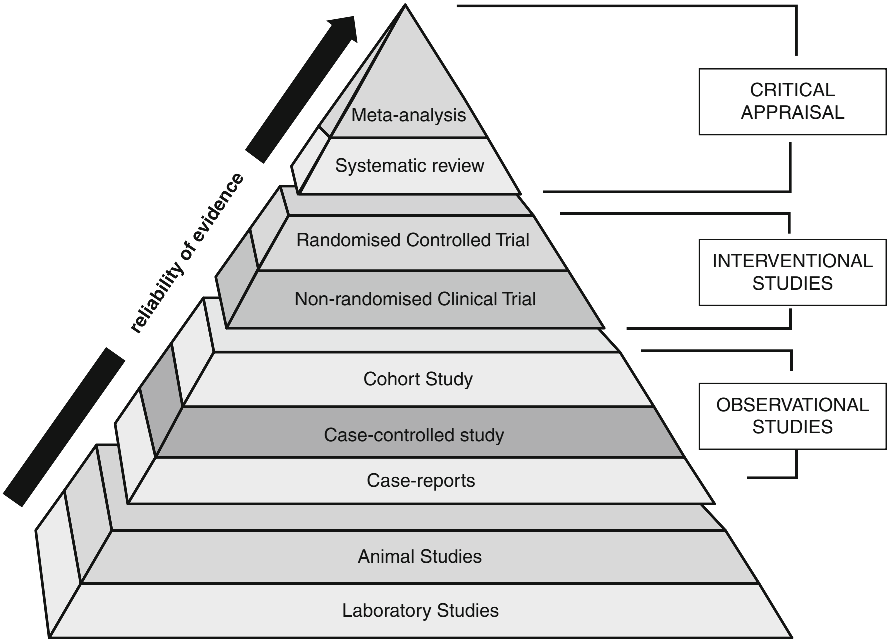


Fig. 20.1 Evidence pyramid – hierarchy of evidence

constantly and internationally scrutinized. Thus, the standards for trial registration and registries should be internationally defined and relevant.

Development of Trial Registration

Although the need for trial registration 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. Its practical development started around 2000 with two critical boosts in 2004 and in 2006. The 2004 New York State Attorney General vs. Glaxo case [2, 3] inspired the International Council of Medical Journal Editors (ICMJE) [4] and Ottawa statements [5], as well as the recommendations of the Mexico Ministerial Summit [6]. These led to the development of international standards for trial registration by the World Health Organization (WHO), which were launched in 2006 and changed the landscape of trial registration worldwide [7].

By 2004, a number of circumstances had coincided, which enabled the development of trial registration and subsequent standards. These include:

- The Internet-enabled storage and retrieval of large data sets
- The existence and experience of two major registries: International Standard Randomized Clinical Trials Number (ISRCTN), based in the United Kingdom, and ClinicalTrials.gov, based in the United States of America

- The ever-increasing 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 pressure from developing countries to share research data
- The appreciation of the impact of trial registration on knowledge creation, sharing, and translation
- 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 are essential for achieving the goal of evidence-based decision-making. Because they identify existing registries and trials that need to be registered, define the minimum data set, designate the timing of registration, and assign unique numbers to trials, international standards facilitate the comparability of data in various registries and the development of any new national or regional registries. It is important to note that as of 2011, there are no international standards for results disclosure; however, they are likely to be developed in the near future and will create numerous opportunities for informatics and information technology (IT) experts to leverage an apply to new applications. Additionally, further evolution of trial registration and its standards is also expected, again leading to new applications and resources that can impact the development of new research and our understanding of health, disease, and effective therapies.

The goal of research transparency has been to have the protocol documents electronically available. For example, the protocol documents would be posted on the registry website, and all trial-related data from them would be cross-referenced to results and findings. A trial protocol can be very complex and lengthy, which can make finding the needed information difficult. An international group has been working on defining Standard Protocol Items for RandomIzed Trials (SPIRIT) as can be seen on the Enhancing QUAlity and Transparency Of health Research (EQUATOR) website [8]. SPIRIT [9] is expected to contribute to clarity and ensure that needed items are included in the protocol. This might also facilitate public disclosure, especially in combination with the growing use of electronic data management [10]. 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. With developing methodology, increasing requests for transparency, and ongoing analysis and evaluation, trial registration standards will have to be revisited frequently, and trial registries will most certainly expand to include results or cross-references to results databases.

Trial Registries

Many different kinds of clinical trial registries exist in the public and private domains, such as international-, country-, and region-specific registries, and corporate (sponsor-driven) registries. This 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 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 define characteristics of registries and their content.

