Health Informatics

Philip R.O. Payne
Peter J. Embi *Editors*

Translational Informatics

Realizing the Promise of Knowledge-Driven Healthcare

Springer

Health Informatics

Philip R.O. Payne • Peter J. Embi
Editors

Translational Informatics

Realizing the Promise of Knowledge-Driven Healthcare

Springer

Editors
Philip R.O. Payne
Biomedical Informatics
The Ohio State University
Columbus, OH
USA

Peter J. Embi
Biomedical Informatics
The Ohio State University Medical Center
Columbus, OH
USA

ISBN 978-1-4471-4645-2 ISBN 978-1-4471-4646-9 (eBook)
DOI 10.1007/978-1-4471-4646-9
Springer London Heidelberg New York Dordrecht

Library of Congress Control Number: 2014948186

Printed on acid-free paper

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

To our loved ones. Thanks for all the support.

Contents

Contributors

Peter J. Embi, MD, MS, FACP, FACMI Departments of Biomedical Informatics and Internal Medicine, The Ohio State University, Columbus, OH, USA

Philip R.O. Payne, PhD, FACMI Department of Biomedical Informatics, The Ohio State University Wexner Medical Center, Columbus, OH, USA

Indra Neil Sarkar, PhD, MLIS Center for Clinical and Translational Science, University of Vermont, Burlington, VT, USA

Department of Microbiology and Molecular Genetics, University of Vermont, Burlington, VT, USA

Nigam Shah, MBBS, PhD Department of Medicine (Biomedical Informatics), Stanford University, Stanford, CA, USA

Jessica D. Tenenbaum, PhD Duke Translational Medicine Institute, Duke University School of Medicine, Durham, NC, USA

Adam B. Wilcox, PhD Department of Medical Informatics, Intermountain Healthcare, Murray, UT, USA

Part I
The Rationale for Translational Informatics

Chapter 1
An Introduction to Translational Informatics and the Future of Knowledge-Driven Healthcare

Philip R.O. Payne and Peter J. Embi

By the End of This Chapter, Readers Should Be Able to:

- Define grand challenges and opportunities surrounding emerging trends in biomedical research and healthcare delivery, with a focus on the data, information, and knowledge needed to achieve the vision of such a healthcare model;
- Understand important trends related to the need for and potential barriers to creating and sustaining a learning healthcare system in which systems-thinking and precision medicine become normative approaches to both research and care delivery; and
- Discuss future directions for the field of translational informatics with a particular emphasis on the technology, cultural, and policy issues that must be addressed to realize the promise of knowledge-drive healthcare.

1.1 Introduction

The field of biomedicine has undergone, and continues to undergo, massive and some might argue tectonic changes, particularly over the past decade. At the core of these changes is a confluence of trends related to the ways in which biomedical

P.R.O. Payne, PhD, FACMI (✉)
Department of Biomedical Informatics, The Ohio State University Wexner Medical Center, Columbus, OH, USA
e-mail: philip.payne@osumc.edu

P.J. Embi, MD, MS, FACP, FACMI
Departments of Biomedical Informatics and Internal Medicine, The Ohio State University, Columbus, OH, USA
e-mail: peter.embi@osumc.edu

P.R.O. Payne, P.J. Embi (eds.), *Translational Informatics: Realizing the Promise of Knowledge-Driven Healthcare*, Health Informatics,
DOI 10.1007/978-1-4471-4646-9_1,

education, research, and clinical care organizations fund, staff, and operate their enterprises. Such factors are called into sharp relief by simultaneous pressures exerted at the economic, governmental, and policy levels. When taken as a whole, this environment and the changes it has experienced can be defined by a number of critical challenges and opportunities, as enumerated below:

- *How can we transform the delivery of both clinical care and wellness promotion, such that the quality, safety, and efficacy of such efforts are improved, while simultaneously decreasing the costs and complexity of doing so?*
- *How do we accelerate the speed with which discoveries in the basic sciences are translated into actionable and widely utilized diagnostic and/or therapeutic strategies?*
- *How can we drive the process of discovery as a natural outgrowth of patient care by leveraging data that is collected through clinical interactions?*
- *How will a biomedical workforce, with appropriate levels and types of training and related career trajectories, both evolve and be sustained in a manner aligned with such rapidly changing needs?*

A shift in emphasis and thinking, relative to the ways in which the broad healthcare community addresses the collection, storage, management, analysis, and dissemination of data, information, and knowledge, will be required to address these questions. Such a shift will be impossible to achieve without significant cultural change. It must de-emphasize application or domain-specific silos in favor of integrative and systems-level translation between and among disciplines and driving biological or clinical problems. For readers not familiar with the premise of such systems-level approached, we will define the concept more fully in Sect. 2.2. When taken as a whole, we have labeled this paradigm as "Translational Informatics" or "TI", and will argue that the definition and application of TI principles is essential to the realization of a knowledge-driven healthcare enterprise capable of addressing the aforementioned challenges and opportunities. Building upon this overall motivation, and in order to contextualize the remainder of this book, in this chapter, we will:

- *Describe in greater detail the motivating factors for the formulation of the TI vision, including the promise of translation, an emergent shift in scientific investigation away from reductionism and towards systems thinking, and the evolving central dogma of the discipline of Biomedical Informatics*;
- *Introduce exemplary trends that serve to illustrate the importance of the preceding factors, including an increasing emphasis on the creation of learning healthcare systems, the evolution of precision medicine, and the dawning era of "big data"*;
- *Propose a set of next steps related to the advancement of critical strategic research foci, implementation science best practice, and workforce development, in response to such motivating trends*; *and*
- *Introduce a hypothetical environment in which all of the preceding factors are present, spanning a spectrum from patients to policy makers. This environment*

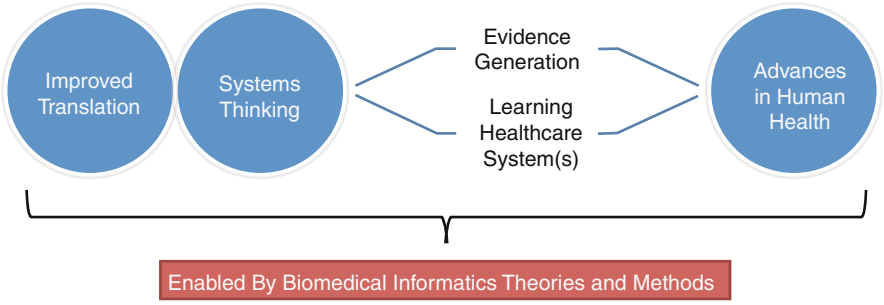

Fig. 1.1 Foundational argument for the vision of TI, wherein improved translational capabilities, combined with systems thinking, and enabled by Biomedical Informatics theories and methods, can facilitate critical advances in human health as provided by knowledge-driven paradigms such as rapid-cycle evidence-generation predicated on the existence of learning healthcare system(s)

will be used throughout the remainder of book to contextualize major areas of innovation that contribute to attainment of a TI model and the vision for knowledge-driven healthcare.

When taken as a whole, we believe that the foundational argument for the vision of TI is that improved translation and systems thinking, enabled by Biomedical Informatics theories and methods, will yield a platform and "way forward" towards critical advances in human health made possible by fully knowledge-driven workflows and practices (Fig. 1.1).

1.2 Motivation for a Translational Informatics Vision

In the sections that follow, we will review three critical and motivating factors underlying our vision for Translational Informatics (TI), namely: (1) the promise of new models for clinical and translational research (collectively referred to as translational science and described in more detail in Sect. 2.1), particularly as applied to biomedicine; (2) an emergent trend away from reductionism and towards systems thinking; and (3) the formalization of a central dogma for the broad domain of Biomedical Informatics.

1.2.1 The Promise of Translational Science

As was introduced at the outset of this chapter, increasingly, the biomedical and healthcare communities have experienced a shift away from narrowly focused and individualized research programs towards a model that emphasizes team-based

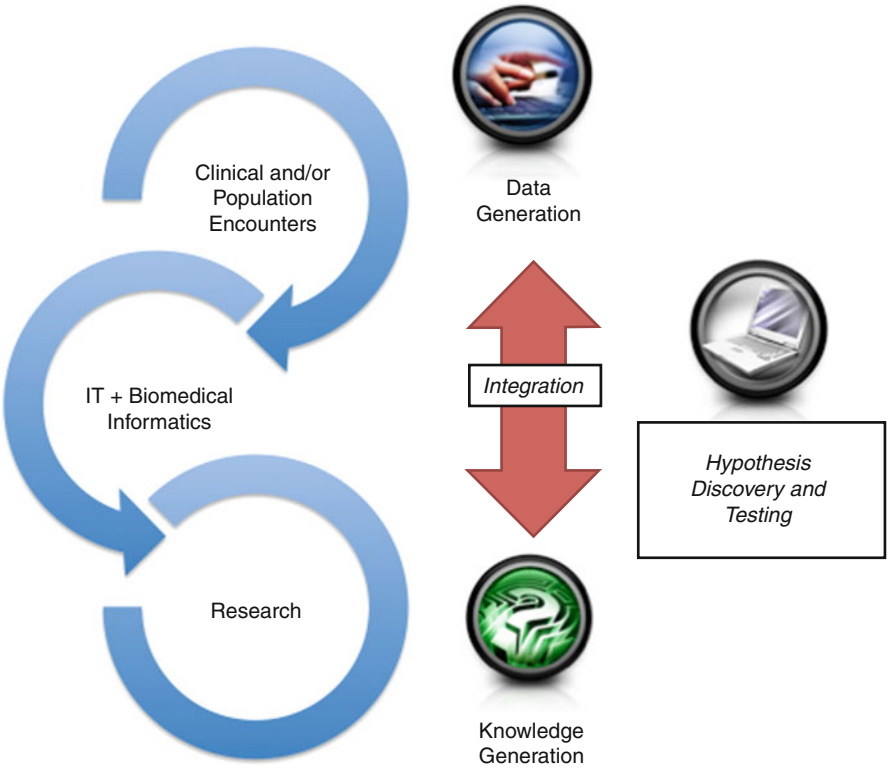

Fig. 1.2 Overview of the "Translational Science" paradigm, in which an integrative and cyclical approach to knowledge discovery, research, and evidence generation seeks to accelerate the translation of new scientific discoveries into clinical and/or population-based healthcare practice

approaches to complex problems that often span traditional organizational boundaries. Such a change in thinking and research practice has been described in many venues as "Translational Science", and has been broadly defined (and redefined) in a variety of reports and books [1–3]. Another critical dimension of this shift in thinking has been towards a model in which research and clinical care are tightly and iteratively connected, whereby data generation and evidence generation spanning these two complementary areas are synergistic, integrated, and highly efficient. Ultimately, such an approach to knowledge discovery, research, and evidence-generation are intended to overcome what has been described as a highly inefficient or even dysfunctional paradigm in which new basic science discoveries can take up to two decades to be translated into broadly utilized clinical care or population health practices. An illustration of the integrative and cyclical nature of the "Translational Science" paradigm is provide in Fig. 1.2 and described below:

- *In the "Translational Science" paradigm, a variety of* ***data, information, and knowledge resources are generated*** **via** ***either clinical or population-level***

***encounters**, wherein patient- or cohort-level features are instrumented, codified, and represented in ways that support and enable their reuse* [1–4];

- *In a similar manner and as part of the "Translational Science" paradigm, **basic science research is conducted in a manner that emphasizes the analysis of the data generated in the lab as contextualized by driving clinical problems** and related data sets derived from the preceding facets of the model as are concerned with instrumenting the healthcare delivery environment to support research endeavors* [5–11];
- *Spanning the two preceding areas is the combined use of information technology (IT) and Informatics theories and methods, such that **data, information, and knowledge generation are enabled in an efficient and harmonized manner*** [2, 4, 6]; *and*
- *Finally, employing such an integrative view of basic science, clinical, and population-level data, information, and knowledge resources, **investigative teams are able to both discover and test high-impact hypotheses that can contribute to the overall biomedical and healthcare delivery knowledge base, and quickly deliver such knowledge in terms of standard-of-care practice guidelines and interventions*** [2, 4, 6].

While the "Translational Science" approach we have described is indeed difficult to achieve given current technical, cultural, and policy-based constraints, the ability to overcome such barriers opens a pathway towards a number of promising benefits, including both:

- *The ability to **break-down conventional barriers** between critical components of the research process, such that the "hand-off" of data, information, and knowledge between such activities becomes timely and efficient, and perhaps more importantly, an expected and valued aspect of such efforts*; *and*
- *The facilitation of **rapid cycling and recycling of data, information, and knowledge** between complementary scientific disciplines, thus creating a systematic whole that is greater than the sum of its constituent parts*;

Ideally, the combination of these benefits can lead to the timelier, resource- efficient, and impactful delivery of scientific evidence at the point-of-care, thus improving the health of patients and populations in any number of critical areas from disease prevention and control to the diagnosis and treatment of complex and heretofore unaddressed pathophysiological states.

1.2.2 Systems Thinking in Biomedicine

In a manner analogous to the shift towards a translational science paradigm as previously introduced, there is also a shift occurring relative to the fundamental thought processes used to conceptualize and execute both research and care in the biomedical and health sciences domains. Traditional models of thinking, tracing their origins

back to the first discovery of cellular-level phenomena, have emphasized a reductionist approach to science [12]. In such a reductionist approach, large and complex problems are broken down into small, manageable units for investigation and inquiry. The size and scope of such units have historically been dictated by a combination of inherent human cognitive limitations and the capabilities of available instruments and data management methods, the latter spanning a spectrum from paper to early computers to modern cloud based technologies. The belief system surrounding reductionist thinking has been and continues to be predicated on the idea that if we can understand the structure, function, or other features of interest of a given unit of investigation, we can then reassemble the knowledge gained from such investigation with similar understanding of complementary or co-occurring units within broader settings (such as biological or organ systems, or at a higher order, organizations and populations). In effect, the reductionist viewpoint is that knowledge of a system can be built from "building blocks" of knowledge concerning the sub-units of that system, as studied in isolation. Such a mindset is quite pervasive in the biomedical and broader scientific communities, dictating aspects of those fields including the ways in which we describe and label various sub-disciplines from an educational and professional standpoint, to the ways in which we organized publication venues and funding programs (e.g., in a manner aligned with organ, disease, or higher-order systems or foci, decomposed from broader systems such as human beings or populations). However, recent scientific endeavors have begun to elucidate a number of critical flaws in this type of reductionist thinking, namely [12]:

- *The elemental units that may make up a complex biological system rarely operate in isolation, and instead, are highly interrelated from a structural and functional standpoint with any number of other entities making up the broader whole. As a result, by studying such units in isolation, the likely outcome is that we will: (1) not fully understand the phenomena of interest that can characterize that unit; and (2) not understand or measure the important interrelationships and dependencies between and across units, thus limiting the ability to reassemble unit-level knowledge into an understanding of the greater system*;
- *Given emerging evidence that many if not all biological systems behave in a similar manner when evaluated as networks of interacting entities and processes (what is referred to in the scientific community as "scale free network theory" – an explained in greater detail by Barabasi and colleagues* [6]*), it can be concluded that the most important targets for the disruption or manipulation of those systems are the nodes or components that are the most highly interrelated with other nodes or components (i.e., having a high degree of "nodality" or in more lay-level terms, serving as "hubs" between individual or groups of nodes). However, by not studying the interrelationships and/or dependencies between the units that comprise systems of interest, the ability to identify such high-value targets, which could inform diagnostic and/or therapeutic strategies, is significantly diminished*; *and*
- *The building body of knowledge generated by the systems biology and medicine communities (e.g., scientific communities who are applying the preceding*

systems thinking principles to complex biological or clinical problems) is beginning to demonstrate that by taking a systems level approach, which is now possible given new and extremely high-capacity instrumentation and data management tools not previously available, we can in fact study complex systems as a whole. These systems-approaches allow for the realization of benefits related to network-level analyses of the interactions between important entities in a biological or disease system and their role in high impact end points such as identifying new uses of existing drugs, identifying important markers for risk of disease; or developing novel therapeutic strategies for a broad spectrum of pathophysiological states.

Thus, when examined in a similar systems level perspective, prevailing approaches to the pursuit of biomedical and healthcare research are showing signs of moving from a historically motivated tendency for reductionism, towards a systems thinking model, with all of the aforementioned and resultant potential benefits.

1.2.3 Towards a Central Dogma for Biomedical Informatics

Finally, in a manner that is crosscutting and underlies the role of Biomedical Informatics as it pertains to both translational science and systems thinking, the increasing maturity of the field is leading to the recognition of a "working" central dogma for Biomedical Informatics. In this "working" definition of the broad purpose for Biomedical Informatics as a scientific field, core theories and methods that span the discipline that can collectively be seen as contributing to the translation of raw data into information through the provision of context and subsequently the translation of such information into knowledge by rendering it in a manner that is actionable. For example, given a clinical data point such as a laboratory value, through the addition of metadata (e.g., context) via the use of technical standards and knowledge engineering methods, we are able to transform that data into information. Subsequently, by representing and communicating that information to a clinical decision support system that has been designed and validated using decision modeling and analysis methods and frameworks, we can render it actionable as knowledge for clinical decision making at the point-of-care. As another example, given a transcriptome sequencing dataset, we can normalize and structurally and/or functionally annotate up- or down-regulated genes using available knowledge bases and deep semantic reasoning methods so as to create an information resource that characterizes a given sample. We could then apply advanced data visualization and interactive data analytic frameworks and methods to deliver a graphical representation of such an information resource. Investigators can then identify and refine patterns or motifs of interest relative to their experimental paradigm, thus rendering the underlying information actionable as a knowledge resource. This overall working "central dogma" is illustrated in Fig. 1.3.

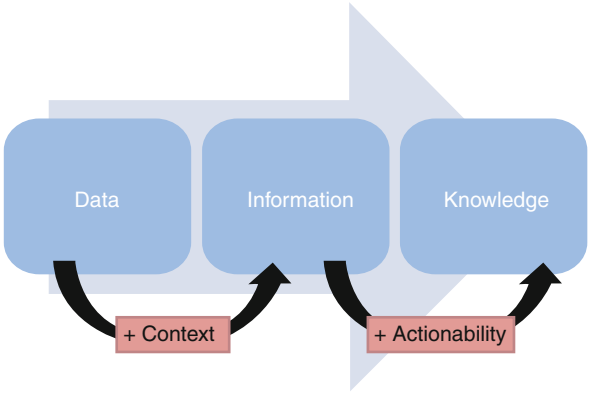

Fig. 1.3 A diagrammatic representation of an emerging and "working" central dogma for Biomedical Informatics, in which theories and methods collectively contribute to the generation of information from data through the addition of context, and subsequently to the generation of knowledge via the delivery of such information in an actionable format/mechanism

1.3 Emerging Trends and Their Implications for Translational Informatics

Building upon the three motivating factors introduced in the preceding sections, we will now go on to discuss three emergent trends that serve to exemplify and illustrate the challenges and opportunities afforded by TI, specifically: (1) an increasing emphasis on the creation and operation of learning healthcare systems; (2) the evolution of precision medicine as a means of improving wellness promotion as well as clinical care; and (3) the expanding role and impact of "big data" in biomedicine.

1.3.1 Learning Healthcare Systems

Increasingly, at the local, regional, national, and international levels, emphasis is being place on the creation of what are being referred to as "Learning Healthcare Systems" (LHCs). Such LHCs are characterize by a number of dimensions [13, 14], including:

- *The instrumentation of clinical care activities such that data that is useful for both care delivery and the investigation of phenomena of interest that could yield new evidence concerning health and wellness, is collected and made accessible to all members of healthcare delivery and research teams. Of note, a major aspect of this particular dimension involves the engineering and/or re-engineering of Electronic Health Record (EHR) systems to support the systematic capture, extraction, and reporting of high-value structured, semi-structured, and un-structured data that can support and enable primary (clinical) and secondary (research) use cases*;
- *The execution of pragmatic research programs, in which large numbers of patients are engaged in the investigation of clinically-relevant hypotheses via their participation in minimally invasive registry or outcomes research programs*

that leverage data sets being created during the course of standard-of-care activities;
- *The rapid and cyclical feedback of observations, findings, and evidence between clinical care and research teams, utilizing the aforementioned capabilities*; *and*
- *The rationalization of regulatory frameworks and policies so as to maximize the balance between clinical, research, and patient privacy/confidentiality needs in an equitable, transparent, resource efficient, and timely manner.*

When viewed collectively, these dimensions that serve to characterize LHCs are predicated on an emerging model in which research and clinical care are both inexorably and desirably intertwined, and where barriers between such activities are mitigated or removed. Further, in the LHC construct, there is a fundamental transition away from a unidirectional model in which research informs practice (evidence based practice or EBM), and towards a bi-directional or cyclical model in which practice informs research that in turn informs practice (which can be described as evidence generating medicine or EGM). This new approach can be thought of as enabling a model in which every patient encounter is an opportunity to improve the care of that patient, their family, and their community, by bringing together health and wellness care with the best possible and appropriately contextualized science, and by recognizing and valuing the role of clinicians, researchers, patients, and their communities as equal partners in such an endeavor.

1.3.2 The Evolution of Precision or Personalized Medicine

The objective of precision medicine is to ensure that each patient has the best clinical outcome by tailoring both preventative measures and treatments to meet his or her unique needs and characteristics. Achieving such a vision requires not only the collection and application of the best possible data, information, and knowledge during each patient encounter, but also, learning from each encounter and engaging patients and their families in the healthcare process, as has been described previously in the context of the emerging LHC model. An innovative and paradigm-shifting approach to conceptualizing precision medicine has been described by Weston and Hood using the moniker of "P4 Medicine" – where it was proposed that our fundamental approach to disease prevention, diagnosis, and treatment must transition from being a primarily reactive model to one that is predictive, personalized, preventive and participatory [11]. In this model, it is envisioned that our fundamental approach to the delivery of healthcare will be shifted from an emphasis on treating illness to the early and continuous prevention of disease and the promotion of wellness. Furthermore, under this paradigm, the patient becomes an integral part of the healthcare delivery ecosystem, taking an active role in the identification and modification of disease-related risk factors, while also assuming responsibility for critical aspects of their ongoing care (moving from being a passive consumer of clinical care to an active member of the overall healthcare team). Unfortunately, it is widely noted that the current healthcare delivery workflows (including essential

data, information, and knowledge management methods) are not well aligned with the P4 paradigm, thus impeding the implementation of the model [3–5, 15, 16].