Standards, Policies, and Principles

Because trials are conducted internationally, 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, and various nations. It is important to note that individual countries often implement international standards by adopting and extending them with additional fields to host more information as needed for 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 dataset that needs to be provided to the registry, often 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 the 20-item minimum data set containing the following:

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

In order to foster the implementation of these standards, to facilitate creation of new registries, and to identify the best practice for trial registration, 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) search portal is a unique global portal to the trials in registries that meet criteria (i.e., WHO primary registries and ClinicalTrials.gov) but does not provide access to the full extent of registries' data. Instead, the predefined 20-item data set provided in English by the registries is displayed. 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 and users to get the full picture of a given trial, from start to finish.

Characteristics and Design Features of Trial Registries

Patient registries are described in depth in Chap. 13. Although patient and trial registries might be confused, as they both capture certain disease-related information and often use Internet-based depositories, these two types of registries are quite different. Patient registries 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 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 leading for 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 between a few and several thousand participants. In a patient registry, each entry is an individual patient with the same disease or a condition of the same group, usually chronic diseases. For example, there are cancer, psychosis, and rare disease patient registries.

The most important difference between trial and patient registries is the purpose, which is reflected in the status. 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 are developed in order to answer epidemiological questions such as incidence and prevalence, natural course of disease(s), and disease-related lethality.

Some trial registries also aim at informing the potential trial participants in order to enhance recruitment. Besides being a transparency tool, registries are also a learning tool, and one could argue that they may help improve the quality of the protocol and, as a result, the quality of conducted trials. 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, termed *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 s*tatus* 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 the trail registry user.

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 adhering to international standards tend to add more fields to meet their registry-specific, often country-specific, needs. Regardless of these additional fields, the essential 20 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 20-item fields. Those registries that collect more data typically have more detailed and quality data for each trial record and are potentially more useful. 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 often including drop-down lists [11].

In order to identify best practices and improve this tool for entering new trial protocol records, as well as to provide support in case of the development of new registries, WHO formed the Working Group on Best Practice for Clinical Trial Registries in 2008 [12]. The working group includes primary and some partner registries. As of May 2011, there were 13 WHO primary registries and the ClinicalTrials.gov registry that directly provide data to the search portal. As can be seen from their geographic distribution shown in Fig. 20.2, the network includes at least one registry per continent.

Clinical trial registries 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 raw data repositories. These links are expected to increase as results databases and raw data repositories are 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 institutional ethics approval is obtained.

Although international standards require registration prior to recruitment of trial participants, this is still not fully implemented [11, 13]. 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 of the unique ID to each trial upon registration and its subsequent use enables any stakeholder to easily find what interests them.

Some countries hesitate to simply "import" the international standards or policies out of fear that they might change and put the country (regulator, funding

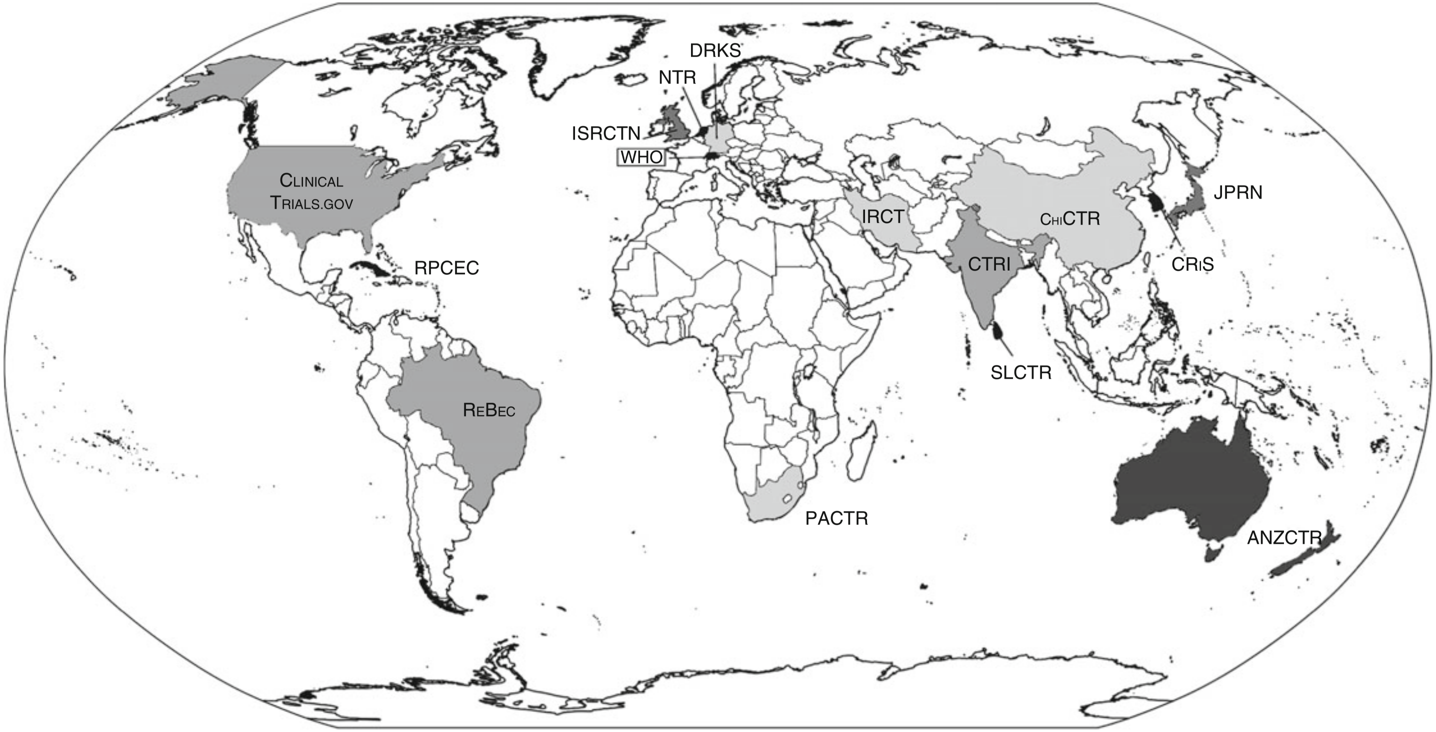


Fig. 20.2 Network of registries providing data to WHO search portal and the WHO portal. This map provides the worldwide distribution of registries that directly provided data to WHO as of May 2011. *ANZCTR* Australian New Zealand Clinical Trials Registry, *ReBec* Brazilian Clinical Trials Registry, *ChiCTR* Chinese Clinical Trial Registry, *CRiS* Clinical Research Information Service, Republic of Korea, ClinialTrials.gov (USA), *CTRI* Clinical Trials Registry, India, *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, *SLCTR* Sri Lanka Clinical Trials Registry, *WHO* WHO Search Portal, Geneva

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) [14], which obliges physicians via their national medical associations and is thus implicitly implemented. Notably, DoH calls specifically for registration and results disclosure of trials [15].

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 on the quality and accuracy of data and the timing of reporting. 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 [16].

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 data [11, 17, 18]. 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 of attaining an acceptable level of data quality. (The practical and theoretical aspects of data quality are described in Chap. 10.)

The ongoing and numerous analyses and evaluations of the implementation of the existing 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) [19]. 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, or 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. 9.

One of concerns for trial registries is the issue of duplicate registration. Duplicate registration of trials, especially of multicenter and multicountry 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 primary registries and 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 intraregistry 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 and 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) [20], 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 [16].

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 ClincalTrials.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 (2, 3). 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 and 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 at meeting the needs of all involved in order to bring 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. In the USA, registration in ClinicalTrials.gov is required by law [21]. 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 [22], was launched in February 2011. It is expected that such registries will function based on similar principles as trial registries. For example, PROSPERO will prospectively register a systematic review (i.e., its design and conduct, protocol, or equivalent) and will later display a link to eventual publication of the completed review. All the information will be 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 will be 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.

Besides 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 [16].

Creation and Management of a Trial Registry: The User Perspective

Design of Trial Registries

As mentioned earlier, each primary trial registry contains fields for a 20-item minimum dataset defined by the international standards and usually a few more. These include the fields for the ID assigned by any other registry, the unique trial registration number (UTRN) assigned by WHO, ethics approval(s), trial website URL, publications, etc. The 20 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 dataset, arguments have been built for publishing the full protocol, and some journals have already started doing so. It will be particularly interesting to have electronic versions of structured, computerized protocols; however, even when that happens, the data provided in trial registries will be useful as a summary of the protocol. These two tools of protocol transparency will probably attract different users.

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 [11]. These instructions, inspired by the WHO standards, might be developed by regulators in combination with the registry and/or journal editors. 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 enable the entering of additional specific data. Registries usually have levels of compulsory completion of fields that cannot be skipped. Furthermore, they might indicate which fields/items are required by the WHO standards and/or by their national regulator.

Data Fields

Design of fields 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, cross-over, 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 a RCT but will also indicate if it is a parallel or cross-over 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"

[23]. 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, but it has also shown improvement in this area over time [24, 25]. 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 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 an *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. Currently, 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 registry 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: Repositories

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. Public disclosure of results in such repositories will complement publication in peer-reviewed journals, and it is an integral part of the transparency tool set.