As can be readily ascertained, the continuum of data, information, and knowledge management that is central to the premises underlying P4 medicine is directly aligned with the emerging and "working" central dogma for Biomedical Informatics that we have previously described. Unfortunately, current approaches to basic science research, clinical care, and biomedical informatics are often poorly integrated, yielding clinical decision-making processes that do not take advantage of up-to-date scientific knowledge and capabilities afforded by Biomedical Informatics theories and method [5, 12, 15]. There are an increasing number of systems modelling and in-silico knowledge synthesis techniques that can provide investigators with the tools to address such information needs, but their adoption and evaluation remains an area of early and open research [3, 4, 8, 12, 16]. Given increasing concerns over barriers to translating discoveries from the laboratory to the clinic or community, such high-throughput informatics methods are highly desirable, and in our opinion, central to the P4 paradigm [4, 6–8, 12]. As such, the on-going evolution in precision or "P4" medicine has been and continues to be focused on overcoming fundamental barriers that serve to prevent or impede rapid and systematic translation between research and clinical care. Such barriers include a lack of unification between data generation environments, as regularly occurs in the laboratory, clinical, and community settings, and knowledge generation, which is the fundamental pursuit of research. The lack of unification is attributable to a number of factors as introduced previously, including innate technical limitations to current clinical decisions support systems, socio-technical and regulatory barriers, as well as a lack of sufficiently robust and widely adopted informatics platforms intended to "shorten the distance" between data and knowledge generation [3, 4]. Such a state of affairs is both promising, in terms of the potential benefits of precision or "P4" medicine that can be enabled through the use of Biomedical Informatics theories and methods, and also challenging, given a landscape that remains somewhat misaligned with this vision for the future of knowledge-driven healthcare.

1.3.3 The Role of "Big Data" in Biomedicine

Finally, and of note, there is an expanding focus in a variety of technical and information-intensive domains on what has been called "Big Data." In contemporary discussions of trends in "Big Data", it has been argued that the definitional characteristics of a data set that is "Big" can be summarized via the three "Vs" [17, 18], namely:

- ***Volume****: the data set is of sufficiently large in scale or volume that it requires specialized collection, storage, transaction, and/or analysis methods*;
- ***Velocity****: the speed with which the data is generated is such that it requires specialized collection, storage, or transaction technologies*; *and*

- ***Variability****: the syntactic and/or semantic nature of the data is highly variable, requiring specialized knowledge management and engineering approaches in order to support the analysis of such resource*

It is widely held that a data set or resource demonstrating any one or more of the aforementioned characteristics is "Big Data." As can readily be concluded, any number of data types commonly encountered in the modern biomedical environment can be classified as "Big Data", such as patient-derived phenotypes extracted from EHRs, sensor data used to understand patient or population characteristics outside of the clinical care environment, and the data resulting from modern bio-molecular instrumentation such as that associated with exome or whole genome sequencing.

The primary challenges that have prompted the definition of what is (and is not) "Big Data", and that have catalyzed such widespread interest in "Big Data" analytics, include [17]:

- *The absence of well accepted and/or validated tools and methods capable of reliably and efficiently supporting or enabling the collection, storage, transaction, and analysis of "Big Data" in a timely and cost-effective manner (e.g., not requiring specialized, costly, and time-intensive computational tools and approaches)*;
- *The minimal understanding of how common quantitative science measurements, such as conventional statistical significance testing, scale to indicate and quantify patterns or phenomena of interest in "Big Data" constructs, especially for those that exhibit two or more of the "3Vs" and thus could include considerable sparsity or "noise"*; *and*
- *The innate challenges of delivering "Big Data" to human end-users in a manner that is comprehensible and capable of leveraging innate cognitive capabilities relative to higher-order pattern recognition and semantic reasoning.*

Because of these, and many other compelling computational, quantitative science, and Biomedical Informatics challenges associated with "Big Data", this area has emerged as a rapidly growing and dynamic area of research and investigation, likely to greatly influence and contribute to the overall TI vision introduced here.

1.4 A Path Forward for Translational Informatics and Knowledge-Based Healthcare

As can be ascertained by the preceding survey of the current state of biomedical knowledge and practice, and the major factors and trends that we have emphasized, it can be argued that the realization of a TI vision is beset by significant challenges and opportunities. As such, we will contend in subsequent chapters of this book that there are three major areas that must be addressed in order to advance the TI knowledge-base from both a basic and applied perspective, namely: (1) the pursuit of strategic, TI-relevant strategic and research foci; (2) an increased emphasis on the

tight coupling of implementation science and technology-based intervention strategies; and (3) enhanced and expanded workforce development relative to the creation of a TI-focused community consisting of both innovators and practitioners. We will explore each of these areas in the following sub-sections:

1.4.1 Strategic and Operational Foci

In order to advance and sustain both the knowledge base and downstream practices incumbent to the TI vision, it will be essential for the health and life science communities, as well as the broader community of interested parties, to identify and engage in the systematic pursuit of core strategic and operational foci. It is only though collective effort and the combinatorial effect of the information and knowledge gained from such endeavors that we can realize the potential benefits of the TI vision and a knowledge-driven approach to healthcare. These focus areas include:

- *The support and pursuit of scientific programs that combine Biomedical Informatics and driving biological or clinical problems in order to achieve translational science end-points*;
- *The re-alignment of scientific and applications-level policies and cultural norms with a systems-thinking approach to decision making, hypothesis generation/testing, research funding, care delivery, and career development for individuals pursuing such efforts*;
- *Continued development and validation of core Biomedical Informatics theories and methods that can contribute to filling in gaps in knowledge and practice pertaining to the aforementioned "working" central dogma for the field*;
- *The creation and demonstration of "real world" LHCs that can positively impact the quality, safety, efficiency, and outcomes of clinical care, as informed by virtuous cycles of evidence generation and practice*;
- *The refinement of core theories, methods, and models by which precision medicine paradigms can be used to ensure that clinical care is informed by the best possible science*; *and*
- *The ongoing development and utilization of "Big Data" theories and methods that will enable a broad spectrum of individuals to ask and answer meaningful questions in a high throughput manner relative to all types of data exhibiting one or more of the "3Vs"*

1.4.2 The Role of Implementation Science

At the same time that the preceding strategic and research foci are pursued, it is of great importance that a parallel and complementary application of implementation science principles be pursued. In this context, we define implementation science as:

> …the study of methods to promote the integration of research findings and evidence into healthcare policy and practice… (Source: National Library of Medicine)

Given such a definition, it is clear that implementation science theory and practice will be central to the ability to understand and optimize the pathways by which knowledge flows among the myriad disciplines and settings involved in the broader TI ecosystem. The use if implementation science theories and methods can and will ultimately assist in the analysis of key socio-cultural and environmental factors that may serve to influence or predispose the way in which the vision of TI, and ultimately knowledge-driven healthcare, is operationalized. However, it is also important to note that implementation science, at least in the biomedical and healthcare domains, is an emergent and relatively "young" discipline, and thus an area defined by a broad spectrum of open research questions. Despite such limitations, if the lessons learned from the intersection of implementation science and other disciplines (such as the social science and public health domains) are taken as exemplars [19], the application of these principles lead to a deeper and richer understanding of the complex "real world" issues that ultimately decide the fate of knowledge translational from research to practice to widespread adoption.

1.4.3 Workforce Development

Finally, and of equal importance to the preceding areas, given the paucity of individuals with training or expertise in both basic and applied Biomedical Informatics, particularly with relevance to the pursuit of TI, it will remain imperative for the biomedical and healthcare communities to pursue and sustain workforce development activities in such domains. The types of stakeholders to be engaged, and the acuity of knowledge relative to Biomedical Informatics competencies needed by such trainees, should be used to appropriately inform such workforce development programs. At a high level, such stakeholders and training acuity levels can be stratified into three major categories:

- **Acculturation**: At this level of training, individuals should become familiar with high level definitions and methodological frameworks such that they are able to prioritize and direct both basic and applied research, development, and systems evaluation activities. Examples of stakeholders at this level of acuity include C-level executives, project managers, and perhaps most importantly, clinicians and other healthcare providers.
- **Applied Knowledge**: At this level of training, individuals should understand how to select and apply theories and frameworks to satisfy use-case specific information needs. Examples of stakeholders at this level of acuity include technical architects, software engineers, and technology deployment/support staff.
- **Foundational Knowledge**: At this level of training, individuals should have a thorough theoretical and methodological grounding, as well as expertise in

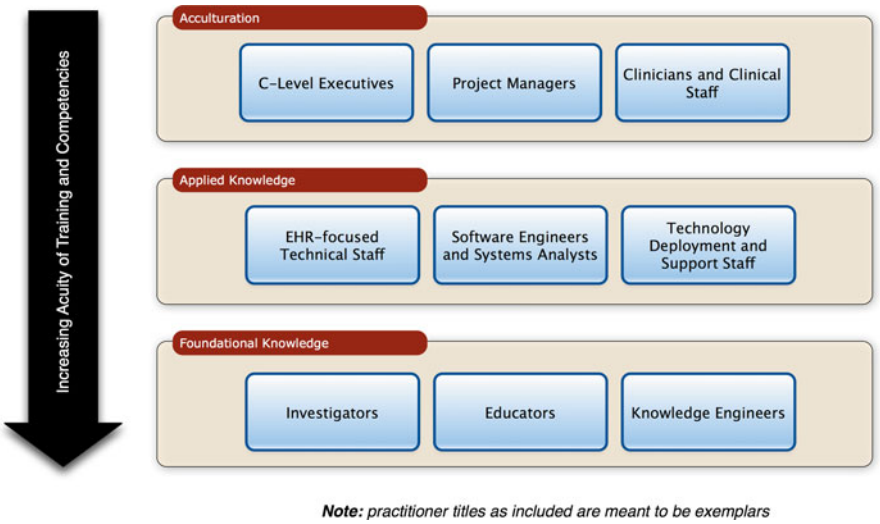

Fig. 1.4 Overview of levels of training acuity in TI relevant Biomedical Informatics theories and methods, including alignment of stakeholder types

scientific investigation and research methods, such that they can conceptualize, design, and evaluate novel theories and methods. Examples of stakeholders at this level of acuity include investigators, educators, and knowledge engineers.

This overall model for workforce development is illustrated in Fig. 1.4.

1.5 Conclusions

As was introduced at the beginning of this chapter, the premise for this book is that a shift in emphasis and thinking, relative to the ways in which the broad biomedical healthcare community addresses needs concerning the collection, storage, management, analysis, and dissemination of data, information, and knowledge, is necessary but difficult to achieve without significant cultural change. This shift should be one that de-emphasizes application or domain-specific silos, and instead, emphasizes integrative and systems-level translational between and among disciplines and driving biological or clinical problems. This emerging paradigm, which we have described as "Translational Informatics" or "TI", is essential to the realization of a knowledge-driven healthcare enterprise capable of addressing the aforementioned challenges and opportunities. Given these basic concepts, throughout the remainder of this book, we review a spectrum of Biomedical Informatics theories, methods, and use cases that serve to inform and exemplify the TI vision on scales from bio-molecules to patients to populations. In doing so, we hope to

provide readers with the basic knowledge and understanding needed to apply TI principles in multiple settings, from policy making to research to clinical practice to population health. Ultimately, such a goal is, in our opinion, critical to achieving the transformation of our healthcare system into one that uses the best possible science to inform and enable a shift from an emphasis on episodic "sick care" to a new and more effective and desirable emphasis on longitudinal wellness and health maintenance.

Discussion Points

- What are motivating and contemporary factors for the formulation TI vision?
- What trends in the broad biomedical and healthcare domains are influencing or otherwise motivating the preceding factors?
- What are the critical strategic research foci, implementation science best practices, and workforce development needs associated with the vision of TI and knowledge-based healthcare?
- How does the TI vision relate to our ability to transform the delivery of both clinical care and wellness promotion, such that the quality, safety, and efficacy of such efforts are improved, while simultaneously decreasing the costs and complexity of doing so?
- How does the TI vision relate to our ability accelerate the speed with which discoveries in the basic sciences are translated into actionable and widely utilized diagnostic and/or therapeutic strategies?

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. PubMed PMID: 19261934. Epub 2009/03/06.
2. Embi PJ, Payne PR. The role of biomedical informatics in facilitating outcomes research: current practice and future directions. Circulation. 2009;120:2393–9.
3. 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. PubMed PMID: 15974245, Epub 2005/06/25.
4. Payne PR, Embi PJ, Sen CK. Translational informatics: enabling high throughput research paradigms. Physiol Genomics. 2009;39(3):131–40.
5. Auffray C, Chen Z, Hood L. Systems medicine: the future of medical genomics and healthcare. Genome Med. 2009;1(1):2.1–2.11.
6. Barabasi AL, Oltvai ZN. Network biology: understanding the cell's functional organization. Nat Rev Genet. 2004;5:101–13.
7. Butcher EC, Berg EL, Kunkel EJ. Systems biology in drug discovery. Nat Biotechnol. 2004;22(10):1253–9.
8. Hood L, Perlmutter RM. The impact of systems approaches on biological problems in drug discovery. Nat Biotechnol. 2004;22(10):1215–7.
9. Lussier YL, Chen JL. The emergence of genome-based drug repositioning. Sci Transl Med. 2011;3(96):1–3.

10. Sadimin ET, Foran DJ. Pathology imaging informatics for clinical practice and investigative and translational research. N Am J Med Sci. 2012;5(2):103–9. PubMed PMID: 22855694. Pubmed Central PMCID: 3407842.
11. Weston AD, Hood L. Systems biology, proteomics, and the future of health care: toward predictive, preventative, and personalized medicine. J Proteome Res. 2004;3(2):179–96.
12. Ahn AC, Tewari M, Poon CS, Phillips RS. The limits of reductionism in medicine: could systems biology offer an alternative? PLoS Med. 2006;3(6):709–13.
13. Embi PJ, Payne PR. Evidence generating medicine: redefining the research-practice relationship to complete the evidence cycle. Med Care. 2013;51:S87–91.
14. Friedman CP, Wong AK, Blumenthal D. Achieving a nationwide learning health system. Sci Transl Med. 2010;2(57):57cm29.
15. Hood L, Friend SH. Predictive, personalized, preventive, participatory (P4) cancer medicine. Nat Rev Clin Oncol. 2010;8:184–7.
16. Schadt EE, Bjorkegren JL. Network-enabled wisdom in biology, medicine, and health care. Sci Transl Med. 2012;4(115).
17. Jacobs A. The pathologies of big data. Commun ACM. 2009;52(8):36–44.
18. Lynch C. Big data: how do your data grow? Nature. 2008;455(7209):28–9.
19. Damschroder LJ, Aron DC, Keith RE, Kirsh SR, Alexander JA, Lowery JC. Fostering implementation of health services research findings into practice: a consolidated framework for advancing implementation science. Implement Sci. 2009;4(1):50.

Additional Reading

Ahn AC, Tewari M, Poon CS, Phillips RS. The limits of reductionism in medicine: could systems biology offer an alternative? PLoS Med. 2006;3(6):709–13.

Ash JS, Anderson NR, Tarczy-Hornoch P. People and organizational issues in research systems implementation. J Am Med Inform Assoc. 2008. PubMed PMID: 18308986.

Auffray C, Chen Z, Hood L. Systems medicine: the future of medical genomics and healthcare. Genome Med. 2009;1(1):2.1–2.11.

Barabasi AL, Oltvai ZN. Network biology: understanding the cell's functional organization. Nat Rev Genet. 2004;5:101–13.

Cantor MN. Translational informatics: an industry perspective. J Am Med Inform Assoc. 2012;19(2):153–5. PubMed PMID: 22237867. Pubmed Central PMCID: 3277629.

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. PubMed PMID: 19261934. Epub 2009/03/06.

Embi PJ, Payne PR. The role of biomedical informatics in facilitating outcomes research: current practice and future directions. Circulation. 2009;120:2393–9.

Hood L, Friend SH. Predictive, personalized, preventive, participatory (P4) cancer medicine. Nat Rev Clin Oncol. 2010;8:184–7.

Kohane IS, Churchill SE, Murphy SN. A translational engine at the national scale: informatics for integrating biology and the bedside. J Am Med Inform Assoc. 2012;19(2):181–5. PubMed PMID: 22081225, Pubmed Central PMCID: 3277623.

Lussier YL, Chen JL. The emergence of genome-based drug repositioning. Sci Transl Med. 2011;3(96):1–3.

Payne PR, Embi PJ, Sen CK. Translational informatics: enabling high throughput research paradigms. Physiol Genomics. 2009;39:131–40.

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. PubMed PMID: 15974245. Epub 2005/06/25. eng.

Payne PR, Pressler TR, Sarkar IN, Lussier Y. People, organizational, and leadership factors impacting informatics support for clinical and translational research. BMC Med Inform Decis Mak. 2013;13:20. PubMed PMID: 23388243. Pubmed Central PMCID: 3577661.

Searl MM, Borgi L, Chemali Z. It is time to talk about people: a human-centered healthcare system. Health Res Policy Syst. 2010;8(35).

Sittig DF, Singh H. A new socio-technical model for studying health information technology in complex adaptive healthcare systems. Qual Saf Health Care. 2011;19 Suppl 3:i68–74.

Weston AD, Hood L. Systems biology, proteomics, and the future of health care: toward predictive, preventative, and personalized medicine. J Proteome Res. 2004;3(2):179–96.

Yip YL. Unlocking the potential of electronic health records for translational research. Findings from the section on bioinformatics and translational informatics. Yearb Med Informa. 2012;7(1):135–8. PubMed PMID: 22890355.

Chapter 2
A Prototype of Translational Informatics in Action

Philip R.O. Payne

By the End of this Chapter, Readers Should Be Able To

- Understand the multiple types and levels of stakeholders, and their respective activities that can and should be impacted by the Translational Informatics paradigm;
- Describe an exemplary clinical scenario that illustrates the scope and impact of Translational Informatics on the delivery of knowledge driven healthcare; and
- Apply the preceding context to subsequent chapters that explore various dimensions of the tight coupling of Biomedical Informatics and the domains of biomedical research and clinical care delivery in order to achieve the vision of Translational Informatics.

2.1 Introduction

As was presented in Chap. 1, the vision for Translational Informatics (TI) is predicated on three critical and synergistic dimensions, namely:

1. The promise of translational science, particularly as applied to biomedicine [1, 2];
2. An emergent trend away from reductionism and towards systems thinking [3–5]; and
3. The formalization of a central dogma for the broad domain of Biomedical Informatics.

These dimensions have broad ranging and significant impacts on a variety of actors and their activities that serve to make up the biomedical research and

P.R.O. Payne, PhD, FACMI
Department of Biomedical Informatics,
The Ohio State University Wexner Medical Center, Columbus, OH, USA
e-mail: philip.payne@osumc.edu

P.R.O. Payne, P.J. Embi (eds.), *Translational Informatics: Realizing the Promise of Knowledge-Driven Healthcare*, Health Informatics,
DOI 10.1007/978-1-4471-4646-9_2,

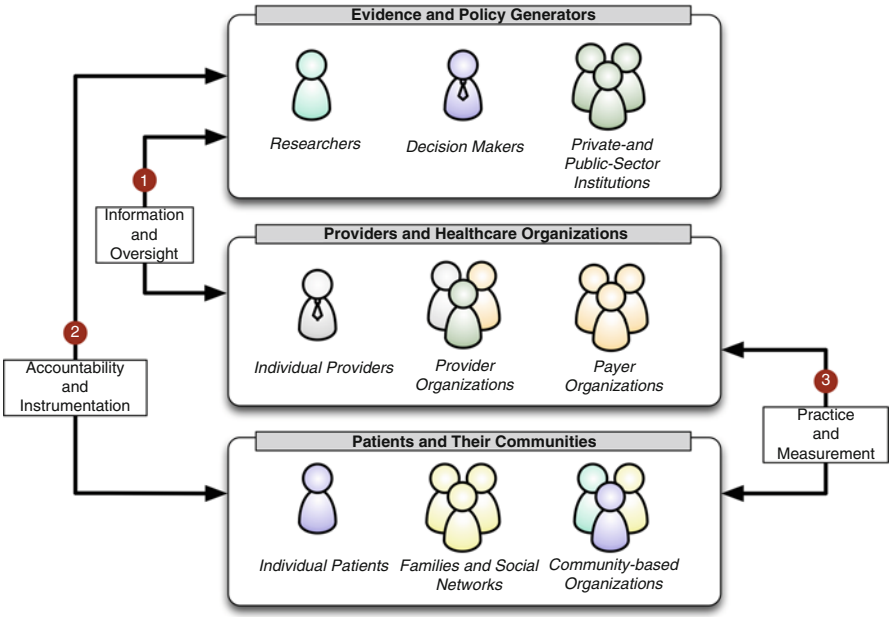

Fig. 2.1 Conceptual model for the "real world" entities that are influenced and/or affected by the TI and knowledge-driven healthcare paradigms

healthcare delivery environments. At a high level, we can classify such actors as belonging to one or more of the following categories:

- ***Evidence and policy generators****, who direct, support and pursue the creation of scientific evidence and frameworks that can catalyze and sustain a TI-centric environment (*e.g., *government agencies, academic health centers, and individual researchers and their laboratories);*
- ***Providers and healthcare organizations****, who fund clinical care, deliver it to patients and populations, and measure its impact (*e.g., *integrated delivery networks, hospitals, clinical practices, third-party payers, and individual physicians or other healthcare professionals); and*
- ***Patients and their communities****, who are provided with, support or consume healthcare knowledge and services (*e.g., *patients, their families, community organizations of like-minded individuals, and other social support structures)*

In this categorization, there exist critical relationships (labeled as [1] in Fig. 2.1) between evidence and policy generators, and providers and healthcare organizations, that serve to set a scientific and operational framework in which knowledge-driven healthcare can and is delivered. Similarly, a there are relationships (labeled as [2] in Fig. 2.1) between those same evidence and policy generators, and patients and their surrounding communities that provide both a means of accountability for the aforementioned framework, as well as for the ability to

instrument large-scale and population level "signals" that inform research and policy making activities. Finally, there are relationships (labeled as [3] in Fig. 2.1) between providers and healthcare organizations, and patients and their surrounding communities that provide both the basis for the practice of knowledge driven healthcare as well as the determination of process and outcomes-oriented measurements of such activities, which can again inform research and policy making activities. This overall model is illustrated in Fig. 2.1 and will be described in further detail throughout this chapter.

It is extremely important to note that while such a categorization and relationship schema would appear to indicate concrete and discernable barriers between such roles and activities, in "real world" settings, such intersection points are often quite "fuzzy", with individuals and organizations fulfilling multiple and simultaneous roles. Thus, readers should be aware of this complexity when attempting to apply the described model to assess and evaluate "real world" problem domains.