Results databases or repositories are complex, and they might include aggregate data, metadata, and microdata (i.e., individual participant dataset, also known as raw data). Similarly, to trial registries, results repositories are expected to build hyperlinks, the most important ones being between the given trial in the registry and related publications or systematic reviews. As of 2011, results databases and repositories are 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 [26], there are numerous issues to be resolved in order to get the results data, especially microlevel data sets, publicly disclosed.

There are no international standards for public disclosure of trial results, and there are no standards or repositories for meta- or microdata. However, there is a lot of discussion on how these should be designed and some initiatives that have been contributing to accumulation of experience [15, 18, 26–30]. The journal *Trials* started posting them on the Internet as the series "Sharing clinical research data," edited by Andrew Vickers [31].

When talking about results disclosure, there is a whole spectrum, from aggregate to full raw data sets. Public disclosure of aggregate data and findings beyond those published in peer-reviewed journals seems to be a starting point. Since 2008, individual experiences have been accumulated by ClinicalTrials.gov and by the European Medicines Agency (EMEA) [18, 32]. ClinicalTrials.gov is implementing the FDAAA of 2007 [21] and has developed fields for aggregate results of registered trials, which are being cross-referenced to its registry [18]. Apparently, it seems that ClinicalTrials.gov and EMEA are working on developing comparable data fields which might inform future development of international standards.

Some results disclosure issues are comparable to those related to trial registries and include the development of 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 microlevel datasets. Those include the cleaning of data, quality of data, accountability, defining who is the guarantor of truth, privacy issues, and issues related to depersonalization efforts and intellectual property rights. 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 issues, such as cleaning of raw data, are becoming less of an issue as they take place simultaneously with the data collection. A lot can be learned from the experience of genome data, for which many have shown that data sharing has boosted the development of the field [10, 33–35].

However, the problems with microlevel data involving individual trial participant data sets are far from being solved. There is no single repository, and there are no standards. These data exist, but they are either protected in the hands of regulators or might be shared with systematic reviewers upon request and only under certain conditions. In order to facilitate systematic reviews and meta-analysis, several journal editors [28–30] are now encouraging data sharing upon publication of trial findings. There have been research-type, experimental efforts to create trial results repositories, some in collaboration with journals [36, 37]. Several dilemmas will have to be studied and resolved, including the balance of privacy versus transparency. Many of these issues will require revisiting and modernization. In the initial stage, one might expect varying levels of accessibility to more or less "de-indentified" microlevel data. All these elements will create a new challenge, a need for interdisciplinary work, and an opportunity for clinical research informatics and information technology experts.

Data sharing is becoming more and more appealing to all stakeholders. Earlier hesitation has been gradually lightening, and we are witnessing increased transparency and a change in the research paradigm. One illustration of this is the increasing registration of phase I trials by major pharmaceutical companies. By the time this book is printed, there will be even more developments in this area as it is constantly and rapidly evolving.

Conclusion

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 (Fig. 20.3). Although it has not as yet been completely accomplished, there is a clear tendency to move in that direction, which will ensure a high level of transparency.

Trial registries host selected defined 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 international standards. Furthermore, there is a strong push for publication of the full protocol, either in the registry or elsewhere. If this were to happen, the international data set that is currently available in registries will be a valuable summary and will include links to the full protocol, publications, trial website, systematic review, and results repositories. It is

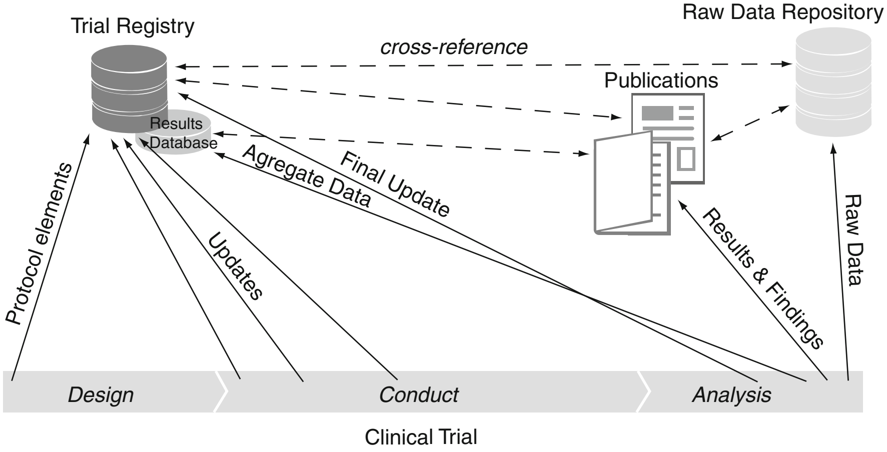


Fig. 20.3 Anticipated flow of data from clinical trial to public domain

expected that even when full protocols become publicly available, registries will continue to provide summaries of protocols and thus continue to play an important role in achieving trial transparency.

Results repositories are in their early stage of development, and they currently lack international standards. The aggregate data repositories are being formed by trial registries or regulators and aim at providing timely aggregate data in predefined tables. For example, ClincialTrials.gov displays results of trials it registered in a results database in accordance with the FDAAA.

There are no international standards to govern the public disclosure (of de-identified data) at the individual participant level data (i.e., microdata or raw data), but there is growing interest and even prototype implementations in this level of data accessibility. Development of such standards will require thorough planning, analysis of quality control, resources, as well as revisiting the privacy and proprietary rules and practices. Furthermore, it is expected that existing systematic reviews will be updated with the results of a given trial to inform various levels of decision-making with the updated evidence. BioMed Central (BMC) and its journal *Trials* opened a discussion about the public disclosure of individual participant datasets of clinical trials and started publishing them [28, 29]. 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.

Disclaimer The views expressed here are the author's and do not represent the views of the Canadian Institutes of Health Research or any other organization.

Acknowledgments The author would like to thank Ludovic Reveiz for support in content, Penny Ashwanden for editorial support, and Nevena at Apropo Media for graphic design.

References

1. Krleza-Jeric K. Clinical trial registration: the differing views of industry, the WHO, and the Ottawa Group. PLoS Med. 2005;2(11):e378.
2. Bass A. Side effects: a prosecutor, a whistleblower, and a bestselling antidepressant on trial. 1st ed. Chapel Hill: Algonquin Books of Chapel Hill; 2008.
3. Gibson L. GlaxoSmithKline to publish clinical trials after US lawsuit. BMJ. 2004;328:1513.
4. De Angelis C, 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. Lancet. 2004;364:911–2.
5. Krleza-Jeric K, Chan AW, 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. 2005;330(7497):956–8.
6. World Health Assembly – Ninth plenary meeting. WHA 58–34 ministerial summit on health research. 2005 May 25. http://www.who.int/rpc/meetings/58th_WHA_resolution.pdf.
7. World Health Organization. International clinical trials registry platform (ICTRP). 2011. http://www.who.int/ictrp/en/. Accessed 26 Jan 2012.
8. EQUATOR Network. Enhancing the quality and transparency of health research. 2011. http://www.equator-network.org. Accessed 2 Feb 2009.
9. Chan A-W, Tetzlaff J, Altman DG, et al. The SPIRIT initiative: defining standard protocol items for randomized trials [conference abstract]. German J Evid Qual Health Care. 2008; 102(Suppl):S27.
10. El EK, Jonker E, Sampson M, Krleza-Jeric K, Neisa A. The use of electronic data capture tools in clinical trials: web-survey of 259 Canadian trials. J Med Internet Res. 2009;11(1):e8.
11. Reveiz L, Chan AW, Krleza-Jeric 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. 2010;5(8):e12484.
12. World Health Organization. The WHO registry network. International clinical trial registries platform (ICTRP). 2011. http://www.who.int/ictrp/network/en/. Accessed 3 Jan 2011.
13. Reveiz L, Krleza-Jeric K, Chan AW, De AS. Do trialists endorse clinical trial registration? Survey of a Pubmed sample. Trials. 2007;8:30.
14. World Medical Association. World Medical Association Declaration of Helsinki; ethical principles for medical research involving humans. 2009. http://wwwwmanet/e/policy/pdf/17cpdf. Cited 18 Jun 2009.
15. Krleza-Jeric K, Lemmens T. 7th revision of the declaration of Helsinki: good news for the transparency of clinical trials. Croat Med J. 2009;50(2):105–10.
16. 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. 2011;30(1):87.
17. Rising K, Bacchetti P, Bero L. Reporting bias in drug trials submitted to the food and drug administration: review of publication and presentation. PLoS Med. 2008;5(11):e217.
18. Tse T, Zarin DA. Clinical trial registration and results reporting: ClinicalTrials.gov and FDAAA. Update [1]:18–22, Washington, D.C., FDLI; 2009.
19. Clinical Data Interchange Standards Consortium. 2011. http://www.cdisc.org/. Accessed 26 Jan 2012.
20. World Health Organization. Unique numbering system for clinical trials. International clinical trial registries platform (ICTRP). 2009. http://www.who.int/ictrp/news/utn/en/index.html. Accessed 15 May 2010.
21. Food and Drug Administration Amendments Act of 2007. Public Law 110–85. 27 Sept 2007.
22. National Institute for Health Research (NHS). PROSPERO: international prospective register of systematic reviews. http://www.crd.york.ac.uk/PROSPERO/ Accessed 26 Jan 2012.
23. Australia New Zealand Clinical Trials Registry. Data item definition/explanation. http://www.anzctr.org.au/docs/ANZCTR%20Data%20field%20explanation%20V5.pdf. Accessed 26 Jan 2012.