2.2 A Prototype of Translational Informatics in Action

Our example case of TI in action will focus on the most common type of colon cancer, known as colon adenocarcinoma [6]. This is a diagnosis that affects 140,000 people per year in the United States according to the National Cancer Institute (NCI), with a incidence of approximately 40 cases for every 100,000 people [6]. People age 50 years and older are at the highest risk for this type of colon cancer. In some but not all cases, colon adenocarcinoma can be aggressive, metastasizing from the primary site in the colon to other parts of the body, including organ systems as well as the lymphatic system. While highly treatable (with more than 50 % of patients surviving for 5 years or more after therapy), the likelihood of a positive clinical outcome post-treatment increases dramatically if the cancer is detected early (with long term survival rates of 80 % or higher given early diagnosis and treatment). One of the challenges relative to such early detection and treatment is that colon adenocarcinoma grows slowly at first, and patients can often remain symptom free for up to 5 years. Symptoms that lead to the diagnosis of colon adenocarcinoma often include gastrointestinal bleeding or blockages, as well as general abdominal pain, fatigue, shortness of breath, and angina. A typical screening or diagnostic protocol for the detection of colon cancer includes digital rectal exam, blood testing, and a colonoscopy. Confirmation of a colon adenocarcinoma diagnosis usually involves the collection of a biopsy that is then reviewed by a pathologist using a microscope or equivalent imaging modality. If pathology studies confirm the diagnosis, additional imaging studies maybe used to both stage the primary cancer (e.g., assess its size and severity) as well as to detect any metastases. Treatments available for patients with confirmed colon adenocarcinoma include chemotherapy, radiation therapy and/or surgery. Of these options, surgery is usually the "front line" treatment, especially for those cases that are detected early. Of note, while less than 4 % of colon adenocarcinoma cases are directly attributable to familial genetics, an

additional 20 % of cases are significantly associated with multifactorial and rare variant genetic lesions that are not always easily recognized. Identification of such genetic traits in the later group requires complex genotyping and mapping to both patient and familial medical history as well as additional individual level clinical phenotype-related measurements (e.g. laboratory measures, patient reported outcomes, etc.). Individuals with both of the aforementioned genetic predispositions for colon adenocarcinoma are both at higher risk of the disease and may present with more aggressive cancers than other types of patients [6, 7].

Building upon this overall clinical context, let us consider a prototypical environment in which cases of colon adenocarcinoma may occur. We will define this environment in terms of the following axes: (1) evidence and/or policy generators; (2) healthcare providers and healthcare delivery organizations; and (3) patients with the disease or risk therein and their surrounding families and communities. Each axis is explored below. At the outset, we will situate our example is a medium sized North American city with a population of approximately one million people. In this city, there is a higher-than normal rate of colon adenocarcinoma with an incidence of 80 cases per 100,000 people in the community, or twice the average for the United States. That is, we can expect that at any given time, there may be up to 800 cases of colon adenocarcinoma; the exact number could be variable, depending on additional demographic and environmental factors that we will not go into detail about in this discussion.

2.2.1 Evidence and Policy Generators

In our prototypical city, a variety of researchers and policymakers have taken note of the increased incidence of colon adenocarcinoma and are investigating the basis for this rate of cancer as well as policy or other measures that could be taken to enhance or improve early detection and/or diagnosis of colon adenocarcinoma so as to decrease costs and quality of life impact associated with the disease and its treatment. Such measures could include but are not limited to: (1) funding research programs to identify the genetic or other bases for the increased cancer incidence; (2) launching a public education and awareness campaign concerning the importance of screening and early detection of colon cancer; and (3) taking measures across and between healthcare providers and delivery organizations to help primary and specialty care providers to readily identify those patients at greatest risk for colon adenocarcinoma (e.g., individuals with a high familial incidence and genetic predisposition).

2.2.2 Healthcare Providers and Organizations

In our same city, there are two to three integrated healthcare delivery organizations (e.g., combined inpatient and outpatient care providers, with associated supporting services), where at least one of those organizations is also part of an

academic health center (AHC) this is involved in additional teaching and research activities. Further, there is network of community based primary and specialty (including oncology) care providers, operating in small practices that are loosely aligned with the preceding integrated healthcare delivery organizations. Within this community, there is also a health information exchange (HIE), which allows for the flow of basic diagnostic and treatment history data between and among the Electronic Health Record (EHR) systems used by all of these care providers. Finally, the AHC includes a Biomedical Informatics department or center, as well as a genomic sequencing and bio-specimen collection/storage facility, which collectively are capable of retrospective and prospective sequencing and analysis of complex patient-specific genotypes for both research and clinical decision making purposes.

2.2.3 Patients, Families and Communities

Given the increased incidence rate of colon cancer introduced previously, it is likely the case that individuals, their families and various sub-parts of the community in this city will be concerned about the number of colon cancer cases surrounding them, the likelihood that they or their family members or friends might have such a diagnosis, and the need for more aggressive measures to identify individuals at risk or exhibiting the disease as early as possible in order to ensure optimal treatment outcomes and survival. Further, those individuals in the community who do have a diagnosis of, are undergoing treatment for, or have survived colon adenocarcinoma, may have an even more significant concern about the potential familial or genetic components of the disease that could put their family members at risk of similar diagnoses. As such, we can expect that individuals, affected families and community groups, such as churches or social entities, may engage in both information seeking and/or advocacy activities intended to address concerns surrounding the preceding high incidence of colon cancer, and the need for both early detection and/or better treatment options.

2.2.4 Putting the Pieces Together

In our prototypical (and idealized) setting, a number of measures and initiatives could be establish under the auspices of a TI paradigm, which can be aligned into the major categories detailed below:

- **Applying the emerging central dogma of Biomedical Informatics**: Given the confluence of public interest at the individual and community levels, as well as the capabilities of the incumbent AHC and regional HIE, and finally the desire of evidence and policy generators to take proactive measures to address the

increased incidence of colon adenocarcinoma, a research program could be launched with the multiple aims of:

1. Collecting and analyzing both bio-specimens and clinical phenotype data from patients with either diagnosed colon cancer or who have family members with such a diagnosis, in order to determine if rare genetic variants or other clinical factors may be responsible for the higher than normal number of cancer cases;
2. Simultaneously working with community members and organizations to determine if additional socio-demographic or environmental factors that may have been observed outside of the clinical setting may correlate with incidence of colon adenocarcinoma; and
3. Leveraging Biomedical Informatics theories and methods to "package" the information and knowledge gained from the data sets associated with items (1) and (2) as well as that available in the public domain (e.g., literature, public data sets, etc.) in order to deliver highly tailored clinical decision support to regional healthcare providers that would promote early screening activities as well as personalized treatments for individuals at risk of or having colon cancer that can be characterized using genomic, clinical, socio-demographic and environmental factors. Such approaches are concerned with capitalizing on the role of Biomedical Informatics as a conduit for translating a variety of raw data sources into contextualized information and ultimately actionable clinical knowledge.

- **Realizing the promise of translational science**: Achieving the vision articulated above relative to the conduct and delivery of wide-ranging research activities that have immediate and demonstrable clinical actionability, is an example of translational science in practice. However, such activities will require the coordination and collaboration of community members, clinicians, basic science researchers (e.g., geneticists), biomedical informaticians, clinical researchers, public health researchers and practioners, policy-makers, and funders. This type of multidisciplinary team must be assembled and operated in a manner that overcomes traditional "translational blocks" in order to rapidly move data, information and knowledge between and among such individuals. Further, this type of activity requires the design of clinical studies, care delivery guidelines and public health interventions that are based on the best possible and integrated scientific knowledge base. Findings and data should be rapidly translated between parties involved in the establishment and tracking of such initiatives.
- **Employing systems thinking**: Finally, as opposed to a traditional view of clinical care, research and public health in which those activities occur in at best a loosely coordinated manner, the scenario described here requires a systems level approach that bridges such "silos". By quickly engaging the community, clinicians, researchers, public health professionals and policy makers in a team based set of initiatives that combine large amounts of distributed and heterogeneous data, information and knowledge, we are ultimately enabling the type of hypothesis discovery and testing called for by a systems thinking model. Further, by

employing downstream solutions such as clinical guidelines for screening and treatment of colon cancer as well as public health interventions to promote early diagnosis of the disease, the solution to the increased incidence of colon adenocarcinoma takes on a systems-level configuration (as opposed to interventions limited to the aforementioned silos).

Unfortunately, as promising as the scenario described above appears, it is the exception rather than the norm relative to current approaches to healthcare delivery, research and policy formulation. The reasons for such challenges are at a high level a function of traditional reductionist thinking paradigms coupled with social, regulatory and technical barriers, and are further explored in Sect. 2.3 of this chapter.

2.3 Implications of This Prototype for Knowledge-Driven Healthcare

As noted in Sect. 2.2, while the ideal and prototypical scenario of using TI to overcome a higher-than-average incidence of colon cancer is highly desirable, it is also extremely uncommon given the current state-of-the-art in healthcare delivery, research and associated policymaking. This of course raises the question of why such a model is not widely seen in the "real world". Broadly speaking, there are a number of organizational, social, regulatory and technical barriers that serve to define this space, and that can be attributed to traditional and reductionist viewpoints. The major and contributing areas that make up current "silos" in the healthcare domain are illustrated in Fig. 2.2 and described below:

- At the core of the current healthcare paradigm is a **unidirectional care delivery model** in which clinicians diagnose and/or treat patients in episodic encounters, which are increasingly codified and recorded using Electronic Health Record (EHR) systems. In this approach, the patient is usually a passive recipient of care, and does not necessarily contribute substantive data or information to the decision making process that occurs during clinical encounters;
- **Clinicians who engage in clinical care utilize a variety of knowledge sources that are delivered to them in a similar unidirectional manner to them with variable time-frames for the delivery of that knowledge**. For example, healthcare educators train clinicians in terms of prevailing basic and clinical science. Such training is usually augmented and informed over time by new evidence that is generated by a variety of researchers. Finally, policy makers at the local, regional, national and international levels may generate guidelines or other policies corresponding to clinical best practices and reimbursement that serve to influence or constrain the decision making of a clinician. Of note, all of the aforementioned relationships to the clinician tend to be unidirectional and the generators of such knowledge or policies are rarely the recipients of data, information or knowledge "feedback" from the point-of-care unless they are engaged in highly targeted and formalized research programs.

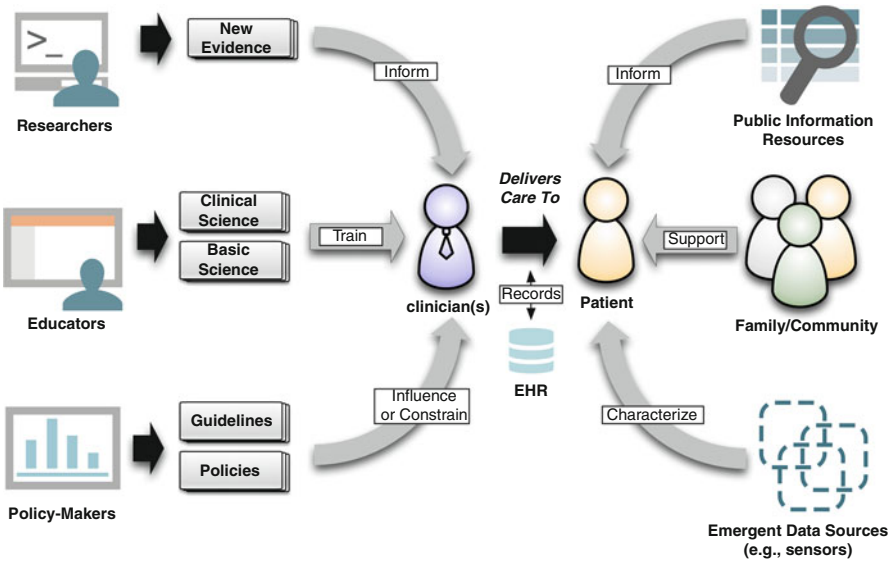

Fig. 2.2 Traditional model of healthcare delivery, in which various activities occur within complementary but isolated silos relating to research, education, policy-making, and the variety of constructs and data sources that surround and characterize a given patient

- Simultaneous to the preceding unidirectional care delivery and knowledge propagation paradigm, **patients are also supported and characterized by a number of constructs and data sources**, including:

  1. Their family members and surrounding community entities that often provide the bulk of both wellness promotion and healthcare delivery for individuals while they are outside of the clinical environment;
  2. Any number of sensors or other ubiquitous computing technologies (e.g., smartphones, etc.) that capture data that may be pertinent to measuring the health status of individuals; and
  3. Public data sources that both inform the socio-demographic aspects of an individual's health, as well as provide access to health and wellness information that may be acted upon by those individuals in a manner disparate and independent from the care delivered by their clinical providers. Examples include large public databases such as those associated with census or other survey data, and/or data derived from prior research programs and made available for reuse.

Unfortunately, models and mechanisms for bringing all of these items to the forefront of clinical decision-making are rare and in many cases, not well understood or trusted by care providers.

In contrast to the issues and concerns noted above, the prototypical case described in Sect. 2.2 provides a glimpse of what has been commonly referred to as a "Learning

Healthcare System" [8] in which an "Evidence Generating Medicine" [9] paradigm is adhered to. In this model:

- **Clinicians, patients, family members and community entities populate a healthcare "ecosystem"**;
- **All of the interactions between the individuals and entities that exist in the "ecosystem" are characterized using Biomedical Informatics theories, methods and technologies**, resulting in the population of numerous accessible and reusable information resources such as EHRs, Personalized Health Records (PHRs), emergent data generators such as the sensors and ubiquitous computing devices described earlier, and a variety of public data and information repositories;
- **These foundational resources in-turn catalyze evidence-generation via the conduct of clinical and translational science programs** that involve multi-disciplinary teams of researchers, educators and policy makers, all of whom can use the knowledge gained from such research to quickly inform their respective roles in terms of advancing science, educating the healthcare workforce or population-at-large, and informing large-scale policies and best practices; and
- **Such evidence generation supports and enables a systems-level approach** to analytics, the creation and delivery of actionable knowledge that can be delivered based on patient-specific characteristics, and the instantiation of decision support tools that bring the best possible knowledge to the point-of-care and wellness promotion, such as in the patients home or in community settings. This type of systems thinking ultimately delivers on the promise of personalized medicine [10], and informs the healthcare "ecosystem" introduced earlier.

Finally, and perhaps most importantly, all of these activities occur within a virtuous and rapid cycle (Fig. 2.3), via which every encounter that occurs in the healthcare "ecosystem" is an opportunity to learn and improve the care and wellness promotion delivered to patients, their families and their communities, building upon all of the preceding components of the "Learning Healthcare System".

2.4 Conclusions

The fundamental vision for Translational Informatics (TI), which has been introduced and elaborated upon in the first two chapters of this book, is predicated on the interaction of three critical and synergistic dimensions, namely:

1. The promise of clinical and translational research in the biomedical and healthcare domains;
2. The adoption and utilization of systems thinking approaches; and
3. The application of an emergent central dogma for the broad domain of Biomedical Informatics concerned with bridging the gaps between data, information and knowledge.

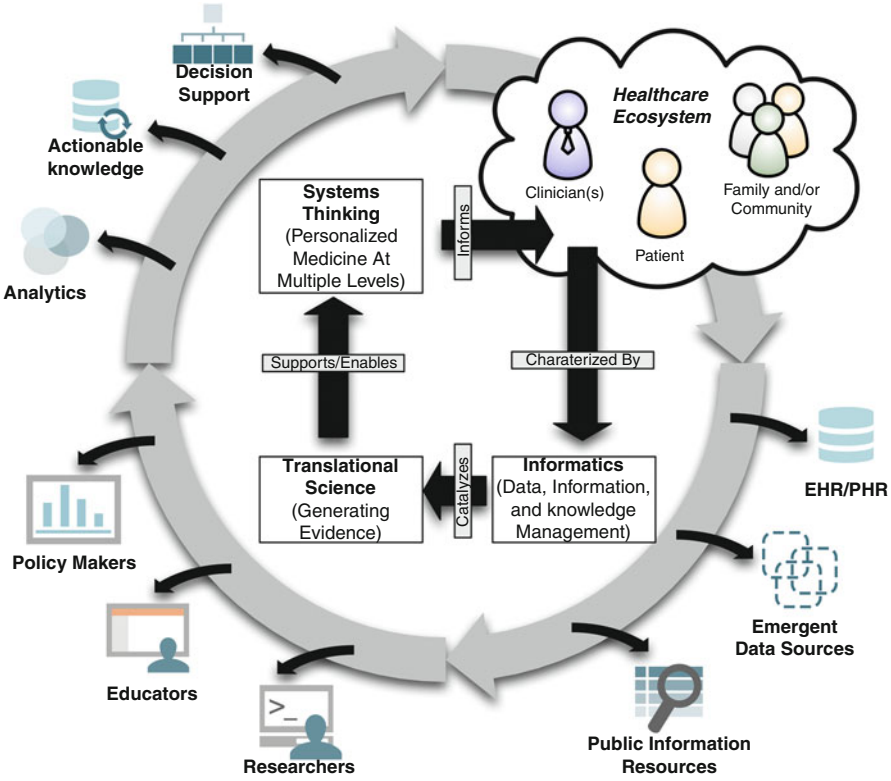

Fig. 2.3 Translational informatics vision for knowledge-driven healthcare, in which a virtuous and rapid cycle of informatics theories/methods, translational science, and systems thinking can be used to optimize every encounter occurring in the healthcare "ecosystem"

Further, these dimensions have broad ranging and significant impacts on a variety of actors and their activities that serve to make up the biomedical research and healthcare delivery environment, including: (1) evidence and policy generators; (2) providers and healthcare organizations; and (3) patients and their communities. In our prototype case that describes a community with higher-than-average incidence of colon cancer, we have shown how all of these factors, when combined, in order to achieve what has been called a "learning healthcare system", can have profound and important impacts on health and wellness that is pertinent to all of the preceding types of individuals and roles. Further, we have also introduced how traditional and reductionist approaches to the multiple areas that make up such a vision are impediments to its realization, and therefore can and should be mitigated. Building upon all of these arguments, in the ensuing chapters, we will survey a number of critical theories, methods, and examples that illustrate a path forwards to achieving this vision. Finally, we will revisit the impact of each chapter's contribution to the activities of such actors at the conclusion of each such discussion.

Discussion Points

- What are the major actors and activities impacted by the vision of Translational Informatics? How are they interrelated?
- What barriers exist to achieving this vision and what are their origins?
- How do the frameworks related to the creation of "learning healthcare systems" that exhibit an "evidence generating medicine" paradigm impact or influence the Translational Informatics and knowledge-driven healthcare vision?

References

1. Payne PR, Embi PJ, Sen CK. Translational informatics: enabling high throughput research paradigms. Physiol Genomics. 2009;39:131–40. Epub Sept 2009.
2. 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. PubMed PMID: 15974245, Epub 2005/06/25. eng.
3. Ahn AC, Tewari M, Poon CS, Phillips RS. The limits of reductionism in medicine: could systems biology offer an alternative? PLoS Med. 2006;3(6):709–13.
4. Auffray C, Chen Z, Hood L. Systems medicine: the future of medical genomics and healthcare. Genome Med. 2009;1(1):2.
5. Schadt EE, Bjorkegren JL. Network-enabled wisdom in biology, medicine, and health care. Sci Transl Med. 2012;4(115):115rv1.
6. Colon cancer: colon adenocarcionma. In: Pathologists CoA, editor. College of American Pathologists; 2011. http://www.cap.org/apps/docs/reference/myBiopsy/ColonAdenocarcinoma.pdf
7. Fearnhead NS, Wilding JL, Winner B, Tonks S, Bartlett S, Bicknell DC, et al. Multiple rare variants in different genes account for multifactorial inherited susceptibility to colorectal adenomas. Proc Natl Acad Sci U S A. 2004;101(45):15992–7.
8. Friedman CP, Wong AK, Blumenthal D. Achieving a nationwide learning health system. Sci Transl Med. 2010;2((57):57cm29.
9. Embi PJ, Payne PR. Evidence generating medicine: redefining the research-practice relationship to complete the evidence cycle. Med Care. 2013;51:S87–91.
10. Hood L, Friend SH. Predictive, personalized, preventive, participatory (P4) cancer medicine. Nat Rev Clin Onc. 2010;8:184–7.

Additional Reading

Ahn AC, Tewari M, Poon CS, Phillips RS. The limits of reductionism in medicine: could systems biology offer an alternative? PLoS Med. 2006;3(6):709–13.
Auffray C, Chen Z, Hood L. Systems medicine: the future of medical genomics and healthcare. Genome Med. 2009;1(1):2.
Embi PJ, Payne PR. Evidence generating medicine: redefining the research-practice relationship to complete the evidence cycle. Med Care. 2013;51:S87–91.
Friedman CP, Wong AK, Blumenthal D. Achieving a nationwide learning health system. Sci Transl Med. 2010;2(57):57cm29.
Hood L, Friend SH. Predictive, personalized, preventive, participatory (P4) cancer medicine. Nat Rev Clin Onc. 2010;8:184–7.
Payne PR, Embi PJ, Sen CK. Translational informatics: enabling high throughput research paradigms. Physiol Genomics. 2009;39:131–40. Epub Sept 2009.
Schadt EE, Bjorkegren JL. Network-enabled wisdom in biology, medicine, and health care. Sci Transl Med. 2012;4(115):115rv1.

Part II
Foundations of Translational Informatics

Chapter 3
Personalized Medicine

Jessica D. Tenenbaum

By the End of This Chapter, Readers Should Be Able to

- Understand the meaning of the term "personalized medicine";
- Become familiar with the underlying biology and novel technologies that have enabled and enhanced the realization of personalized medicine;
- Learn about "hot topics" and future directions in the area of personalized medicine; and
- Understand the implications of personalized medicine from the viewpoint of patients and their communities, healthcare providers, and policy makers.

3.1 Introduction

Personalized medicine is a key concept in discussions regarding the transformation of health care. At the highest level, personalized medicine means health care, and disease prevention, that is targeted to the individual at the appropriate time. By that definition, though, any senior physician will tell you she has been performing personalized medicine for decades. So what has changed?