24. Zarin DA, Tse T, Ide NC. Trial registration at ClinicalTrials.gov between May and October 2005. N Engl J Med. 2005;353(26):2779–87.
25. Zarin DA, Ide NC, Tse T, Harlan WR, West JC, Lindberg DA. Issues in the registration of clinical trials. JAMA. 2007;297(19):2112–20.
26. Krleza-Jeric K. International dialogue on the public reporting of clinical trial outcome and results – PROCTOR meeting. Croat Med J. 2008;49(2):267–8.
27. Ottawa Group. Principles of operationalisation of results reporting. 2006. http://ottawa-groupohrica/docs/Results_Dublinpdf. Cited 14 Jun 2009.
28. Hrynaszkiewicz I, Altman DG. Towards agreement on best practice for publishing raw clinical trial data. Trials. 2009;10:17.
29. Hrynaszkiewicz I, Norton ML, Vickers AJ, Altman DG. Preparing raw clinical data for publication: guidance for journal editors, authors, and peer reviewers. Trials. 2010;11:9.
30. Laine C, Goodman SN, Griswold ME, Sox HC. Reproducible research: moving toward research the public can really trust. Ann Intern Med. 2007;146(6):450–3.
31. Vickers A. Thematic series: sharing clinical research data. Trials. 22 Nov 2010; 2006–2011. http://www.trialsjournal.com/series/sharing. Accessed 15 Dec 2011.
32. Smyth RL. Making information about clinical trials publicly available. BMJ. 2009;338:b2473.
33. Collins F. Has the revolution arrived? Nature. 2010;464(7289):674–5.
34. Collins FS, Green ED, Guttmacher AE, Guyer MS. A vision for the future of genomics research. Nature. 2003;422(6934):835–47.
35. Edwards AM, Bountra C, Kerr DJ, Willson TM. Open access chemical and clinical probes to support drug discovery. Nat Chem Biol. 2009;5(7):436–40.
36. Sim I, Rennels G. A trial bank model for the publication of clinical trials. Proc Annu Symp Comput Appl Med Care. 1995(1):863–7.
37. Sim I, Owens DK, Lavori PW, Rennels GD. Electronic trial banks: a complementary method for reporting randomized trials. Med Decis Making. 2000;20(4):440–50.

Chapter 21
Future Directions in Clinical Research Informatics

Peter J. Embi

Abstract 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 need to be developed to achieve this, and this innovation is what will drive the field forward in the coming years.

Keywords Clinical research informatics • Biomedical informatics • Phases of translation research • Electronic health records • Future trends • US policy initiatives • Health IT infrastructure

As evidenced by the production of this text and reflected in its chapters, clinical research informatics (CRI) has clearly emerged 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 to be expected. As described throughout the text, the myriad stakeholders in CRI, and their roles in the healthcare, research, and informatics enterprises, are continually evolving, fueled by technological, scientific, and socioeconomic changes. These changing roles bring new challenges for research conduct and coordination but also bring potential for new

P.J. Embi, M.D., M.S.
Department of Biomedical Informatics and Internal Medicine,
The Ohio State University Medical Center,
3190 Graves Hall, 333 W. 10th Avenue, Columbus, OH 43210, USA
e-mail: peter.embi@osumc.edu

R.L. Richesson, J.E. Andrews (eds.), *Clinical Research Informatics*, Health Informatics,
DOI 10.1007/978-1-84882-448-5_21, © Springer-Verlag London Limited 2012

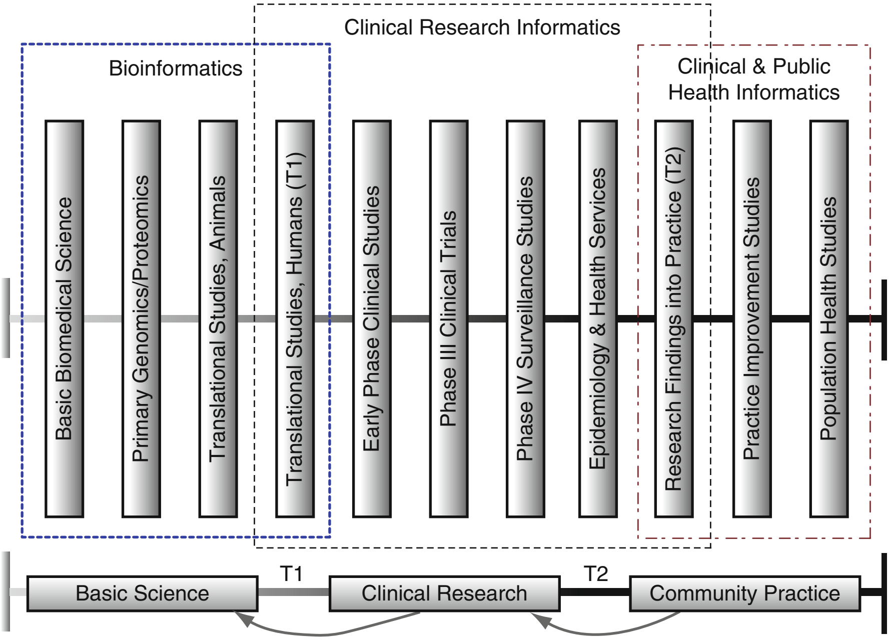

Clinical Research Informatics
Bioinformatics
Clinical & Public Health Informatics
Basic Biomedical Science
Primary Genomics/Proteomics
Translational Studies, Animals
Translational Studies, Humans (T1)
Early Phase Clinical Studies
Phase III Clinical Trials
Phase IV Surveillance Studies
Epidemiology & Health Services
Research Findings into Practice (T2)
Practice Improvement Studies
Population Health Studies
Basic Science
T1
Clinical Research
T2
Community Practice

Fig. 21.1 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)

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. 21.1 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.

Policy Trends

It should therefore come as no great surprise that recent years have seen the emergence of several national and international research and policy efforts to foster advances in CRI by supporting CRI professionals' efforts to address the inherent challenges and opportunities that motivate the subdiscipline. Focused on accelerating and improving clinical research capacity and capabilities in the biomedical sector, 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 Cancer Institute (NCI), such as the Cancer Biomedical Informatics Grid (caBIG) [2–5], to the National Institutes of Health's (NIH) Clinical and Translational Science Award (CTSA) [6, 7] programs [8, 9]. In recent years, the CTSA program in particular has had fostered significant growth in both the practice and science of CRI as well as fostering professional development of CRI, given one of its major emphases the advancement of CRI, and the closely related domains of translational research informatics and translational bioinformatics. Further, other NIH institutes like the National Library of Medicine, as well as funders like the Agency for Healthcare Research and Quality (AHRQ), are also driving advances in research data methods and techniques for CRI-related efforts, including comparative effectiveness and health services research.

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 IT (ONC), and the US Centers for Medicare and Medicaid Services (CMMS) are serving to incentivize the adoption and "meaningful use" of electronic health records (EHRs). Such movement toward more widespread health IT infrastructure, while initially focused primarily on improving patient care, is meant ultimately to lead an interoperable infrastructure that will enable a national health information network in the United States. Once in place and enabled via appropriate health information interchange standards, such a network is envisioned to leverage the reuse of data and information from clinical care for improvements in public health and research – to create the learning health system [10]. Just as biomedical informatics approaches and resources are essential to realizing the potential of such systems for enhancing clinical care, so too will CRI methods, theories, and tools be critical to realizing the potential of such a system for enabling discovery through acceleration and enhancement of clinical research.

Data Management and Quality

Indeed, fully leveraging our healthcare and research investments to advance human health will require even more emphasis on making sense of the ever increasing amounts of data generated through healthcare 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 healthcare data for population level science and health services research. Importantly, these advances will 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 [11]. By facilitating all of the information-dense aspects of clinical research, CRI methods and resources enable the conduct of such research programs to generate new and impactful knowledge. In fact, the truly "meaningful use" 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. However, realizing this promise will require the attention and efforts of experts focused on advancing the domain of CRI.

As the preceding chapters also demonstrate, advances in CRI have already begun to enable significant improvements in the quality and efficiency of clinical research [8, 9, 12]. 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 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.