Personalized medicine in the post-genome era represents a confluence of factors in which informatics plays a key role, enabling the transformation of increasingly voluminous amounts of data into information, and subsequently translating this information into actionable knowledge both in research and at the point of care. First, the sequencing of the human genome and the technologies that project has

J.D. Tenenbaum, PhD
Duke Translational Medicine Institute,
Duke University School of Medicine, Durham, NC, USA
e-mail: jessie.tenenbaum@duke.edu

P.R.O. Payne, P.J. Embi (eds.), *Translational Informatics: Realizing the Promise of Knowledge-Driven Healthcare*, Health Informatics,
DOI 10.1007/978-1-4471-4646-9_3,

enabled have triggered a paradigm shift in biological inquiry. Instead of focusing solely on one gene or protein of interest, technologies developed over the past few decades have enabled collection of data across tens of thousands of molecules at a time. In this way, biology has become more of a systems-oriented, data-driven discipline. In addition, increasing adoption of health information technology in the form of electronic medical records, enables a learning health care system in which information collected through clinical care, at different points in time and by different providers, can be more easily consolidated and retrieved for use in both aggregated analysis and clinical decision making. Increased information about an individual, increased access to that information at the point of care, and more precisely targeted guidelines help to enable delivery of the right treatment to the right patient, at the right time.

3.2 Not Your High School Teacher's Biology

3.2.1 Refresher: The Central Dogma of Biology

In order to give context to the sections below, what follows is a brief refresher in the basics of molecular biology. Recall that DNA (deoxy-ribonucleic acid) exists as a double helix in the nucleus of the cell. Its shape is often compared to a ladder, though it might be better compared to a spiral staircase in which the steps are made up of pairs of molecules known as nucleotides or "bases," specifically, cytosine (C), guanine (G), adenine (A), and thymine (T). A small portion of the genome, on the order of 1.5 %, is devoted to the specification of protein sequences [1], while the vast majority of human DNA does not code for proteins. Of the remaining 98.5 %, some is known to be structural, e.g. ribosomal RNA, which makes up part of the ribosome, and some is used for gene regulation. For much of it, though, the function remains unknown. The ENCODE project is a large NIH-funded initiative intended to decipher the purpose and function of the rest of the genome, beyond the protein coding segments. While it has long been suspected that use of the term "junk DNA" for the remaining 98.5 % was a misnomer, the ENCODE consortium surprised many researchers with the assertion that they had identified biochemical functions for over 80 % of the genome.

DNA has the property that specific bases only pair with specific other bases, specifically G with C, and A with T (see Fig. 3.1). This leads to the ability for a single strand of DNA to be used as a template to create a matching (complementary) copy, which can in turn be used to create an exact copy of the initial sequence. This template-based replication is what takes place when cells divide, creating two new cells with the same genome as the original cell. This property can also be exploited for a number of other purposes, from amplifying DNA (i.e. making many copies of a specific DNA molecule), to capturing and identifying DNA, to its use as a "barcode" to identify different species of organisms.

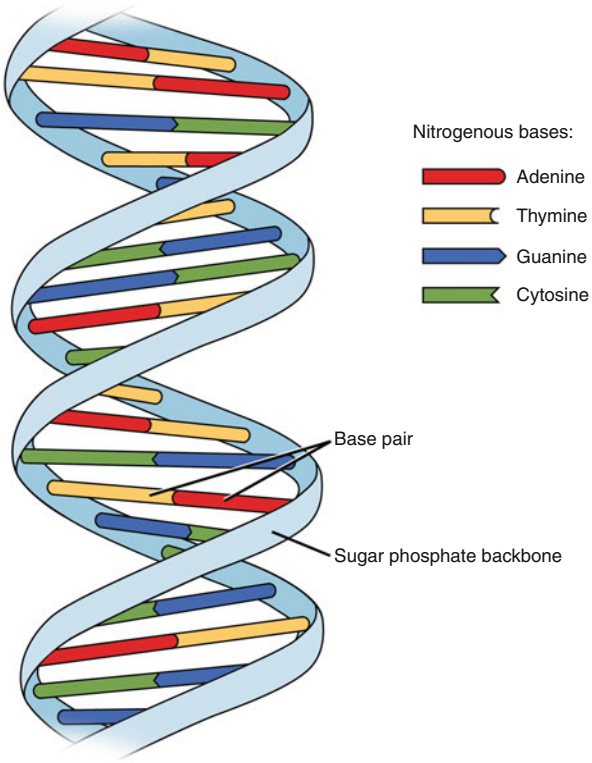

Fig. 3.1 DNA takes the form of a double helix, with nucleotide pairs as the "rungs." Adenine pairs with Thymine, and Guanine pairs with Cytosine (Reprinted with permission from: OpenStax College. Organic Compounds Essential to Human Functioning, OpenStax_CNX Web site. http://cnx.org/content/m46008/1.4/, Jun 27, 2013)

When a given gene is activated or "expressed", the information from that portion of the DNA is transferred to a complementary strand of *ribonucleic acid*, or RNA, through a process known as *transcription*. RNA is made up of nucleotides similar to those in DNA, but with uracil (U) instead of thymine (T). G still pairs with C, and A with U. This complementary strand of RNA undergoes certain processing (e.g. splicing out of introns, addition of a "poly-A tail", etc.) and is then *translated* into a polypeptide sequence, or chain of amino acids. The polypeptide then folds in three dimensions and may undergo *post-translational modifications*, e.g. additions of functional molecular groups, resulting in a functional protein molecule. Ultimately, proteins are broken down into amino acids and derivative compounds (metabolites) and eliminated from the cell. Figure 3.2 summarizes this process. The term genomics may be used to refer to data regarding DNA sequence or gene expression (A, B). The term transcriptomics refers to gene expression (B). Proteomics may be used to refer to all aspects of proteins, from sequence to shape to modifications (C, D). And metabolomics is used to refer to metabolites, the smaller molecules that are generated as a result of the breakdown of proteins (E).

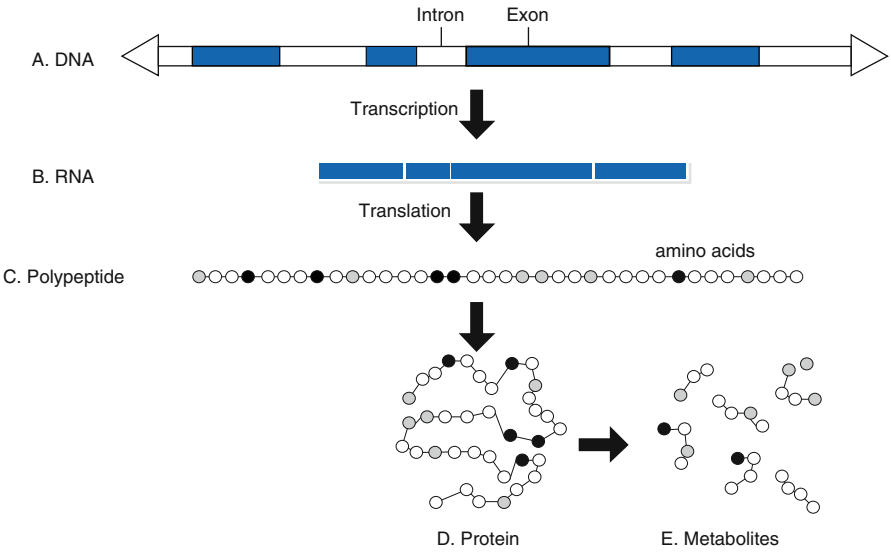

Fig. 3.2 Molecular biology refresher. (*A*–*C*) The central dogma of biology: DNA exists in the cell nucleus. Through a process known as transcription, the DNA is "read" and a matching strand of RNA is created. After further processing of this messenger RNA (*mRNA*), the specific sequence is "translated" into a corresponding sequence of amino acids, forming a polypeptide chain. (*D*) This chain folds in a highly specific manner to form a three-dimensional protein structure. This 3D structure, and consequently the functionality of the protein, may be affected by post-translational modifications, e.g. the addition of a phosphate group (PO_4) (*E*). Proteins are broken down into metabolites to be eliminated from the cell

One's DNA sequence, or genome, is generally the same across every cell in the body, while changes in the expression levels of genes, and the quantity and state of gene products (i.e. proteins) determine both the cell type and the biological state of the cell. The ability to measure which genes and gene products are active in different tissues gives critical clues regarding the underlying mechanisms of biological processes and how these processes can be disrupted in disease.

3.2.2 The "Omics" Revolution

The amount that is known about the human genome, and the tools at our disposal to learn more have expanded significantly since most of today's adults last took a course in biology. The mid-1990s saw the advent of DNA microarray technology, which could be used to quantitate the expression of tens of thousands of genes in a single assay. In 2001, a complete draft of the human genome sequence was announced by two different competing groups. The public effort, led by the NIH, was carried out with the intention to share all data from the start [1]. Celera

Genomics, headed up by J. Craig Venter, performed a parallel effort [2]. This private effort initially intended to retain their sequence data as proprietary, but later made it available for non-commercial use [3]. With the completion of these draft sequences, which took approximately 10 years and $100 million [4], scientists had the "parts list" for human cellular biology. However, in much the same way that a parts list for a Boeing 747 would not be sufficient for building an airplane, or troubleshooting it once it was built, a genomics parts list is only the beginning. The key information is how all the different parts work together. Upon sequencing the human genome, then, the focus shifted from cataloging the parts to understanding what each of those parts does both individually and collectively. This transition helped to bring about a radical change from the reductionist approach used in the past to a more holistic and dynamic systems approach used widely in biological research today.

The word "genome" was coined in 1930 as a combination of the words "gene" and "chromosome." Despite the term's venerability, the field of *genomics* arguably only came into its own in the mid-1990s with the invention of DNA microarrays, which enabled measurement of gene expression across tens of thousands of genes at one time. These chips, no bigger than a microscope slide, are printed with specific DNA sequences from known genes. The general approach entails labeling cellular RNA (i.e. the RNA that is present in cells for genes that are activated or "turned on") with fluorescent dyes and then washing the labeled RNA over the chips. Through complementary base pairing, the labeled RNA only sticks to spots with a matching sequence (Fig. 3.3). Spots that light up when viewed with a laser scanner indicate

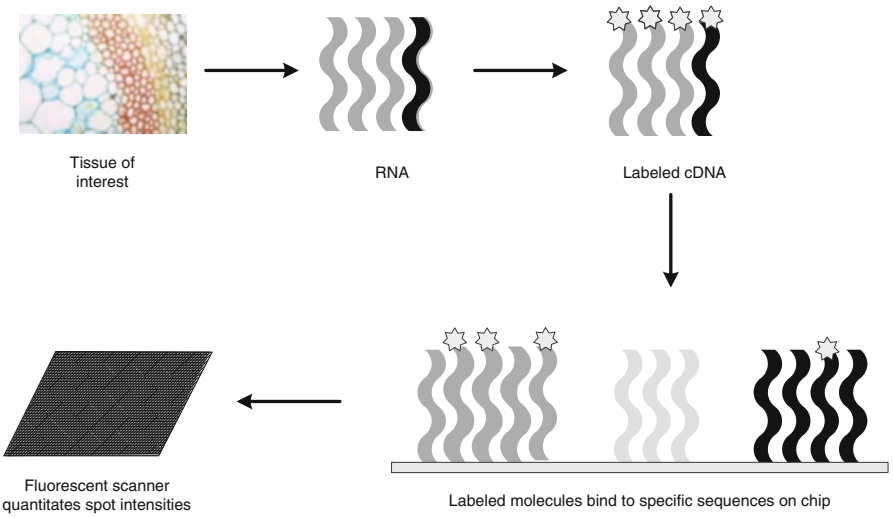

Fig. 3.3 Overview of DNA microarrays. Messenger RNA is extracted from the cells of interest and serve as a template for the creation of labeled cDNA. Those labeled molecules are then washed over a chip printed with known, gene-specific sequences at known coordinates. A fluorescent laser scanner is then used to detect the intensity of each spot, which corresponds to the amount of RNA that was present in the cell for each of the genes represented on the chip

that the corresponding gene is activated in the sample in question. Brighter fluorescence indicated a higher degree of activation. Spots that do not light up, because no labeled RNA stuck, indicate that the corresponding gene was not active in the given sample. In this way, researchers can compare the cellular activity of, for example, skin vs. muscle, or healthy lung vs. cancerous lung tissue. Functional relatedness can also be hypothesized for genes that are expression at the same time, under the same conditions.

A similar technique can be used to determine specific genotypes. Both versions of the variable portion of a gene are printed on an array, and then the relative levels of bound RNA compared, suggesting heterozygosity (two different nucleotides, one from each parent, e.g. AG) or homozygosity (two of the same nucleotide), and for which base (i.e. AA vs. GG). This information can then be used for genome-wide association studies (GWAS) in which a population with a given phenotype is compared to a control population. Each observed SNP (single nucleotide polymorphism) is evaluated for statistical enrichment in cases versus controls. Enrichment for a given genotype in one group or the other suggests a causal mutation *in that area* of the genome. This approach has a number of limitations. One is that only a finite number of SNPs are printed on a given array (on the order of one to two million), generally reflecting relatively common variants that have already been identified as polymorphic. Only those SNPs that are printed on the array can be directly observed, but they may not be the actual causal mutation, or even in the same gene as the true mutation. In this sense, GWAS only provides a "guilt-by-association" approach to deciphering the underlying mechanism of disease. Due in part to this drawback, as well as the rapidly dropping cost of sequencing, genotyping is increasingly being performed through sequencing as an alternative to the chip-based approach. In addition, GWAS suffers from what is known in quantitative sciences as the "curse of dimensionality." That is, by its very nature it entails multiple hypothesis testing—as many as one to two million hypotheses, in fact. Correcting for this degree of multiple hypotheses can make it very difficult to detect actual signal.

Another key technological advance, widely adopted beginning in 2004, was next generation ("NextGen") sequencing. NextGen sequencing offers significant advantages over Sanger sequencing, the method used for both early human genome sequencing projects. In both cases, DNA is amplified and then cut into millions of short, overlapping fragments. These individual fragments are sequenced, and then informatics techniques are used to "stitch" together the "reads," or short sequences, into long contiguous sequences. Sanger sequencing is based on "DNA chain termination," which relies on selective incorporation of chain-terminating bases, and then separation of different sized molecules using gel electrophoresis. The process is relatively slow and expensive, but it is still used for small-scale projects and in cases where long contiguous reads are desired since the Sanger approach can produce reads up to 1,000 bases in length. NextGen sequencing enables sequencing to be done in a massively parallel manner, speeding up the process to where an entire human genome may be sequenced in a matter of days. Instead of identifying the sequence of bases through chain-termination followed by separation in a gel, the various flavors of NextGen sequencing identify nucleotide sequences by synthesiz-

ing DNA, adding one base at a time and observing which base is incorporated in any given step. In January 2014, Illumina announced that through the release of its HiSeq X Ten, the long sought-after $1,000 genome barrier had been broken [5]. One downside to NextGen sequencing is that the read lengths are relatively short (on the order of 100 bases). This makes it impossible to accurately sequence some portions of the genome, particularly highly repetitive regions.

Third generation sequencing is aimed at taking the next big leap: continuous single-molecule sequencing. Instead of determining a very long DNA sequence by sequencing millions of short reads and then aligning them, third generation sequencing reads the individual based pairs from a single molecule [6]. Major advantages to this approach include the small sample size required and increased accuracy in highly repetitive regions of DNA. As the price of sequencing has dropped, these techniques are increasingly being used not only for DNA but for other assays too, e.g. RNA expression (RNA-seq) and transcription factor binding assays (ChIP-seq).

With these technological advances for generating massive amounts of omics data, biologists can no longer glean new scientific knowledge simply by looking at a gel or a spreadsheet. Rather, computational tools are required to make sense of tens of thousands of data points from a single assay. Visualization tools, statistical methods, and machine learning techniques are all critical for turning genomic data into knowledge. Thus was born the field of bioinformatics, opening up a whole new frontier of scientific inquiry. As described above, DNA microarrays enabled elucidation of underlying pathways by showing which genes had similar expression patterns. Genome-wide associated studies (GWAS) enable researchers to identify genotypes that are associated with a given condition, providing hints regarding the area of the genome where causal mutations are found. And whole genome and whole exome sequencing have enabled fine-grained detection of rare, disease-causing mutations.

It is worth noting that a significant proportion of the field of personalized medicine, and hence the content of this chapter, is devoted to genomics, despite the fact that the real action in biology tends to happen downstream of genes, at the level of proteomics and metabolomics. Unfortunately, proteins and metabolites do not have DNA and RNA's intrinsic base-pairing quality, so specific quantitation can be more difficult. Protein microarrays have been developed with which proteins can be captured using antibodies and fluorescent dyes. But by far the most popular approach to quantification of proteins, peptides, and metabolites is the use of mass spectrometry (MS), a technology that has actually been in existence since the first half of the twentieth century. MS involves the separation, ionization, and detection of molecules and their sub-components. This enables researchers to compare the observed particle sizes with known molecular weights and thus deduce which molecules are present in a given sample and their relative quantity. MS may be used in a discovery fashion to detect all molecules above a minimum level of abundance in a given sample, or specific molecules may be labeled with a radioisotope, enabling quantification of the molecule in question relative to the labeled molecule of known concentration.

Together, these various omic technologies have provided biomedical researchers, and increasingly health care providers, with new instruments with which to query the state of an individual. In much the same way as the microscope and stethoscope changed the way doctors surveyed the conditions of their patients, so too has the omic era expanded the scope of possible clinical observations, enabling a much more fine-grained picture of what is going on with a patient.

3.3 Medicine by Any Other Name

Between the medical literature, a major report from the Institute of Medicine, and mainstream media, personalized medicine has been extolled as the future of health care. In this section we discuss what is meant by personalized medicine, as well as some issues and caveats around it.

*3.3.1 P**

While there is near consensus regarding the need for personalized medicine, far less agreed upon is what to call it. A 2011 report by the Institute of Medicine describes a path toward "Precision Medicine" [7]. Others refer to P4 medicine (predictive, preventive, personalized, participatory) [8], individualized [9], targeted [10], genomic [11], or stratified medicine [12]. The details or emphasis for each of these concepts varies, e.g. a focus on genes and gene products, timely intervention, patient proactivity, or grouping individuals into similar cohorts, but common across all of these concepts is improving our ability to target the right intervention for the right patient at the right time (See Fig. 3.4.).

3.3.2 Biomarkers

Biomarkers may be defined as any biological phenomenon that gives information regarding some underlying biological state [14]. Macroscopic biomarkers have been used for millennia—fever, pain, rash. More recently, in the latter half of the twentieth century, cellular and molecular data points, e.g., glucose, antibodies, and prostate specific antigen (PSA), have been used in this way. The various omics technologies described above have opened up a new frontier of potential biomarkers. Often these biomarkers represent not a single data point but a multidimensional signature of biomarkers. These signatures may be used to further stratify individuals, beyond the limited diagnosis or prognosis enabled by more traditional methods.

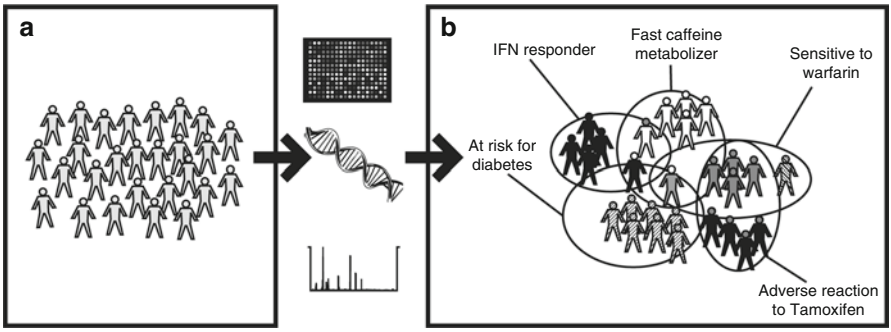

Fig. 3.4 Personalized medicine: (**a**) A group of individuals may appear homogeneous upon macroscopic observation through traditional methods. (**b**) Novel molecular assays enable discernment of underlying physiological differences. This stratification can inform decisions regarding lifestyle, disease prevention, and therapeutic interventions (Reprinted from [13], used with permission from Springer)

3.3.2.1 Predictive vs. Mechanistic Biomarkers

Different players have different motivations for probing underlying biological state through the use of biomarkers, and different types of biomarkers may be used for different purposes. Health care providers are primarily interested in *predictive* biomarkers. The biomarker itself may be causal for a disease or phenotype, or simply correlated with the phenotype because both are physiologically downstream of the actual cause. As long as the marker can be used as a reliable indicator for, e.g. diagnosis, prognosis, or therapeutic response, it can be useful from a clinical perspective. *Mechanistic* biomarkers, in contrast, are useful to researchers to devise methods for intervention and prevention. Understanding underlying disease mechanism can help identify causal pathways, which may in turn suggest new directions for hypothesis-driven research or new leads for drug targeting.

3.3.2.2 From Statistical Significance to Clinical Utility

In any discussion of biomarkers, it is critical to recognize the differences between statistical and clinical significance, analytic and clinical validity, and clinical utility. Statistical significance is a measure of confidence that a finding from a statistical test is actually true. Some common statistical tests performed in the biomedical context include the t-test, Chi-squared test, or analysis of variance (ANOVA). In each case, the test essentially asks whether different sets of values likely came from the same distribution, or different distributions. Statistical significance is often expressed in the form of a p-value, which represents the probability of being wrong if one were to conclude that the distributions are indeed different. (Drawing this conclusion is commonly described as “rejecting the null hypothesis”, i.e. the hypothesis that there

is no difference between the groups.) A *p*-value of 0.05 suggests that the observed results could be attributed to chance only 5 % of the time [15]. The value of 0.05 is an arbitrary, but commonly accepted, threshold below which a test is said to be *statistically significant*. Typically, the more samples one uses to make a determination, of significance the more statistically significant a finding will be. Clinical significance, on the other hand, reflects whether that difference has any impact on clinical care. As an example, one might hypothesize that a given genotype confers increased risk for heart attack. If, hypothetically, one were to test that gene in one million cases and one million controls, it may be shown that individuals with the genotype in question were 1.1 times as likely as those without it to have a heart attack. With such a large sample size, the *p*-value for this finding could be <0.00001, a statistically significant result. However, from a health care perspective, it changes nothing. This hypothetical test thus lacks clinical significance.

Analytical validity refers to the accuracy and reliability of a test itself, e.g. how well a given test predicts the presence or absence of a particular genetic change. Criteria for analytic validity include not only sensitivity and specificity of the test itself, but also factors such as reproducibility, quality control, and limits of quantitation [16]. *Clinical* validity, on the other hand, is a measure of the degree to which the test result reflects the presence, absence, or risk of disease. Clinical validity relies on analytical validity, but also incorporates disease penetrance and prevalence and the concepts of positive and negative predictive value, i.e. if the test indicates a person has the disease, how likely is it that the person *actually* has the disease, and vice versa.

Clinical utility is related to the concept of clinical significance. It is a measure of how useful that information actually is in informing clinical care. A biomarker test may be 100 % accurate, and perfectly indicative of a given disease, but if no treatment for that condition exists, the test lacks clinical utility. Alternatively, a test may be suggestive for specific treatment course, but cost of treatment or severity of side effects may outweigh that recommendation in light of uncertainty. In addition, clinical utility must be evaluated in the larger clinical context. An advanced biomarker is only useful if it provides additional information beyond what could be gleaned through standard, more easily obtained, observations. For example, in order to have clinical utility, a molecular biomarker for risk of heart attack would need to be more accurate than a standard clinical model that takes into account BMI, smoking status, family history, etc.

3.3.3 Stratification

Biomarkers can be used to stratify patients along a number of different axes. This stratification may or may not be actionable. For example, some people respond well to certain drugs while others do not respond at all, or worse, have an adverse event. Knowing which of these groups a patient belongs to can help inform pharmaceutical intervention. Biomarkers may also help to group people by diagnosis where macroscopic observations cannot differentiate. For example, it can be difficult to know whether flu-like symptoms are caused by a virus or a bacterial infection, but if

patients with a virus can be designated as such, unnecessary prescription of antibiotics may be avoided [17]. One of the most promising areas of application for disease stratification is in cancer, where molecular information can provide clues regarding which pathways have been dysregulated, and thus what therapy is likely to work best, or not at all. Less specific, but still potentially quite valuable, is grouping patients by prognosis. Knowing that someone has a very poor prognosis can help to inform decisions regarding how aggressive to be with treatments that are known to have unpleasant side effects.

Even when a distinction is not medically actionable through a specific therapeutic decision, biomarkers to differentiate between different groups can be useful. If a biomarker signature can be used to predict symptomatic flare-ups in advance, then even if no treatment is available, a patient can use this information to inform decisions regarding work schedule or recreational plans. Relapsing remitting multiple sclerosis is a good example of such a condition [18]. Similarly, if a cohort with a given diagnosis can be stratified by likelihood of disease progression, this can help increase the financial feasibility of carrying out a clinical trial on potential leads. As an example, currently available drugs for osteoarthritis treat only the symptoms and not the disease [19]. This is true in part because it can be hard to predict which patients' disease will progress and whose will remain static, thus requiring very large numbers of participants in a clinical trial in order to obtain sufficient statistical power. In this case, biomarkers for likely disease progression can be used to enrich the population being tested with likely progressors, enhancing statistical power and ultimately lowering the high cost of bringing a drug to market.

3.3.4 Beyond Biomarkers

It is important to note that personalized medicine is not solely about physiological biomarkers. Equally important are intangible factors such as personal preferences and values. A risk-averse person may choose a low-risk intervention that will only partially resolve a condition rather than a radical option that could effectively cure the condition, but comes with a higher risk of mortality. Likewise, a person might make different decisions, or differently timed decisions, if he or she is starting a new job, about to get married, or about to become a grandparent for the first time. Unfortunately, financial resources and insurance coverage factor in as well. Other considerations might include past prescription compliance, environmental conditions, literacy, and presence of caregivers and a support network.

3.3.5 Timing

One theme throughout the different approaches to personalized medicine, particularly in the P4 version, is that of timeliness. Intervention before a patient is symptomatic, or better yet before a person is sick, is far less expensive than the therapies

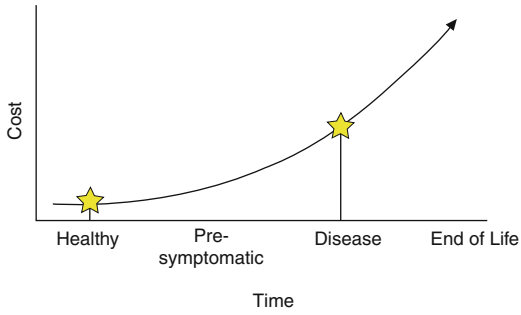

Fig. 3.5 The cost of health care is far lower for interventions performed before a person falls ill

and procedures that are required once disease has set in (see Fig. 3.5). The ability to target interventions based on who is at risk for a given disease, and ideally to prevent the disease from manifesting in the first place, would significantly raise an individual's quality of life and also lower health care costs. To this end, personalized medicine aims to detect individuals who are at risk for a particular disease so that, for example, diet and lifestyle may be changed before a high risk person has a heart attack at age 45. Of course, everyone knows that exercise and a healthy diet are beneficial to one's health, and yet few people practice these guidelines. Studies are ongoing to detect whether knowledge of one's personal risk provide the additional motivation required to catalyze actual change [20].

3.4 Translational Bioinformatics in Personalized Medicine

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, and genomic data, into proactive, predictive, preventive, and participatory health" [21]. It is clear, then, how TBI plays a critical role in personalized medicine, as described above. We describe here some "hot topics" in personalized medicine for which TBI methods play an integral role.

3.4.1 Pharmacogenomics (PGX)

One of the most successful areas of application for personalized medicine approaches has been in pharmacogenomics, or how a person's genes affect his or her response to drugs. Interestingly, while attempts to identify genes responsible for specific diseases have been somewhat disappointing, genes affecting the body's ability to process and metabolize drugs have been more readily discovered. This may be in part because there has been selective pressure against disease-risk genes

over many millennia, while mutations affecting drug response have only relatively recently become relevant [22].

Warfarin is a prime example of a drug for which genetic information may affect prescribing. It is an anticoagulant, initially developed as rat poison in the mid-1900s. The drug itself has a very narrow therapeutic index: too low a dose and there is no therapeutic effect, too high a dose and the patient may bleed out. Two genes in particular, VKORC1 and CYP2C9, are known to affect the activity of warfarin, causing the FDA to include this information on the drug label. Various resources (e.g. websites, smartphone "apps") exist to provide clinical decision support that incorporates genotype data for initial dosing, though clinical studies to date have reached different conclusions regarding improved outcomes through use of genetic information in dosing [23, 24].

Other examples of pharmacogenomics success stories in recent years include targeted use of ivacaftor (Kalydeco™) to treat cystic fibrosis in patients with G551D CFTR mutations [25], approval of crizotinib (Xalkori™) for ALK positive non-small cell lung cancer [26], and Vemurafenib, the first drug approved for BRAF-mutant cancer [27].

3.4.2 Direct to Consumer (DTC) Genetic Testing

The direct to consumer (DTC) genetic testing landscape has been an evolving one since deCODE and 23andMe first launched their services in 2007, with others to follow soon after. These companies varied fairly widely in terms of cost, number of mutations tested, and degree of support provided. In general, the customer would provide a biospecimen, e.g. by spitting into a test tube or taking a buccal swab, and send the specimen back to the company to be processed. The test results were then made available to the customer through a secure web interface (see Fig. 3.6). Some companies focused primarily on actionable, disease-related traits. 23andMe, the main player still standing by 2014, included both health traits (at least initially, see below) and other more recreational information, such as wet versus dry earwax, and the ability to smell a metabolite of asparagus in one's urine. They also introduced functionality around tracing ancestry, and the ability to participate in surveys, thereby furthering biomedical research.

DTC testing has been somewhat controversial from a regulatory perspective. It was, and at some level still is, unclear to what extent these services are or should be regulated by the FDA. Through 2010 and 2011, the FDA approached companies individually, with communications suggesting that their services might qualify as devices and therefore require FDA review and approval for use. They requested that the companies either explain why this was not the case, or how they intended to pursue agency approval. In late 2013, the FDA issued a letter to 23andMe ordering that it cease to market its service until FDA authorization was received. Shortly thereafter, the company stopped airing a television advertisement in which they promoted the service's

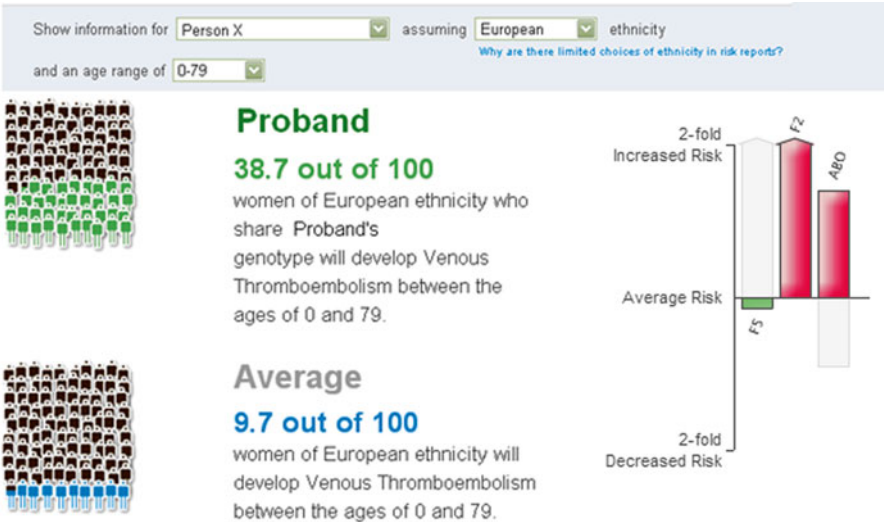

Fig. 3.6 Interface for a 23andMe health report (no longer offered to new customers). The odds calculator shows the estimated incidence of a disease for a person sharing this individual's gender, ethnicity, and genotype for the selected markers. Only markers that are included in 23andMe's gene chip are included in the calculation, and the relative risk shown does not factor in environment or lifestyle, which are both known to play a role in one's risk for disease (©23andMe, Inc. 2013. All rights reserved; distributed pursuant to a limited license from ©23andMe)

ability to inform customers of their risk for specific diseases and conditions, and stopped offering health-related reports for anyone who purchased a kit after the date when the FDA issued its letter at the time of writing, (and take out the work "now"). The company's website now indicates that they provide ancestry information and raw genomic data, but that they have suspended health-related genetic reports [28].

3.4.3 Sequencing in Rare Diseases

Rare diseases are very difficult to detect using GWAS due to the "guilt by association" approach that SNP chips necessitate. DNA sequencing to determine genomic variation can have far greater resolution to pinpoint the actual functional variant. Indeed, there have been some spectacular successes using this approach, some of which have been picked up by the mainstream media. As one example, in 2007, a 2-year-old boy named Nicholas Volker had a mysterious and excruciating bowel condition. Whole exome sequencing allowed researchers to compare his DNA with the human reference genome in order to determine the likely culprit. Researchers were able to identify 1,500 novel mutations in Nic's genome. A causal mutation was identified and, equally as important, a known treatment was available.

A curative cord blood transplant was performed. Two years later, while not without challenges, Nic was eating real food, playing sports, and generally living his life, outside of the hospitable and without the excruciating intestinal episodes that started him and his family down their diagnostic odyssey [29]. Nic celebrated his ninth birthday in late 2013.

In 2011, fraternal twins Noah and Alexis Beery had their genomes sequenced to identify the genetic mutations that were causing severe health problems in both, though manifesting differently in each [30]. At age 5 they had been diagnosed with dopa-responsive dystonia, for which they were prescribed a dopamine precursor. That treatment had helped with symptoms until the twins were 13, when Alexis developed a severe respiratory condition. Genome sequencing identified mutations in a gene called SPR, or sepiapterin reductase, which enables the synthesis of neurotransmitters. This mutation had been previously linked to some cases of dopa-responsive dystonia. Already taking the dopamine precursor, the twins were additionally prescribed 5-hydroxytryptophan, a serotonin precursor. Within a month, this additional therapy resolved the respiratory condition.

In 2005, a woman named Beth McDaniel was diagnosed with a rare T cell lymphoma [31]. Chemotherapy held the disease at bay for 5 years, but in 2010 the tumors under her skin came back. Her son, a molecular biologist, took a leave of absence from his job to devote himself full-time to seeking an answer through DNA sequencing. With considerable effort and resources recruited to the cause, scientists were able to identify a gene fusion event that was causing signals for T cells to stop growing to be interpreted as signals to grow, and vice versa. As luck would have it, a new melanoma drug had been approved that worked by signaling T cells to grow. The hope was that this drug could be used to cause Mrs. McDaniel's T cells to stop growing. Beth's response to the new drug was immediate and striking, but unfortunately short lived. Two months later, the cancer was back, and 2 months after that, it took her life.

One last example is that of Dr. Lukas Wartman, a cancer researcher at Washington University. When he was diagnosed with adult acute lymphoblastic leukemia, his colleagues at the university's genomic institute put other work on hold to sequence his entire genome, both in cancer cells and healthy ones [32]. Through this exercise, they were able to identify a causal mutation for which an FDA approved drug existed. The drug had been approved for use in kidney cancer, not leukemia, but as reported in the New York Times, was successful in driving Dr. Wartman's cancer into remission.

It is worth noting that these heartwarming success stories are the exception, not the rule. In each case, not only were researchers lucky to find a causal mutation, there was also a known, approved therapeutic intervention to target that mutation. That is not always the case. In addition, while a tumor sequencing approach is becoming slightly more common in cancer cases where standard treatment has failed, this approach is generally used only in special circumstances. In almost every case, the patient[s] in question knew someone who had specialized knowledge and who worked in cutting edge research organizations. Dr. Wartman was an expert in the field and worked at a university with a world class genome institute.

The Beery twins' father was CIO at Life Technologies. Timothy McDaniel, Beth's son, worked for Illumina. Though these stories give cause for optimism, we are still a long way from widespread use of sequencing to guide clinical care.

3.4.4 Epigenetics

Epigenetics refers to functionally relevant, heritable changes to the genome that do *not* involve changes to DNA sequence [33]. Some examples include DNA methylation (the addition of a methyl group to a specific site along the DNA molecule) or modification of histones, the proteins responsible for maintaining the 3D structure of DNA in its coiled state in the nucleus. Much remains to be discovered about the mechanisms of epigenetic phenomena, but links have been demonstrated to various biomedical phenomena, such as aging [34] and oncogenesis [35]. It is likely that epigenetic biomarkers will increasingly help to target therapies and interventions in much the same way as the genomic biomarker signatures discussed above.

3.4.5 Gene/Environment Interaction

While genome-wide association studies have revealed a number of locations along the genome that appear to play a role in various different diseases, rarely is the correlation absolute. That is, a given genotype may be more common in people who have a given disease than those who do not, but by no means does everyone with that genotype develop the disease. Of course, many diseases are complex, with numerous genes and pathways playing their respective roles. But even with identical twins who generally have the same genotype across all genes, one may become sick while the other remains healthy.

It is important to note that genes do not manifest themselves in a vacuum. The interactions between people's genes and the environment to which they are exposed are critical in determining downstream phenotype. Take smoking as an example. Different people's genotypes predispose them to the risk of cancer to varying degrees. Within a set of identical twins, the genetic risk is essentially the same. However, if one twin smokes regularly, and the other is a life-long non-smoker, the smoker is far more likely to develop lung cancer. As another example, exposure to sunlight is known to increase the risk of skin cancer. However, sun exposure is confers greater risk for light skinned individuals than for dark skinned individuals. Recent work has begun to uncover some molecular mechanisms underlying these gene-environment observations. For example, in post-traumatic stress disorder (PTSD), it has been shown that FKBP5, a stress response regulator, is more likely to be "demethylated" (i.e. the DNA is without a "methyl group" attached) in children exposed to trauma. In this case, therefore, the underlying mechanism for this gene-environment

interaction turns out to be epigenetic. This change persists, enhancing expression of FKBP5, causing higher risk of developing PTSD in adulthood [36].

3.4.6 The Microbiome

It has long been known that our gut, skin, and mucosal membranes are host to an entire ecosystem of microbial organisms that live with us in a symbiotic fashion. Less commonly recognized is the fact that human beings carry around approximately ten times as many microbial cells as human ones [37]. In attempting to target therapies to the individual, it would be ill-advised to ignore such a significant contributor to our constitution. The Human Microbiome Project is a large-scale NIH-funded initiative with the goal of characterizing microbial communities found on several different sites of the human body. Ultimately this effort seeks to understand the relationship between diseases and changes in the human microbiome.

The microbiome has been shown to play a role in a wide range of diseases, including autism, depression, inflammatory bowel disease, type 1 diabetes, and various other autoimmune related illnesses [38]. Again, new sequencing techniques have enabled deeper exploration of these microbial communities, fostering a greater understanding of the make-up of these microbial ecosystems, the consequences of microbial imbalance, and potential therapeutic interventions to restore a healthy population. As one somewhat surprising example, a procedure known as fecal microbiota transplant, or FMT, has been performed with great success for the treatment of *C. difficile* infection [39]. Medicine does not get much more personal than that.

3.4.7 Clinical Decision Support

The human brain has a finite capacity to integrate different data sources [40]. This fact can be limiting even in the traditional practice of medicine, as it has been performed for decades. Increasing the number of facts or parameters that must be taken into account to inform health care decision making only serves to exacerbate this problem. The introduction of novel, high dimensional data types into clinical care (Fig. 3.7) takes us down this path.

Informatics, more specifically bioinformatics, introduces the challenge of a data deluge. Informatics also helps address this challenge in the form of clinical decision support (CDS). The information technology system underlying electronic health records can effectively integrate dozens, hundreds, even thousands of data points to make a recommendation regarding therapeutic decisions. Novel visualization techniques help to deliver these recommendations, along with the underlying reasoning, to time-constrained clinicians at the point of care.

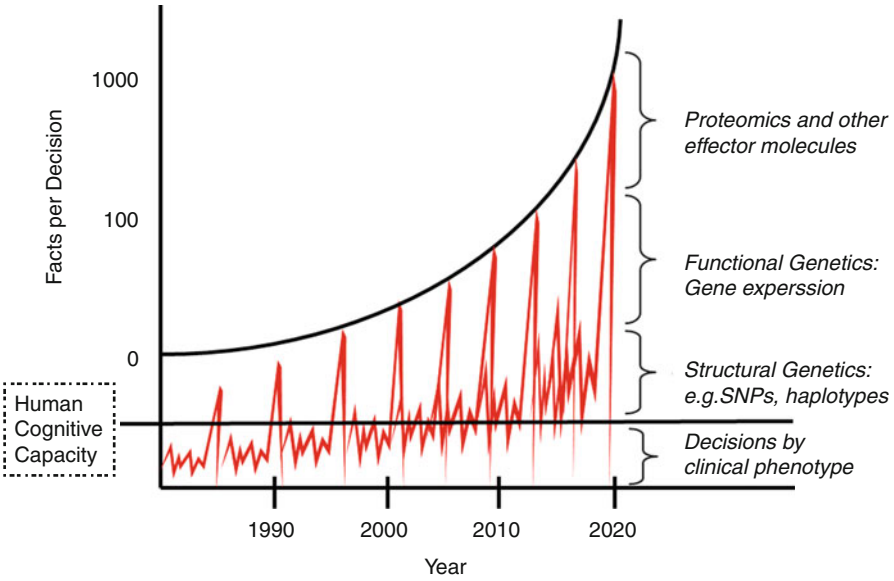

Fig. 3.7 The human mind is only able to incorporate a limited number of data points in decision making. High dimensional data in biomedicine increasingly exceeds that cognitive limitation. Electronic clinical decision support is required to address this problem (Used with permission from Masys D, personal communication 2014)

3.5 Discussion

In the preceding sections, we have described what is meant by the term "personalized medicine" (and its many variations), and have presented a number of hot topics within the field. As suggested above, these topics have both necessitated and guided the development of a number of subfields within informatics. Below we discuss the implications of these topics and their accompanying informatics advances on areas outside of research and the delivery of health care *per se*.

3.5.1 Economic Issues

Ultimately, personalized medicine should help drive down the cost of health care. By focusing on the right treatment for the right patient at the right time, fewer resources are wasted on ineffective treatments, and fewer costly treatments are required due to early intervention and prevention of disease in the first place. But a number of caveats must be considered in the context of reimbursement. First, in the era of personalized medicine, reimbursement will be required not only for the treatment, but also up front for various different diagnostic tests. Justification for that

expense requires significant evidence that outcomes are improved. As an example, CMS will currently only approve genetic testing for warfarin metabolism if the testing is performed in the context of a clinical study. In addition, all of the benefits described above apply *on average*. Individual cases may raise ethical quandaries. As an example, consider pharmacogenomics. If a terminally ill patient has a 50 % chance of responding to a drug, should insurance cover it? What about 5 %? 0.05 %? What if another drug is available, but causes more unpleasant side effects? What if the patient in question is your child?

One stakeholder for whom personalized medicine is mixed news is the pharmaceutical industry. On one hand, stratification of the patient population makes blockbuster drugs less likely. Drug companies stand to make greater profits when a drug is prescribed across the largest possible number of people. On the flip side, diagnostic companion tests may enable FDA approval for whole classes of drugs that would not have been seen as successful across the population. In some cases, a test may help rule out patients who are likely to have adverse events. In other cases, the test may help single out people who are particularly likely to respond well to a drug. Focusing on the right segment of the population can make clear the benefits of a drug that, on average, would not outperform the current standard of care.

3.5.2 Ethical, Legal, and Social Issues

3.5.2.1 Data Sharing

From the data generation and policy perspective, personalized medicine necessitates "big data" approaches. The sheer number of variables means that larger numbers of subjects (i.e. "bigger N") are required to support scientific conclusions. In light of the need for larger N, data sharing and re-use becomes increasingly important. Research funding cannot sustain the sample size that is required in the increasingly high-dimensional domain of biomedical research [41]. Those generating data must share the data in a manner that makes them accessible and comprehensible to those who would use them to further biomedical discovery. Makers of policy, including publishers and funders, have already established a number of guidelines for good data sharing practices. And increasingly these policies are actually be enforced.

Of course, a mandate to share data necessitates the creation of somewhere to put the data. Publicly available databases are also needed as repositories for researchers to deposit and access data. Free, unobstructed access to well-annotated, high quality data enables collaborations and data re-use and reduces obstacles to research. dbGap and TCGA (The Cancer Genome Atlas) are two examples of such repositories, though arguably there is some room for improvement in ease of use of the interfaces for data access [42]. In addition, after researchers demonstrated that it was possible to identify the presence or absence of an individual in a complex mixture of DNA samples, the NIH limited access to GWAS data to eligible researchers who are

required to apply for access. On the plus side for data access, there has been a recent movement to enable people who want to share their data to do so. Sicker patients tend to be less concerned with data privacy and more concerned that their data be made available to anyone working on a potential cure. Initiatives like the Portable Legal Consent [43] and 23andWe [44] aim to empower individuals to share their personal health data in meaningful ways.