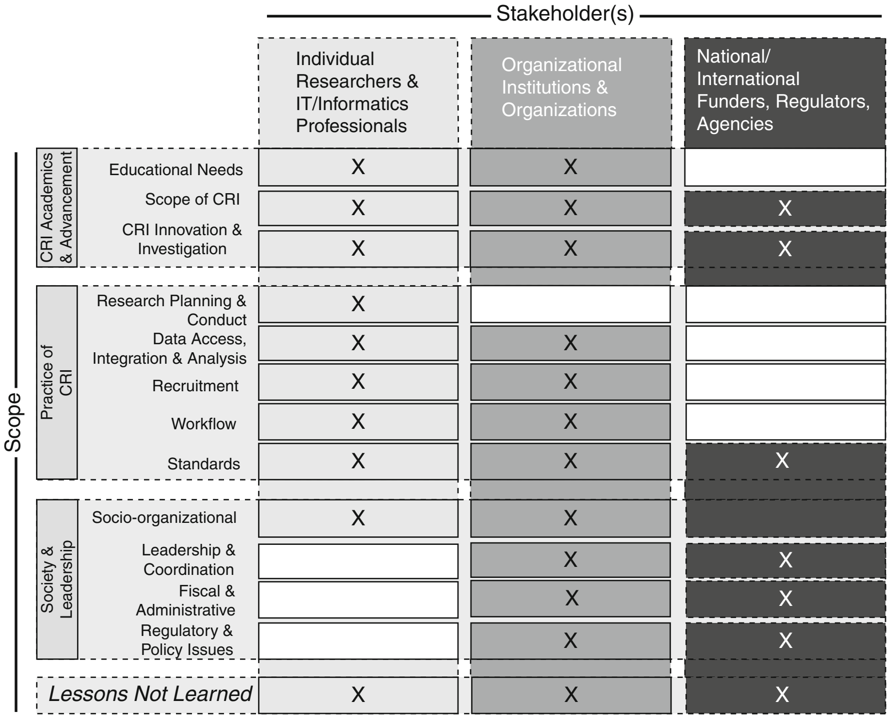
Multidisciplinary Collaboration

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 one such well-recognized organization. Working groups focused on CRI within organizations like AMIA have seen considerable growth in interest and attendance over the past decade. More recently, scientific conferences dedicated to CRI and the closely related informatics subdiscipline of translational bioinformatics (TBI) have been launched by AMIA to great success among the informatics and clinical/translational research communities. AMIA's journal, JAMIA, has also recently acknowledged the importance of CRI, with the addition of editorial board members and allotted journal space to the important topics in CRI, as have others. 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, and a myriad of professional medical societies) also increasingly provide coverage of 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 stakeholders groups (Fig. 21.2) [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.



Scope		Stakeholder(s): Individual Researchers & IT/Informatics Professionals	Organizational Institutions & Organizations	National/ International Funders, Regulators, Agencies
CRI Academics & Advancement	Educational Needs	X	X	
	Scope of CRI	X	X	X
	CRI Innovation & Investigation	X	X	X
Practice of CRI	Research Planning & Conduct	X		
	Data Access, Integration & Analysis	X	X	
	Recruitment	X	X	
	Workflow	X	X	
	Standards	X	X	X
Society & Leadership	Socio-organizational	X	X	
	Leadership & Coordination		X	X
	Fiscal & Administrative		X	X
	Regulatory & Policy Issues		X	X
	Lessons Not Learned	X	X	X

Fig. 21.2 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)

Among the many challenges 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 study of 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 conversant 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 healthcare 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 [13]. 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 and 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. Through CRI advances, discovery, quality improvement, and the systematic generation of evidence will become as routine and expected a part of the healthcare system and practice in the years to come as advances in clinical informatics in years past have helped foster the systematic application of evidence into healthcare practice.

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 healthcare, 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 need to be developed to achieve this, and this innovation is what will drive the field forward in the coming years. It is a great 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. 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.
3. 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.
4. 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. Medinfo. 2007;12(Pt 1):330–4. PMID: 17911733.
5. 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.
6. Zerhouni EA. Translational and clinical science – time for a new vision. N Engl J Med. 2005;353(15):1621–3.
7. Zerhouni EA. Clinical research at a crossroads: the NIH roadmap. J Investig Med. 2006;54(4):171–3.
8. Payne PR, Johnson SB, Starren JB, Tilson HH, Dowdy D. Breaking the translational barriers: the value of integrating biomedical informatics and translational research. J Investig Med. 2005;53(4):192–200.
9. Sung NS, Crowley Jr WF, Genel M, et al. Central challenges facing the national clinical research enterprise. JAMA. 2003;289(10):1278–87.

10. Friedman CP, Wong AK, Blumenthal D. Achieving a nationwide learning health system. Sci Transl Med. 2010;2(57):57cm29.
11. 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.
12. Chung TK, Kukafka R, Johnson SB. Reengineering clinical research with informatics. J Investig Med. 2006;54(6):327–33.
13. 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 July 2011.

Index

R.L. Richesson, J.E. Andrews (eds.), *Clinical Research Informatics*, Health Informatics, DOI 10.1007/978-1-84882-448-5, © Springer-Verlag London Limited 2012