3.5.2.2 Data Privacy

The unintended consequence of increased data sharing (genomic and otherwise) is an increased chance of a data breach, and of this information becoming available to those who would use it against the individual from whom it was derived. One often cited example is an insurance company refusing to cover someone with increased risk of a given disease, or charging exorbitant rates for such coverage. GINA, the Genetic Information Nondiscrimination Act, was passed in 2008 to address some of these issues, but it only covers employment and healthcare insurance. It does not extend to life, disability, or long term care insurance [45]. Another more dramatic scenario, as offered up in the consent form for the Personal Genome Project [46], is that if a person's sequence is known, it could in theory be used to create artificial DNA to be planted at the scene of a crime.

The 18th HIPAA identifier is defined as "Any other unique identifying number, characteristic, or code, unless otherwise permitted by the Privacy Rule for re-identification" [47]. And yet, as of early 2014, the US Department of Health and Human Services has not issued clear guidance on how the HIPAA Privacy Rule applies to genetic data nor whether DNA sequence is considered an identifier under HIPAA [48].

3.5.2.3 Return of Results

For a patient or research participant who has his or her genome sequenced, should that information be provided back to the individual? Surely a person should have access to his or her own data if desired. On the other hand, so much of the data is of questionable clinical utility or part of the "incidentalome," analogous to incidental findings in radiology [49]. Undiscovered, these traits would remain benign. Instead, their observance can lead to costly tests, emotional distress, or worse. Will providing hints of uncertain risk do more harm than good in causing mental or emotional stress? While most findings from DTC testing are benign, some are not. For potentially troubling results such as ApoE carrier status or BRCA mutations, 23andMe does not show these results along with the others but requires the user to click an extra link to "unlock" those results. It is appropriate to enable people to get their genomic data without the involvement of a clinician? Are people ready to learn about a life-altering genetic mutation over the internet? Conversely, is it overly paternalistic to think people should have their own data kept from them? This is an area of ongoing investigation, but studies to date regarding people's

ability to handle genetic bad news without long-term distress give cause for optimism [50, 51].

The American College of Medical Genetics and Genomics (ACMG) issued a set of guidelines in 2013 regarding incidental findings [52]. The guidelines recommend that disease-causing variants from 57 known genes, related to 24 disorders, be returned to the ordering physician regardless of patient preference. Various commentaries expressed concern over these recommendations, leading to a clarification statement on the ACMG's website which emphasized, among other things, that variants of unknown significance were not included in the recommended list. Rather, the list included only a set of variants for which there is "significant potential for preventing disease morbidity and mortality if identified in the presymptomatic period" [53].

But there are other questions that the guidelines do not address. What are the researcher's or health care provider's obligations to follow up with the individual as new knowledge accrues? If a person's genotype is of unknown significance today, but understood and actionable 10 years later, should he be made aware? What if the significance is known and dire, but *not* actionable? These will surely be topics of debate in the coming years.

3.5.3 Training

As biology and biomedicine become increasingly data driven, the curricula for training both biomedical researchers and clinicians will need to include more quantitative components. Basic statistics and computer science will be mandatory. Researchers with some basic programming skills will be at a significant advantage, both for basic data formatting tasks and use of common software packages such as R and Bioconductor. Perhaps the most basic, but also the most important, addition to training will be increasing the skill of *numeracy*, that is, the ability to reason with and apply numerical concepts. Researchers, care providers, and patients alike will all benefit from greater numeracy, enabling, for example, interpretation of probabilities.

On the clinical side, recent surveys have shown that only slightly more than half of primary care physicians report feeling confident in interpreting genetic test results, and 20 % report having had no genetics education [54]. In addition to quantitative sciences, medical training programs will need to make genetics, biomarkers, and other "molecular medicine" courses part of their core curricula in order to meet the needs of clinicians in the post-genome era.

3.5.4 Participatory Medicine

Patients need not just sit back and wait for the benefits to accrue. There are a number of ways in which they can maximize the likelihood that they themselves will see the benefit of personalized medicine. First and foremost, patients can make sure they

are informed. So much information is available online—individuals can learn about everything from the human genome to mechanisms of disease to how to interpret a *p*-value. Reputable medical websites such as WebMD[1] and the Mayo Clinic[2] offer information across a broad set of conditions. In addition, many disease-centric societies offer information to patients and their families. While one must exercise common sense and restraint, not believing everything one reads on the internet, and not bombarding a time-constrained physician with reams of internet-derived medical wisdom, better informed patients will be better able to ask the right questions and to have a meaningful dialog with their care providers. This, in turn, will enable the patient to feel more empowered and to play an active role in formulating a treatment plan. Cancer survivor, Dave deBronkart, also known as "ePatientDave," has taken participatory medicine to a new level with his website, blog, social media presence, and keynote presentations on patient engagement, with particular emphasis on HealthIT.[3] Taking a broader view, patient communities can advocate for causes that increase the likelihood, and speed, that personalized medicine will become a reality, including research funding, data sharing, and science education.

3.6 Implications for Stakeholders

It can be seen that each of the different stakeholders described in Chap. 2 benefits from realization of the vision of personalized medicine: physicians provide better care, patients are healthier and receive better care when they do get sick, and payers get more for their money. But there are other implications as well. Key among those are:

Evidence and Policy Generators

- Researchers must continue to **develop and master new technologies and statistical techniques** to generate new data types, convert that data into information, and extract new biomedical knowledge from the information, informing new guidelines for clinical care.
- Policy makers must **consider ethical**, **legal**, **social**, **and economic issues** around the new capabilities that are enabled through personalized medicine and translational bioinformatics.

Providers and Healthcare Organizations

- Medicine is poised to shift away from the macroscopic classifications that have been employed for centuries toward a more **precise and personalized approach to health and disease**.

[1] http://www.webmd.com/

[2] http://www.mayoclinic.com/health-information/

[3] http://www.epatientdave.com/

- As the number of factors affecting clinical decisions increases, **information technology and the clinical decision support** it can provide becomes increasingly important.

Patients and Their Communities

- Patients and their advocates can help clinicians do a better job providing care by being **informed and engaged**.

Finally, across all stakeholders but particularly clinicians and researchers, training programs will need to evolve to increase general **knowledge in genetics**, **computer science**, **and quantitative methods**.

3.7 Conclusions

This is a very exciting time for the field of personalized medicine, and informatics plays a critical part in both enabling its achievements and addressing its challenges. Ongoing advances in biomedical techniques, health information technology, and informatics methodologies will continue to accelerate progress in this important area of translational research and clinical practice. Ultimately, realization of personalized medicine will benefit stakeholders across the biomedical enterprise.

Discussion Points

- How does personalized medicine differ from how medicine has been practiced for decades? Haven't doctors always taken the specific individual into account? What has changed?
- How can personalized medicine reduce health care costs even as more treatments are being discovered? How could it increase them? What new costs could be introduced that did not apply in the past?
- What are the respective wins from personalized medicine for the various stakeholders? What are some potential challenges it raises? Consider health, economics, ethics, and workforce training.
- How can personalized medicine facilitate patient engagement? What are some ways in which patients can help realize the participatory aspect of P4 medicine?
- What are some other potential future directions for personalized medicine?

References

1. Burgoon LD, Boutros PC, Dere E, Zacharewski TR. dbZach: a MIAME-compliant toxicogenomic supportive relational database. Toxicol Sci. 2006;90(2):558–68.
2. Zimmermann P, Schildknecht B, Craigon D, Garcia-Hernandez M, Gruissem W, May S, et al. MIAME/Plant – adding value to plant microarrray experiments. Plant Methods. 2006;2:1.
3. Davies K. The $1000 genome: the revolution in DNA sequencing and the new era of personalized medicine. 1st ed. New York: Free Press; 2010.

4. NHGRI. DNA sequencing costs. 2013. [9/17/2013]. Available from: http://www.genome.gov/sequencingcosts/.
5. Herper M. The $1,000 genome arrives – for real, this time. Forbes [serial on the Internet]. 2014. Available from: http://www.forbes.com/sites/matthewherper/2014/01/14/the-1000-genome-arrives-for-real-this-time/.
6. Schadt EE, Turner S, Kasarskis A. A window into third-generation sequencing. Hum Mol Genet. 2010;19(R2):R227–40.
7. NRC. Toward precision medicine: building a knowledge network for biomedical research and a new taxonomy of disease. 2011.
8. Hood L. Systems biology and p4 medicine: past, present, and future. Rambam Maimonides Med J. 2013;4(2):e0012.
9. Evans WE, Relling MV. Moving towards individualized medicine with pharmacogenomics. Nature. 2004;429(6990):464–8.
10. Burrill GS. Where's the beef? Drug Discov. 2003;4:9.
11. Green ED, Guyer MS. Charting a course for genomic medicine from base pairs to bedside. Nature. 2011;470(7333):204–13.
12. Trusheim MR, Berndt ER, Douglas FL. Stratified medicine: strategic and economic implications of combining drugs and clinical biomarkers. Nat Rev Drug Discov. 2007;6(4):287–93.
13. Shortliffe EH, Cimino JJ. Medical informatics: computer applications in health care and biomedicine. London: Springer; 2013.
14. Killion PJ, Sherlock G, Iyer VR. The Longhorn Array Database (LAD): an open-source, MIAME compliant implementation of the Stanford Microarray Database (SMD). BMC Bioinforma. 2003;4:32.
15. Nuzzo R. Scientific method: statistical errors. Nature. 2014;506(7487):150–2.
16. CDC. [9/10/2013]. Available from: http://www.cdc.gov/genomics/gtesting/ACCE/.
17. Ginsburg GS, Woods CW. The host response to infection: advancing a novel diagnostic paradigm. Crit Care. 2012;16(6):168.
18. Hecker, Michael, et al. Reassessment of blood gene expression markers for the prognosis of relapsing-remitting multiple sclerosis. PloS one. 2011;6(12):e29648.
19. Kraus VB, Burnett B, Coindreau J, Cottrell S, Eyre D, Gendreau M, et al. Application of biomarkers in the development of drugs intended for the treatment of osteoarthritis. Osteoarthr Cartil. 2011;19(5):515–42.
20. Voils CI, Coffman CJ, Edelman D, Maciejewski ML, Grubber JM, Sadeghpour A, et al. Examining the impact of genetic testing for type 2 diabetes on health behaviors: study protocol for a randomized controlled trial. Trials. 2012;13:121.
21. Butte AJ. Translational bioinformatics: coming of age. J Am Med Inform Assoc. 2008;15(6):709–14.
22. Cirulli ET, Goldstein DB. Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nat Rev Genet. 2010;11(6):415–25.
23. Pirmohamed M, Burnside G, Eriksson N, Jorgensen AL, Toh CH, Nicholson T, et al. A randomized trial of genotype-guided dosing of warfarin. N Engl J Med. 2013;369(24):2294–303.
24. Kimmel SE, French B, Kasner SE, Johnson JA, Anderson JL, Gage BF, et al. A pharmacogenetic versus a clinical algorithm for warfarin dosing. N Engl J Med. 2013;369(24):2283–93.
25. Sermet-Gaudelus I. Ivacaftor treatment in patients with cystic fibrosis and the G551D-CFTR mutation. Eur Respir Rev. 2013;22(127):66–71.
26. O'Bryant CL, Wenger SD, Kim M, Thompson LA. Crizotinib: a new treatment option for ALK-positive non-small cell lung cancer. Ann Pharmacother. 2013;47(2):189–97.
27. Bollag G, Tsai J, Zhang J, Zhang C, Ibrahim P, Nolop K, et al. Vemurafenib: the first drug approved for BRAF-mutant cancer. Nat Rev Drug Discov. 2012;11(11):873–86.
28. Annas GJ, Elias S. 23andMe and the FDA. N Engl J Med. 2014;370(11):985–8.
29. Johnson M, Gallagher K. Living on the edge of science. Journal Sentinel. 2012. 6/30/2012.
30. Bainbridge MN, Wiszniewski W, Murdock DR, Friedman J, Gonzaga-Jauregui C, Newsham I, et al. Whole-genome sequencing for optimized patient management. Sci Transl Med. 2011;3(87):87re3.

31. Kolata G. A new treatment's tantalizing promise brings heartbreaking ups and downs. New York Times. 2012. 7/8/2012.
32. Kolata G. In treatment for leukemia, glimpses of the future. New York Times. 2012. 7/7/12.
33. Goldberg AD, Allis CD, Bernstein E. Epigenetics: a landscape takes shape. Cell. 2007;128(4):635–8.
34. Madrigano J, Baccarelli A, Mittleman MA, Sparrow D, Vokonas PS, Tarantini L, et al. Aging and epigenetics: longitudinal changes in gene-specific DNA methylation. Epigenetics. 2012; 7(1):63–70.
35. Dawson MA, Kouzarides T. Cancer epigenetics: from mechanism to therapy. Cell. 2012; 150(1):12–27.
36. Klengel T, Mehta D, Anacker C, Rex-Haffner M, Pruessner JC, Pariante CM, et al. Allele-specific FKBP5 DNA demethylation mediates gene-childhood trauma interactions. Nat Neurosci. 2013;16(1):33–41.
37. Zimmer C. How microbes defend and define us. The New York Times. 2010 July 12. 2010.
38. de Vos WM, de Vos EA. Role of the intestinal microbiome in health and disease: from correlation to causation. Nutr Rev. 2012;70 Suppl 1:S45–56.
39. Kassam Z, Lee CH, Yuan Y, Hunt RH. Fecal microbiota transplantation for Clostridium difficile infection: systematic review and meta-analysis. Am J Gastroenterol. 2013;108(4): 500–8.
40. Stead WW, Searle JR, Fessler HE, Smith JW, Shortliffe EH. Biomedical informatics: changing what physicians need to know and how they learn. Acad Med. 2011;86(4):429–34.
41. Ginsburg GS, Staples J, Abernethy AP. Academic medical centers: ripe for rapid-learning personalized health care. Sci Transl Med. 2011;3(101):101cm27.
42. Doan S, Lin KW, Conway M, Ohno-Machado L, Hsieh A, Feupe SF, et al. PhenDisco: phenotype discovery system for the database of genotypes and phenotypes. J Am Med Inform Assoc. 2014; 21(1):31–6. doi: 10.1136/amiajnl-2013-001882. PubMed PMID: 23989082; PubMed Central PMCID: PMC3912702
43. weconsent.us. Consent to Research. [9/11/2013]. Available from: http://weconsent.us/donate-your-data/data-donation-faq/.
44. 23andMe. 23andWe Research. [9/11/2013]. Available from: https://www.23andme.com/research/.
45. Hudson KL, Holohan MK, Collins FS. Keeping pace with the times–the Genetic Information Nondiscrimination Act of 2008. N Engl J Med. 2008;358(25):2661–3.
46. Church GM. The personal genome project. Mol Syst Biol. 2005;1:2005.0030.
47. NIH. HIPAA privacy rule and its impact on research. [8/27/2013]. Available from: http://privacyruleandresearch.nih.gov/pr_08.asp.
48. HIPAA, the privacy rule, and its application to health research. In: Nass SJ LL, Gostin LO, editor. Beyond the HIPAA privacy rule: enhancing privacy, improving health through research. Washington, DC: National Academies Press; 2009.
49. Kohane IS, Masys DR, Altman RB. The incidentalome: a threat to genomic medicine. JAMA. 2006;296(2):212–5.
50. Green RC, Roberts JS, Cupples LA, Relkin NR, Whitehouse PJ, Brown T, et al. Disclosure of APOE genotype for risk of Alzheimer's disease. N Engl J Med. 2009;361(3):245–54.
51. Bloss CS, Schork NJ, Topol EJ. Effect of direct-to-consumer genomewide profiling to assess disease risk. N Engl J Med. 2011;364(6):524–34.
52. Green RC, Berg JS, Grody WW, Kalia SS, Korf BR, Martin CL, et al. ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. 2013; 15(7):565–74. doi: 10.1038/gim.2013.73. Epub 2013 Jun 20. PubMed PMID: 23788249; PubMed Central PMCID: PMC3727274.
53. ACMG Board of Directors. Points to consider in the clinical application of genomic sequencing. Genet Med. 2012;14(8):759–61. doi: 10.1038/gim.2012.74. PubMed PMID: 22863877.
54. Bernhardt BA, Zayac C, Gordon ES, Wawak L, Pyeritz RE, Gollust SE. Incorporating direct-to-consumer genomic information into patient care: attitudes and experiences of primary care physicians. Per Med. 2012;9(7):683–92.

Additional Reading

Altman RB. Personal genomic measurements: the opportunity for information integration. Clin Pharmacol Ther. 2013;93(1):21–3.
Angrist M. Here is a human being: at the dawn of personal genomics. New York: Harper; 2010.
Ashley EA, Butte AJ, et al. Clinical assessment incorporating a personal genome. Lancet. 2010;375(9725):1525–35.
Butte AJ. Translational bioinformatics: coming of age. J Am Med Inform Assoc. 2008;15(6): 709–14.
Butte AJ, Ito S. Translational bioinformatics: data-driven drug discovery and development. Clin Pharmacol Ther. 2012;91(6):949–52.
Butte AJ, Ohno-Machado L. Making it personal: translational bioinformatics. J Am Med Inform Assoc. 2013;20(4):595–6.
Chen R, Mias GI, et al. Personal omics profiling reveals dynamic molecular and medical phenotypes. Cell. 2012;148(6):1293–307.
Davies K. The $1000 genome: the the revolution in DNA sequencing and the new era of personalized medicine. New York: Free Press; 2010.
Fernald GH, Capriotti E, Daneshjou R, Karczewski KJ, Altman RB. Bioinformatics challenges for personalized medicine. Bioinformatics. 2011;27(13):1741–8.
Ginsburg GS, Willard HF. Genomic and personalized medicine. 2nd ed. London: Elsevier/ Academic; 2012.
Hamburg MA, Collins FS. The path to personalized medicine. N Engl J Med. 2010;363(4):301–4.
Hood L, Friend SH. Predictive, personalized, preventive, participatory (P4) cancer medicine. Nat Rev Clin Oncol. 2011;8(3):184–7.
Overby CL, Tarczy-Hornoch P. Personalized medicine: challenges and opportunities for translational bioinformatics. Per Med. 2013;10(5):453–62.
Sarkar IN. Biomedical informatics and translational medicine. J Transl Med. 2010;8:22.
Sarkar IN, Butte AJ, et al. Translational bioinformatics: linking knowledge across biological and clinical realms. J Am Med Inform Assoc. 2011;18(4):354–7.
Whirl-Carrillo M, McDonagh EM, et al. Pharmacogenomics knowledge for personalized medicine. Clin Pharmacol Ther. 2012;92(4):414–7.

Chapter 4
Leveraging Electronic Health Records for Phenotyping

Adam B. Wilcox

By the End of This Chapter, Readers Should Be Able to:

- Understand phenotyping
- Understand how data from electronic health records (EHRs) can be used for phenotyping
- Identify the key challenges to using EHR data for phenotyping
- Understand current changes that can affect phenotyping from EHR data

4.1 Introduction

Arguably the greatest advancements in biomedicine over the last few decades have been in genetics. From the mapping of the human genome to the development of genetic testing to research biobanks, genomic science has advanced to be a dominant field in biomedical research. Today, research labs across the country are constantly mining the genetic code of thousands of individuals, identifying associations that can change both the way health care is delivered and how disease is prevented. The first human genome was sequenced at a cost of nearly $3 billion [1] – in a few years, scientists expect the cost to drop to around $100. This opens opportunities for even more genetic studies, which will only increase the impact of genetics on how we understand health.

Currently, among the most common studies using genetics are genome-wide association studies (GWAS). These are done by taking a large number of sequenced genomes of patients along with specific characteristics of the patients, and

A.B. Wilcox, PhD
Department of Medical Informatics, Intermountain Healthcare,
Murray, UT, USA
e-mail: adam.wilcox@imail.org

P.R.O. Payne, P.J. Embi (eds.), *Translational Informatics: Realizing the Promise of Knowledge-Driven Healthcare*, Health Informatics,
DOI 10.1007/978-1-4471-4646-9_4,

identifying how the patients' different genes are related to their characteristics. The set of characteristics, or phenomena, for an individual is called a phenotype. Linking genotypes and phenotypes has significantly accelerated genetic discovery. At first, the phenotypes were collected like other research studies when the biosamples were collected: first obtaining consent from the patient, then asking a series of questions to define the phenotype, collecting the sample, and matching it to the phenotype. But this approach was both slow and zexpensive, especially since a large number of subjects need to be included for genotype-phenotype studies. Innovations in both genetic sequencing and consent processes have increased the genotype collection. For phenotypes, the greatest innovation has been to extract the information from data already collected as part of the clinical care process – from electronic health records (EHRs).

The ability to leverage electronic information in health records for genetic research is an excellent example of translational informatics and its influence in the future of biomedical research and healthcare. In this chapter, we discuss the importance of EHRs for creating phenotypes and how they can be used. First we describe a brief history of its use, and recent influences that are affecting its current interest. We review examples of projects that are successfully leveraging EHRs for phenotyping, identifying both their successes and challenges. Finally, we discuss the future of phenotype extraction from EHRs, and the impact on genetic research as well as other health care research domains.

4.2 History of Secondary Use of Electronic Health Data

The idea of using EHR data beyond clinical care is not new. Health care is an information-intensive field, and generates large amounts of information. For decades, researchers have been recommending their use in research studies. Many of the Patient Outcomes Research Teams created by the Agency for Health Care Policy and Research specifically used data from medical records, and identified the importance of using medical records rather than claims and billing information [2, 3]. Early informatics researchers were successful in both demonstrating that EHR data can be valuable to research and identifying the challenges inherent in using it [4–7]. Even when very few health care institutions were collecting electronic health data, the use of electronic health data for secondary use in research was pursued. This interest only increased as data mining and data warehousing in medical databases grew in the late 1990s. The increased use in the 2000s of electronic research databases and the desire to both populate these systems with existing clinical data and facilitate cohort identification and selection [8, 9] demonstrated secondary data use as a core part of the emerging discipline of translational informatics.

Perhaps nothing has been as significant in increasing the need for EHR data for research as genomics in the last few years [10]. A main reason was the emerging need for genome-wide association studies (GWAS). Initial genetic discoveries were from family-based studies. Researchers were studying diseases with a strong

hereditary link and rare genetic variants [11, 12]. But there are also many common disorders caused by many common genetic variants, with weaker links. New types of studies other than family-based studies were needed for common genetic variants. GWAS are population-based studies of common variants with small effects [13]. GWAS have also been used successfully to identify variations in patients' responses to different medications, leading to personalized medicine.

One thing that makes GWAS particularly interesting is that the genotype, once it has been sequenced, is available for studying multiple associations. A challenge is to then have a large enough sample to do population-based studies on common diseases [14]. Various institutions have created biobanks to address this, where biologic samples are collected for use in GWAS studies, many with tens of thousands of samples [15]. The last decade has seen a tremendous growth in sequencing ability, with a corresponding reduction in costs. EHRs can then be used as a source of phenotypic information for genotype-phenotype studies like GWAS.

The emergence of GWAS has changed the potential of using EHRs for research data for many reasons. First, the large number of subjects required for GWAS has increased the overall benefit of using EHRs. When smaller numbers of subjects were needed, alternatives to phenotype extraction from EHRs could be an acceptable option. Second, successes in GWAS have been both significant and rapid. Since the first GWAS study in 2005, thousands of GWAS studies have examined hundreds of traits and diseases [16]. The results and interest in GWAS, and the subsequent need for phenotypic data to expand it, has been explosive. Third, successful GWAS studies have included development of an infrastructure that can be more easily leveraged for successive studies. The creation of biorepositories is comparatively more difficult than extraction methods from electronic health records. Once the biorepository is created, however, its marginal cost of use drops rapidly especially for genotyped samples. The comparative cost-benefit of using EHR phenotypes makes it a more worthwhile pursuit.

Initial demonstrations using EHRs for research has only expanded their perceived potential. Beyond GWAS, they are seen as a method to rapidly identify variables and outcomes of cohorts that can then be applied to similar patients under treatment [17]. Comparative effectiveness research has also increased the demand for using EHR data in research [18]. Effectiveness research focuses on studying interventions in the "real world," or the environment where the interventions would be most likely to be received. This is in contrast to clinical trials that have actively limited the environment of the trial to study efficacy of an intervention, without confounders. EHRs collect data in the world where care is provided, so their data are more relevant to effectiveness studies than data collected in case report forms during efficacy studies. The learning health system is focused on effectiveness, and the use of EHR data to support it is fundamental to its vision [17]. Phenotype extraction from EHRs has become a critical requirement of both research and health care transformation.

Understanding the history of leveraging EHRs for research is important to recognize both the potential benefits and challenges. It has been pursued for a long time, with great successes and even greater potential. Some of the most significant early

successes in genetic studies and clinical informatics arose from secondary data use. Researchers in the discovery of BRCA1 used the Utah Population Database, which includes familial histories linked to data extracted from electronic health databases [11]. Evans et al. demonstrated one of the first examples of a learning health system with the Antibiotic Assistant, which used data in the EHR from similar patients to compute recommendations for antibiotic prescribing [19]. These studies were successful in part because they used very specific data for defined purposes among defined populations. As more data were collected and the data types, use and scope broadened, researchers began to face limitations in secondary use of EHR data. Early research using EHR data for improving adverse drug events (ADEs) showed that structured data were incomplete and underestimated the number of ADEs[20]. Studies in data mining were also limited by methods extracting data from EHRs [21]. Issues of data quality and completeness have continued to challenge EHR data reuse [22, 23]. For many years, the early potential was unmet as researchers identified and faced multiple barriers, even while the collection of data was increasing [6].

4.3 Recent Developments

Now that we have established how the use of EHR data for phenotypes has emerged over time, and why it is interesting now, we can discuss how it is done more clearly.

4.3.1 Extracting Data from EHRs

EHRs contain various types of data that are used for various purposes. Some data are collected primarily for administrative purposes. Demographics information is needed to identify an individual, both for treatment and payment. Diagnosis data indicate the overall condition of a patient, and are generally collected for billing. Procedure data indicate the various actions taken by clinicians, and are also used for billing. Because demographics and billing data are most commonly and consistently collected for patients, they have been a primary source for phenotype information in population research [24]. However, because they are used primarily for non-clinical care, researchers have observed inconsistencies and errors in billing data [2]. Other data are collected primarily to support clinical care. These include medications prescribed, assessments made, tests and activities ordered, results of tests, actions performed, and statements of clinical judgment. Laboratory test results and standard patient assessments (e.g., vital signs) are generally stored in structured form, and can be used to interpret phenotypes for diseases that are specifically indicated by test results [25]. Medication orders and prescriptions are also often structured, but may also be stored as part of unstructured text (e.g., medication history). Medications can also reveal patient clinical conditions by what is being treated [26]. The richest source of clinical information is usually stored as unstructured text in

clinical documentation. This is the most clinically relevant information, representing what was clinically important, but is also the most difficult to extract. Narrative text allows a high degree of expressiveness and flexibility, but this same flexibility makes it difficult to extract information from the records on a large scale. Researchers have for years been refining approaches to natural language processing (NLP) to extract information from narrative text [27], and more recently have been applying this to records specifically for phenotype extraction [28].

Because of the multiple data types in the electronic health record that require different methods of extraction for phenotype representation, a significant amount of research in using EHR data for phenotypes has been done just on extracting data from electronic health records. The SHARPn project, for example, was funded specifically to determine and demonstrate best approaches for extracting data from EHRs for secondary data analysis [28]. Research in medical language processing for extracting phenotypic information has grown substantially to be a significant focus of the field [29]. And multiple initiatives have emerged that focus on defining different phenotypes that can be extracted from the different data sources from EHRs [30, 31]. Usually, the actual extraction algorithm is a set of rules that query for data from different data sources. For example, a diabetes phenotype extraction algorithm is a combination of administrative visit data, laboratory results, diagnosis codes, prescribed medications, and family history data from narrative text [32].

4.3.2 Performing Genome-Wide Association Studies

As mentioned above, when researchers can successfully extract disease phenotypes from EHR data, they can use this information to perform GWAS. GWAS analyze a large number of genotypes and matched phenotypes. Genotypes must be sequenced from biological samples, so the collection of biospecimens determines what genotypes can be done. Currently most genotyping is done through chip-based microarray techniques that identify millions of markers on the genetic code for one individual, but it is anticipated that future sequencing techniques will provide the full DNA sequence within a few years. Currently the markers used in the genotype are single nucleotide polymorphisms, or SNPs. SNPs are small changes in the DNA sequence that occur relatively frequently in the human genome. They typically do not have substantial impact on biological processes, but are helpful for marking genetic variation among individuals. Regardless of the method of genotyping, it is critical that the sample and genotype be matched to an identified subject, so that a matching phenotype can be queried from the data in the EHR [13].

The order of the two tasks (genotyping or phenotyping) is less important, as long as a genotype is linked to a phenotype for the same patient. The methods of the study will often dictate which must be done first based on dependencies. In some cases, the genotype is first collected from all biospecimens in a population of subjects who also have data in EHRs. Then extraction rules for a specific phenotype of interest can be developed, validated and used to query that phenotype for the subjects from

the EHR. In other cases, a large sample of biospecimens is collected first, but is not genotyped. Extraction rules for a specific phenotype are developed, and this is run against the EHR for all the subjects in the biobank. Cases and controls are then selected as a subset based on the matching phenotypes, and then biospecimens from those specific subjects are genotyped. This is a cost-savings approach, because only a sub-population requires genotyping, which is typically the most costly component for the research study.

Genome-wide association (GWA) analysis can occur once the genotypes and phenotypes are created and linked. Analysis is often done with standard statistical testing, such as analysis of variance, contingency tests, or regression analyses. More advanced testing is used to account for covariates. More complicated testing is done when the interest is not just single gene variations, but for interactions among different genes. These multi-locus analyses quickly become computationally difficult, but various methods have been successfully used to filter SNPs for analysis. The results of the analyses are identification of significant associations between gene markers or combinations of genes and specific phenotypes. Like all scientific studies of significance, the result can be further validated by replicated tests, until a recognized association for a trait is made, and a genetic test can then be developed to mark an individual's specific risk for that condition [13].

4.3.3 Pharmacogenetics and Personalized Medicine

A specific example of applying genotype-phenotype analysis from EHRs is in discovering medication efficacy, or pharmacogenetics. Bush and Moore [13] give a good example about warfarin, a blood-thinning medication that helps prevent clots in patients at risk of an embolism. Administering anticoagulation therapy to patients is a delicate process, where the right dose needs to be determined and used. If too low a dose is used, the medication will not prevent potentially fatal clots; if too high a dose is used, the blood can become too thin and the patient risks dangerous internal and external bleeding. When administering anticoagulation therapy, clinicians must carefully watch the patient's clotting activity, or prothrombin time, because warfarin has a very narrow therapeutic window. A GWAS has shown that there is also wide variation due to genetics in a patient's response to warfarin – in some populations greater than any other known factor [33]. Phenotype data for this study were collected from electronic health records, including warfarin doses for the patients, lab tests indicating clotting activity, and demographics. Linear regression was then used to analyze the data. The result of the GWAS is the development of a genetic test for patients to determine their appropriate safe warfarin doses. This type of test, designed to tailor the care provided to an individual based on her genotype, is personalized medicine, and represents an important translation of GWAS to actual patient care. A benefit of personalized medicine being developed from data extracted from electronic health records is that they can be more effective when used with computerized decision

support. Since the clinical data needed to apply the rule (e.g., the prescribing of warfarin and patient demographics) were already used in the discovery of the correlation, its application is simplified.

4.3.4 *Projects Leveraging EHRs for Genotype-Phenotype Studies*

Based on the early results and promise of using EHRs for extracting phenotype information for genetics research, the National Human Genome Research Institute (NHGRI) funded the Electronic Medical Records and Genomics Network, or eMERGE. The NHGRI began working with the International Human Genome Project, and continues to support genome research as one of the National Institutes of Health. eMERGE was created to specifically develop methods and practices for using EHRs for genomic research [15, 34]. It began as a network of five institutions, each with a biobank and an electronic health record. The size of the biorepositories at each site ranged from about 4,000 to 75,000. Each eMERGE site was focused on a particular primary and secondary phenotypic outcome with its subject population. eMERGE later expanded to include nine main research groups or institutions, and additional affiliate institutions.

eMERGE has been significant in advancing the understanding of capabilities and issues of using EHR data for phenotyping. They have published results of GWAS using phenotypes from EHRs in each of the primary conditions studied: cataract and HDL, dementia, electrocardiographic QRS duration, peripheral arterial disease, and type 2 diabetes. Their successes have furthered the interest in using EHRs for GWAS. They also were able to successfully validate and deploy phenotypes developed at one institution using one EHR to other institutions and EHRs across the network [35].

eMERGE has also increased understanding of issues related to using EHRs for phenotypes, that has extended beyond the goals of GWAS. They published results of studies demonstrating how privacy could be both breached and protected when performing research with data from EHRs for GWAS and other studies [36–38]. The timing of these analyses and results was significant – with breach penalties incurred from the HITECH act, many institutions had difficulty in navigating the new rules of privacy and confidentiality, while still sharing data. eMERGE researchers were also able to demonstrate how phenotypes from EHRs could be used for a new type of study beyond GWAS. Rather than scanning genotypes for associations with a defined phenotype, as is done with GWAS, they demonstrated scanning a large set of phenotypes for associations among various genotypes. This new approach created phenome-wide association studies, or PheWAS.

Another product of eMERGE has been resources for other researchers to use in GWAS or PheWAS research. Software to perform PheWAS was made available and distributed to researchers. eMERGE researchers created over 21 different phenotypes, that are made publically available in the Phenotype KnowledgeBase. These

concrete resources are in addition to the methods and lessons learned that were published in the scientific literature.

Other groups have also used EHRs for genotype-phenotype studies. The Pharmacogenomics Research Network (PGRN) includes sites that use EHR data for phenotypes, though the use of EHRs is not defining of the goals as is eMERGE. The Kaiser Permanente Research Program on Genes, Environment and Health used EHRs for phenotypic information linked to genotypes [26]. While not having the breadth of institutional participation or depth of use of EHR data, these projects demonstrate the interest in leveraging EHR data for genotype-phenotype research extends beyond the initial demonstration projects.

The Integrating Informatics and Biology at the Bedside (i2b2) is particularly important to leveraging EHR data for phenotypes because it provides a platform for storing, linking and querying data from EHRs directly by researchers. It also allows cross-institution queries for connected research institutions. Data extracted from EHRs are loaded into the i2b2 data model. Individuals can then create queries based on the organization of the clinical data. It allows linking of genotype information to support GWAS. It also includes a natural language processing engine to extract findings from text reports. Perhaps the greatest indication of importance for i2b2 is its broad use – it has been adopted by over 60 academic health centers internationally [39].

The Measurement to Understand the Reclassification of Disease of Cabarrus/Kannapolis (MURDOCK) Study in North Carolina represents a slightly different approach to creating an infrastructure for genotype-phenotype studies. It focuses on creating a comprehensive database of phenotype information along with biospecimens, and is non-disease specific. Rather than recruiting biospecimens at one time with consent to link to EHR data, MURDOCK enrolls the whole individual. Subjects consent to provide biospecimens, complete annual surveys, link data to their EHRs, and be contacted for participation in other studies. In addition, environmental and geospacial data are collected so that non-clinical factors can be measured for their impact on health and treatment effectiveness. The MURDOCK study is a good example demonstrating how the extraction of phenotype data from EHRs can be extended beyond the genotype, and to fully patient-centered and personalized research.

4.3.5 Projects Leveraging EHRs for Comparative Effectiveness

Some projects have focused on leveraging EHR data for secondary analysis in comparative effectiveness and patient-centered outcomes research, rather than specifically on GWAS or other genetic studies. In 2010, the Agency for Healthcare Research and Quality (AHRQ) funded multiple large projects designed to create data infrastructures using EHR data to perform comparative effectiveness research. Like the eMERGE network, each of these projects collected data for a specific population, focusing on a particular disease. Unlike eMERGE, they each included data

from multiple clinical sites, and addressed issues of merging data from different EHRs to increase both the breadth and depth of the data available. For example, the Scalable Architecture for Federated Translational Inquiries Network (SAFTINet) project in Colorado created a distributed research network across multiple sites of care in multiple states, allowing the creation of large cohorts of patients with EHR data. They also focused on diverse and underserved populations, and initially studied heart and blood vessel conditions, as well as breathing conditions [40]. The SUrveillance, PREvention, and ManagEment of Diabetes Mellitus (SUPREME-DM) project at the Kaiser Permanente Institute for Health Research leverages EHR data to create a longitudinal registry of diabetes care for 1.3 million patients across 11 geographically-disparate integrated delivery systems. They also studied other conditions related to the population, such as gestational diabetes, obesity, and heart conditions. Like SAFTINet, they used a distributed data network [41].

The Washington Heights/Inwood Informatics Infrastructure for Comparative Effectiveness Research (WICER) also leveraged EHR data to create patient phenotypes. Similar to MURDOCK but different than SAFTINet and SUPREME-DM, WICER used EHR data as one source among many for a specific population. The focus was an underserved, immigrant population in New York City. WICER linked data from EHRs of multiple providers (inpatient, ambulatory, home care) to create a comprehensive and longitudinal view of the clinical data for a patient. For a subset of the population, WICER also surveyed individuals at various locations on clinical measures, social attitudes and networks, and health behaviors and beliefs. They also collected biospecimens for some of the population. For patients with biospecimens, EHR data across sites and survey data, WICER created one of the most comprehensive datasets for a research population. By collecting data from multiple sources, WICER investigators are also able to study differences in data quality and completeness across different sources [42].

While eMERGE, i2b2 and other studies have advanced the understanding of EHR data for genotype-phenotype research, these AHRQ-funded studies have been critical in learning important lessons about using EHRs for cohort studies in a population. They have advanced understanding of data display and navigation, data security, primary data collection, distributed queries, research sustainability, and data quality [18, 22, 40, 42–44]. They have also each built an infrastructure that is now being extended for more analyses. These infrastructures demonstrate both the viability and potential of leveraging EHR data to define subjects for clinical research studies.

4.3.6 Challenges and Future Directions in Using EHR Data for Research

Along with these successes, however, are challenges. Secondary use of EHR data is more available and inexpensive than primary research collection, but since it is collected for a different purpose its quality for research is different. For example, EHR data is most complete on patients who have visits to health providers using the

EHR. If a patient is not being treated, the information is not collected. If the provider does not use the EHR, the information is not collected. If the information is not directly relevant to the care being provided, the information may not be collected by the provider. In contrast, with research-based case report forms the specific data elements of interest are defined prospectively, and the data are collected for each subject.

The real effect of these biases is still unknown. Researchers acknowledge that data quality could undermine the ability to use EHR data as a surrogate for primary research collection. Previous studies using billing data have had noted quality issues [2]. Currently, researchers with EHR data for phenotypes are investigating methods for assessing quality [40, 45]. Best practices have been defined for researchers to validate the accuracy of phenotypes that are extracted from EHRs [13]. And studies have shown that the data, while imperfect, are at least similar in specific cases to self-report data.[26] At the same time, a study of differences in phenotype definitions for diabetes showed significant variation in populations depending on the phenotype definition used [25]. Data quality continues to be an area of concern, though it has yet to invalidate the approach.

One issue of data quality from EHRs that is expected to decrease over time is the issue of data completeness. With government incentives for EHR adoption under the Meaningful Use program, institutions have increased their use of EHRs. More significantly, the Meaningful Use criteria have specified certain data types that must be collected above threshold levels [46]. For the data elements that are part of the Meaningful Use regulations, this will undoubtedly increase the consistency of their creation, and will likely increase the consistency of phenotypes defined for extracting that information from the EHRs for research.

4.4 Implications for Stakeholders

As was introduced in Chap. 2, a variety of stakeholders can and will benefit from the ability to leverage data sources, such as EHRs, in order to enable patient- and population-level phenotyping. Critical examples of these benefits stratified by stakeholder type include the following:

Evidence and Policy Generators

Research at the patient and/or population levels requires the derivation and analysis of complex and discrete phenotypes. Such phenotyping is even more critical when attempting to link clinical presentations of health or disease with biomolecular markers, such as the activities introduced in Chap. 3. While the adoption of healthcare IT platforms, such as EHRs, provide a basis for such phenotype generation, the act of computerizing patient records does not represent a complete solution to such information needs. Thus, the application of phenotyping principles such as those described in this chapter are central to generating data critical to research endeavors that will ultimately result in new, actionable knowledge.

In a similar manner, **policy makers must make decisions and set priorities based upon the best available data and knowledge**. To-date, such decision making, when it pertains to clinical data sets, has been limited due to the lack of comprehensive and actionable representations of patient or population level phenotypes in computational tractable formats. Thus, the provision of comprehensive and rigorous phenotyping methods provides a means of enhancing such data-driven policymaking.

Providers and Healthcare Organizations

Providing tailored and contextually appropriate decision support at the point-of-care, such as that which would like clinical and bio-molecular phenotypes in order to inform disease prevention and/or treatment planning needs, requires computationally tractable representations of patient phenotype data. The use of phenotyping methods overcomes the challenges of what is often critical and unstructured data this is not well aligned with these types of information needs, in order to enable such evidence-driven and personalized healthcare delivery.

As healthcare organizations seek to achieve the "triple threat" of lower costs, increased quality, and improved outcomes of care, **the ability to characterize and manage populations of patients requires that we understand the phenotypes of those individuals and groups**. As such, the use of phenotyping algorithms is central to such population management tasks, which are inherently data analytic in their nature.

Patients and Their Communities

Patients often wish to be integral parts of the care delivery, and even better, wellness promotion activities that make up healthcare management at the individual or population levels. By phenotyping patients based upon the contents of their EHR related information, and enabling the linkage of that data with patient-reported outcomes, sensor data, and other non-traditional sources, **we can enable patients to become part of a "data fabric" the facilitates such shared healthcare decision making**.

Finally, communities often wish to understand measures that can be taken to promote health and wellness. **By using EHR-derived phenotypes for community-based and/or participatory research paradigms, we can empower communities to be part of the evidence-generation process** that underlies knowledge-based approaches to achieving optimal health outcomes.

4.5 Conclusion

For years, researchers have been recommending and attempting the use of EHR data for research. These efforts have been met with moderate success on opportunistic projects where the EHR data was complete enough and matched the research goals. Recently, changes in research towards GWAS in biology and comparative

effectiveness in clinical research have opened wide opportunities with increased need for EHR data. The successes and lessons learned of initial research projects in these areas have created both a foundation and a critical mass that is leading to more and more use. While challenges still exist, studies demonstrating either the limits of those challenges or solutions to them have kept momentum strong. Recent developments to increase the consistency and completeness of EHR data will undoubtedly add to that momentum. We are now at a point that research leveraging EHR data for phenotype and subject definition moves from an opportunity, to an accepted approach, to a priority. This will continue to have dramatic effects on the need for translational informatics.

Discussion Points

- Review what a phenotype is. Discuss what types of data are used to define a phenotype, and where that data exist.
- Discuss how different phenotypes could actually be biased according to data.
- Discuss the effect of Meaningful Use, and how data are actually growing (at what rate is it growing?).
- Discuss how data that are not collected in EHRs could be collected.

References

1. Gullapalli RR, Desai KV, Santana-Santos L, Kant JA, Becich MJ. Next generation sequencing in clinical medicine: challenges and lessons for pathology and biomedical informatics. J Pathol Inform. 2012;3:40.
2. Jollis JG, Ancukiewicz M, DeLong ER, Pryor DB, Muhlbaier LH, Mark DB. Discordance of databases designed for claims payment versus clinical information systems. Implications for outcomes research. Ann Intern Med. 1993;119(8):844–50.
3. Donaldson, MS, Capron AM. Patient Outcomes Research Teams (PORTS): managing conflict of interest [Internet]. [cited 2013 Oct 3]. Available from: http://www.nap.edu/openbook.php?record_id=1821&page=17.
4. Einbinder JS, Rury C, Safran C. Outcomes research using the electronic patient record: Beth Israel Hospital's experience with anticoagulation. Proc Annu Symp Comput Appl Sic Med Care. 1995;819–23.
5. Safran C. Using routinely collected data for clinical research. Stat Med. 1991;10(4):559–64.
6. Safran C, Bloomrosen M, Hammond WE, Labkoff S, Markel-Fox S, Tang PC, et al. Toward a national framework for the secondary use of health data: an American Medical Informatics Association White Paper. J Am Med Inform Assoc. 2007;14(1):1–9.
7. Tierney WM, McDonald CJ. Practice databases and their uses in clinical research. Stat Med. 1991;10(4):541–57.
8. Weng C, Bigger JT, Busacca L, Wilcox A, Getaneh A. Comparing the effectiveness of a clinical registry and a clinical data warehouse for supporting clinical trial recruitment: a case study. AMIA Annu Symp Proc. 2010;2010:867–71.
9. Weng C, Batres C, Borda T, Weiskopf NG, Wilcox AB, Bigger JT, et al. A real-time screening alert improves patient recruitment efficiency. AMIA Annu Symp Proc. 2011;2011:1489–98.
10. Kohane IS. Using electronic health records to drive discovery in disease genomics. Nat Rev Genet. 2011;12(6):417–28.

11. Cannon-Albright LA, Skolnick MH. The genetics of familial breast cancer. Semin Oncol. 1996;23(1 Suppl 2):1–5.
12. Kerem B, Rommens JM, Buchanan JA, Markiewicz D, Cox TK, Chakravarti A, et al. Identification of the cystic fibrosis gene: genetic analysis. Science. 1989;245(4922):1073–80.
13. Bush WS, Moore JH. Chapter 11: Genome-wide association studies. PLoS Comput Biol. 2012;8(12):e1002822.
14. Burton PR, Hansell AL, Fortier I, Manolio TA, Khoury MJ, Little J, et al. Size matters: just how big is BIG?: Quantifying realistic sample size requirements for human genome epidemiology. Int J Epidemiol. 2009;38(1):263–73.
15. Gottesman O, Kuivaniemi H, Tromp G, Faucett WA, Li R, Manolio TA, et al. The Electronic Medical Records and Genomics (eMERGE) network: past, present, and future. Genet Med. 2013.
16. Johnson AD, O'Donnell CJ. An open access database of genome-wide association results. BMC Med Genet. 2009;10:6.
17. Best care at lower cost: the path to continuously learning health care in America [Internet]. [Cited 2013 Oct 4]. Available from: http://books.nap.edu/openbook.php?record_id=13444.
18. Randhawa GS, Slutsky JR. Building sustainable multi-functional prospective electronic clinical data systems. Med Care. 2012;50(Suppl):S3–6.
19. Evans RS, Classen DC, Pestotnik SL, Lundsgaarde HP, Burke JP. Improving empiric antibiotic selection using computer decision support. Arch Intern Med. 1994;154(8):878–84.
20. Bates DW, Evans RS, Murff H, Stetson PD, Pizziferri L, Hripcsak G. Detecting adverse events using information technology. J Am Med Inform Assoc. 2003;10(2):115–28.
21. Wilcox A, Hripcsak G. Medical text representations for inductive learning. Proc AMIA Annu Symp AMIA Symp. 2000;923–7.
22. Kahn MG, Raebel MA, Glanz JM, Riedlinger K, Steiner JF. A pragmatic framework for single-site and multisite data quality assessment in electronic health record-based clinical research. Med Care. 2012;50(Suppl):S21–9.
23. Weiskopf NG, Hripcsak G, Swaminathan S, Weng C. Defining and measuring completeness of electronic health records for secondary use. J Biomed Inform. 2013;46(5):830–6.
24. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. Nat Rev Genet. 2012;13(6):395–405.
25. Richesson RL, et al. A comparison of phenotype definitions for diabetes mellitus. J Am Med Inform Assoc. 2013;20(e2):e319–26.
26. Denny JC. Chapter 13: Mining electronic health records in the genomics era. PLoS Comput Biol. 2012;8(12):e1002823.
27. 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(3):328–37.
28. Rea S, Pathak J, Savova G, Oniki TA, Westberg L, Beebe CE, et al. Building a robust, scalable and standards-driven infrastructure for secondary use of EHR data: the SHARPn project. J Biomed Inform. 2012;45(4):763–71.
29. Patrick J, Li M. High accuracy information extraction of medication information from clinical notes: 2009 i2b2 medication extraction challenge. J Am Med Inform Assoc. 2010;17(5):524–7.
30. Carroll RJ, Thompson WK, Eyler AE, Mandelin AM, Cai T, Zink RM, et al. Portability of an algorithm to identify rheumatoid arthritis in electronic health records. J Am Med Inform Assoc. 2012;19(e1):e162–9.
31. Sarkar IN. Methods in biomedical informatics: a pragmatic approach. Waltham, MA: Academic, 2013; 589 p.
32. Kho AN, Hayes MG, Rasmussen-Torvik L, Pacheco JA, Thompson WK, Armstrong LL, et al. Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study. J Am Med Inform Assoc. 2012;19(2):212–8.
33. Cooper GM, Johnson JA, Langaee TY, Feng H, Stanaway IB, Schwarz UI, et al. A genome-wide scan for common genetic variants with a large influence on warfarin maintenance dose. Blood. 2008;112(4):1022–7.

34. McCarty CA, Chisholm RL, Chute CG, Kullo IJ, Jarvik GP, Larson EB, et al. The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. BMC Med Genomics. 2011;4:13.
35. Newton KM, Peissig PL, Kho AN, Bielinski SJ, Berg RL, Choudhary V, et al. Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network. J Am Med Inform Assoc. 2013;20(e1):e147–54.
36. Loukides G, Denny JC, Malin B. The disclosure of diagnosis codes can breach research participants' privacy. J Am Med Inform Assoc. 2010;17(3):322–7.
37. Loukides G, Gkoulalas-Divanis A, Malin B. Anonymization of electronic medical records for validating genome-wide association studies. Proc Natl Acad Sci U S A. 2010;107(17): 7898–903.
38. McGuire AL, Basford M, Dressler LG, Fullerton SM, Koenig BA, Li R, et al. Ethical and practical challenges of sharing data from genome-wide association studies: the eMERGE Consortium experience. Genome Res. 2011;21(7):1001–7.
39. Kohane IS, Churchill SE, Murphy SN. A translational engine at the national scale: informatics for integrating biology and the bedside. J Am Med Inform Assoc. 2012;19(2):181–5.
40. Kahn MG, Batson D, Schilling LM. Data model considerations for clinical effectiveness researchers. Med Care. 2012;50(Suppl):S60–7.
41. Nichols GA, Desai J, Elston Lafata J, Lawrence JM, O'Connor PJ, Pathak RD, et al. Construction of a multisite DataLink using electronic health records for the identification, surveillance, prevention, and management of diabetes mellitus: the SUPREME-DM project. Prev Chronic Dis. 2012;9:E110.
42. Wilcox A, Yoon S, Boden-Albala B, Bigger JT, Feldman PH, Weng C, et al. Developing a framework for sustaining multi-institutional interdisciplinary community participatory comparative effectiveness research. AMIA 2013 Joint Summits on Translational Science. San Francisco; 2013.
43. Sittig DF, Hazlehurst BL, Brown J, Murphy S, Rosenman M, Tarczy-Hornoch P, et al. A survey of informatics platforms that enable distributed comparative effectiveness research using multi-institutional heterogenous clinical data. Med Care. 2012;50(Suppl):S49–59.
44. Wilcox AB, Gallagher KD, Boden-Albala B, Bakken SR. Research data collection methods: from paper to tablet computers. Med Care. 2012;50(Suppl):S68–73.
45. Brown JS, Kahn M, Toh S. Data quality assessment for comparative effectiveness research in distributed data networks. Med Care. 2013;51(8 Suppl 3):S22–9.
46. Blumenthal D, Tavenner M. The "meaningful use" regulation for electronic health records. N Engl J Med. 2010;363(6):501–4.

Additional Reading

Bush WS, Moore JH. Chapter 11: Genome-wide association studies. PLoS Comput Biol. 2012;8(12):e1002822. doi:10.1371/journal.pcbi.1002822.

Denny JC. Chapter 13: Mining electronic health records in the genomics era. PLoS Comput Biol. 2012;8(12).

Gottesman O, et al. The Electronic Medical Records and Genomics (eMERGE) network: past, present, and future. Genet Med. 2013;15(10):761–7.

Sox HC, Sheldon G. Chapter 2: What is comparative effectiveness research. In: IOM (Institute of Medicine), editor. Initial national priorities for comparative effectiveness research. Washington, DC: The National Academies Press; 2009.

Chapter 5
Mining the Bibliome

Indra Neil Sarkar

By the End of This Chapter, Readers Should Be Able to

- Understand the role of literature in biomedicine (the "bibliome");
- Explore approaches to impute knowledge from the bibliome; and
- Demonstrate the potential of bibliome mining for a learning healthcare system

5.1 Data-Driven Knowledge in Biomedicine

Data in biomedicine originate from a plethora of sources that span the continuum of healthcare. Molecular data may originate from whole genome sequencing, identification of gene sequence variants (such as Single Nucleotide Polymorphisms [SNPs]), or transcriptomic signatures associated with gene expression (such as microarrays). Health data may originate from electronic systems that gather data directly from patients (e.g., using "quantified self" technologies or personal health records), as part of the health care system (e.g., through electronic health record systems or as recorded in health care utilization data sets), or as a result of population health initiatives (e.g., vital records or infectious disease surveillance). To be deemed useful, biomedical data must be analyzed and interpreted through a systematic and repeatable process. The results of these processes, and the foundation on which future insights are often based, are ensconced in biomedical literature.

I.N. Sarkar, PhD, MLIS
Center for Clinical and Translational Science, University of Vermont, Burlington, VT, USA

Department of Microbiology and Molecular Genetics,
University of Vermont, Burlington, VT, USA
e-mail: neil.sarkar@uvm.edu

P.R.O. Payne, P.J. Embi (eds.), *Translational Informatics: Realizing the Promise of Knowledge-Driven Healthcare*, Health Informatics,
DOI 10.1007/978-1-4471-4646-9_5, © Springer-Verlag London 2015

5.1.1 Hypothesis Generation and Hypothesis Testing

Without context or purpose, data are essentially meaningless. To elicit the value of data, they must first be transformed into *information*, which in turn needs to be coalesced into *knowledge* before any utility may be ascertained. This knowledge, which may be deemed as "actionable," may be immortalized as *wisdom* that is used to guide future encounters with data. Formally, the process of data transformation is cast as the "Data-Information-Knowledge-Wisdom" (DIKW) framework. The DIKW framework, and its contemporary incarnations largely attributed to Ackoff [1], provides a formal construct to analyze data transformation. Most importantly, the DIKW framework offers both context and purpose as constraints to instill meaning into volumes of data. This architecture is increasingly important as improvements in data generation and acquisition technologies continue to exceed intellectual capacity for interpretation. It is outside the scope of the present discourse to describe the complete process of data transformation associated with the DIKW framework. Nonetheless, the DIKW framework offers a useful construct to describe how data are used within biomedical contexts.

The advancement of technologies across the spectrum of biomedicine has resulted in a new cadre of data that are referred to as "Big Data" Chap. 7 provides more detail about the nuances of Big Data. Within the current context, where the focus is to leverage knowledge that have been recorded in some reusable form, the discussion will be around approaches that are used for one of two purposes: (1) for hypothesis *generation*; or (2) for hypothesis *testing*. Historically, these two purposes can be perceived to be in conflict with each other; the increased ability to generate hypotheses is of no value without completing the testing of those hypotheses that have already been postulated. Indeed, the scientific methodologies for hypothesis generation and testing are largely considered independently (those that generate hypotheses seldom actually test them and those that test hypotheses seldom are focused on hypotheses generation approaches). However, the realities of contemporary scientific inquiry in light of the volume of data that are available require a synergy between the generation and testing of hypotheses.

For the purposes of this chapter, it is not essential to fully understand the philosophical principles of the Baconian Method or Scientific Method, which can be seen as frameworks for respectively formalizing the process of hypothesis generation and testing [2]. Instead, it is useful to consider these as two major scientific philosophies as approaches that involve the use of data. Big Data might be best leveraged through a Baconian process of reduction of highly complex, highly produced, and highly heterogeneous data into tractable units of "actionable knowledge." Similarly, actionable knowledge might be best utilized if subjected to the Scientific Method for validation. Thus, in the context of a learning healthcare system, there is a necessary synergistic relationship between the Baconion Method and the Scientific Method that marry the realms of hypothesis generation and testing.

5.1.2 Role of Biomedical Literature: The "Bibliome"

The array of data repositories collectively provides an infrastructure for cataloguing and providing access to artifacts that are generated as part of the scientific process. By themselves, the data may not directly convey their meaning; the transformation to knowledge is recorded as wisdom that, within the biomedical context, commonly takes the form of publications. The transformation of individual datum points into actionable knowledge results in an array of artifacts that can be catalogued and be deemed "wisdom" that form the foundation for future studies. It is essential to understand that the transformation of data into wisdom is not necessarily linear in structure; Data might lead to wisdom that, in turn, may be perceived as data for another context. In addition to the interpretations of data that impart wisdom for a particular context, it is essential that the process of transformation is catalogued and that the data themselves are preserved in a potentially reusable form. Indeed, it is important to accept that the ultimate purpose of a given set of data may not actually be known at the time of collection. In biomedical research parlance, this is often referred to as "secondary use" of data – e.g., data that may have been collected for the purposes of monitoring the progress of treatment for a given individual may be aggregated with a population of patients with similar conditions or treatments and form the basis of future studies [3–5].

The ultimate utility of archived data is determined not only by the ability to understand associated interpretations, but also by the ability to leverage the data in combination with other data. To facilitate the archiving of data for potential future uses, it is imperative that a defined standard be used for the representation of the data. In biomedicine and health care, there are defined standards for how data should be represented (e.g., as defined by standards organizations such as Health Level 7 [HL7] and the International Health Terminology Standards Development Organisation [IHTSDO]) or as archived in catalogues of standardized nomenclatures or ontologies (e.g., by the National Library of Medicine as part of the Unified Medical Language System [UMLS [6]] or as indexed in the BioPortal maintained by the National Center for Biomedical Ontology [NCBO [7]]).

There are a multitude of data repositories that can be categorized by data type. For example, nucleotide sequence data are commonly archived (often by journal publisher mandate) in a repository that participates in the International Nucleotide Sequence Database Consortium (INSDC, which consists of GenBank [maintained by the National Center for Biotechnology Information at the United States National Library of Medicine], European Nucleotide Archive [maintained by the European Bioinformatics Institute in Europe] or the DNA Databank of Japan). Many data repositories, such as those associated with nucleotide sequence data, are freely accessible either through direct search interfaces or programmatically. Other data repositories may be more restrictive due to either privacy or confidentiality reasons. For example, patient data may be archived in a clinical data warehouse that is associated with a healthcare organization and their access is restricted to only those that have approved human subjects research protocols that are in accordance with

appropriate legal requirements (e.g., as might be defined by institutional or federal guidelines for research use of human subjects data [8]).

Perhaps the most important role of data repositories is their role in providing an accessible archive of raw material onto which subsequent scientific inquiry might be built. As such, publicly accessible data archiving has increasingly become mandated for publication, receipt of funding, as well as deemed an essential aspect of scientific citizenship. Because of their central role, data are commonly archived in non-proprietary formats and supported centrally by governments as an essential aspect of national importance for research and development (e.g., data repositories for much of biomedicine in the United States are maintained by the National Library of Medicine, one of the institutes of the National Institutes of Health and the largest biomedical library in the world [9]).

Perhaps the best-known repository of data is of these publications, which is maintained by the United States National Library of Medicine. This archive, the Medical Literature Analysis and Retrieval System Online (MEDLINE) is an indexed resource that catalogues the sum of wisdom associated with data that may originate from individual experiments by bench researchers, observations by clinicians, or aggregate analyses of patterns across populations by epidemiologists. MEDLINE traces its origins to the *Index Medicus* that was developed and manually curated by the first director of the National Library of Medicine, John Shaw Billings. As of this writing, MEDLINE consists of over 22 million citations that are systematically indexed with Medical Subject Headings (MeSH), which is the biomedical analog to the Library of Congress Subject Headings (LCSH). MeSH was designed with the purpose to provide a granular set of index descriptors specific to biomedicine and have a more detailed coverage than what is available in the LCSH R (Medicine) hierarchy. In total, there are over 27,000 MeSH descriptors that are applied to each MEDLINE entry.

The potential wealth of knowledge that is embedded within MEDLINE is immense. Concomitant with the rise of various "omic" areas of specialization (e.g., the three canonical "omics" that are aligned with the central dogma of biology: genomics [the study of genome data], transcriptomics [the study of gene expression data], or proteomics [the study of protein data]), the study of data embedded in literature sources is termed "bibliomics." The initial introduction of the term can be traced to the early 2000s and was initially presented as the application of computational approaches to discover new knowledge from within growing corpora of biological texts [10].

In contrast to traditional "omic" areas of study, bibliomics is not canonically focused on the analysis of primary biological data. Instead, the emphasis is on the development of techniques to identify potential linkages across reports about primary data in a way that may unveil new linkages. Intuitively, researchers develop many postulations and evaluate the plausibility of potential hypotheses through the deep study of literature. In the context of bibliomics, the use of computational or informatics techniques to catalyze the process of hypothesis discovery is driven by the promise that new knowledge will emerge. However, grand promises such as this as well as its relatively amorphous definition, along with the non-direct relation to

primary biological data, have earned bibliomics the dubious honor of "Bad Omics Word of the Day" [11]. In the present context, the discussion and use of the term bibliomics will be specific to the process of eliciting knowledge from biomedical literature using techniques such as those associated with data mining.

In pursuit of identifying new knowledge from the growing corpora of biomedical literature, bibliomics offers a unique perspective that is essential in the era of big data. As raw data are produced at increasingly unbelievable rates, the complementary literature offers a key distillation of at least some of these data. Minimally, published reports provide a narrative description of the experiments and the original purpose as well as findings relevant at the time of data generation. New types of publications are also emerging that even more specifically focus on description of data and associated experimental parameters that put the data into context (e.g., *Scientific Data* [12]). It is rare that a publication describing an original investigation does not involve data, therefore literature plays an essential role in mediating the interpretations about data. Subsequent analyses that may involve the amalgamation of data sets further extend the meaning that can be conferred from data. Finally, literature presents a distilled view of data that can itself be mined to identify potentially novel relationships that can either lead to the support, refutation, or generation of hypotheses. It is this last role that is the primary focus of bibliomics.

5.2 Eliciting Knowledge from Biomedical Literature

Continued advances in the technical and practical ability to generate data at unprecedented rates has resulted in an avalanche of data that presents a two-fold challenge: (1) interpreting the data themselves; and, (2) leveraging interpretations to support the scientific process. In the context of biomedicine, biomedical literature is a primary source of data interpretations and subsequent application of the interpretations. It is thus the essential role of literature to serve as the repository of complete biomedical wisdom. Such a position also implicates the enormity of challenges that are faced with leveraging biomedical literature to be used subsequently to support scientific endeavors. In part this challenge is due to the historical audience of biomedical literature: human readers. As such, the utility of biomedical literature for "out-of-the-box" knowledge discovery using automated techniques is a Herculean feat. Whilst there will be increasing sets of biomedical literature that are available in digital format that are machine usable (e.g., in a structured format like the eXtensible Markup Language [XML]), the most value of biomedical literature is in the narrative descriptions of how data were used or interpreted in the context of a study that has encoded nuances that only a human reader is capable of appreciating. This knowledge paradox, where human readability confers greater conveyance of knowledge than more machine-readable formats, suggests that as there is increased generation of big data in biomedicine there will be increased need to develop approaches for leveraging biomedical literature to reveal the potential value of the newly rendered atoms of knowledge.

5.2.1 The Challenge of Unstructured Data

In contrast to well-formed, machine-readable data, the majority of content in biomedical literature is in unstructured form. The value of unstructured format is that it permits one to convey concepts, such as interpretations about data, in a way that is most expressive with no limitations except those that are inherent in language. Although outside the scope of the present discussion, it is important to also acknowledge that beyond written language (e.g., narrative text) another significant source of biomedical knowledge within literature is embedded in graphical formats such as embedded figures that summarize data to facilitate human interpretation. The development of computational approaches for leveraging data interpretations presented in graphical format is an active area of research [13], but will not be covered in the context of bibliome mining here.

Unstructured data, which are most often encountered in the form of text-based sources such as biomedical literature, embody the full suite of challenges associated with understanding written language. Aside from core grammatical features (e.g., periods usually represent the full stop of a sentence or thought), the power of language is in its flexibility. This flexibility, which is inherent in the story-telling nature of human discourse, results in myriad complications for developing automated or machine based approaches to sift through the volumes of generated text. Nonetheless, the story-telling aspect of describing data is an essential element in scientific discourse, since it allows for one to present potentially novel interpretations in a manner without requiring complete redefinition of fundamental concepts [14–16].

Amidst the continued attempts to structure unstructured data, there remains a constant need for unstructured formalisms to capture and describe aspects that are simply not structured. The perennial challenge is thus achieving the appropriate balance of structured and unstructured data formats in a way that allows for the necessary efficiency of structured formalisms and still allows flexibility of unstructured formats.

Biomedical literature does have some structure inherent in its presentation. It is uncommon to find a published article that is not organized into sections such as *introduction, background, materials, methods, results, discussionn* and *conclusion*. Furthermore, narrative can be encoded into structured templates, such as the PubMed Central Document Type Definition (PMC DTD) template that is now part of the Journal Article Tag Suite [17], which further facilitates machine interpretation of document components. For example, the use of the PMC DTD enables one to develop automated routines to extract the methods sections across an entire corpus. Even with structured templates such as PMC DTD, the majority of knowledge remains embedded in narrative form and thus requires additional processing for utility in automated discovery frameworks. Because of the flexibility allowed in unstructured narrative to describe potentially never before described interpretations for a set of data, unstructured formats will likely remain the primary modality for conveying knowledge and recording wisdom. Nonetheless, there is a strong

case for leveraging templates to systematically enhance otherwise unstructured documents with structure to support subsequent uses for the document (e.g., for knowledge discovery).

5.2.2 Natural Language Understanding

Appreciating that there will be likely be a significant component of biomedical literature that continues to be represented in narrative form, there will be a continued demand for the development and use of computational approaches to identify potential embedded knowledge. The sheer volume of biomedical literature that needs to be analyzed will perpetually necessitate the use of computational approaches. As mentioned earlier, MEDLINE consists of more than 20 million citations. Even more impressive is the growth rate of MEDLINE – currently exceeding 1.5 million articles a year, up from 500,000 articles a year less than a decade ago. It is not inconceivable, that with the growth of biomedical data generation, that the interpretations that are embodied into biomedical literature will result in continued growth of annual MEDLINE entries. The sheer volume of text represents a challenge that will increasingly depend on automated approaches for the elicitation of embodied knowledge that might be sequestered in text form.

Natural language processing systems are built around algorithms to mediate between unstructured data and human understanding [18]. Natural language processing systems are of two flavors: (1) Natural Language Understanding (NLU); and, (2) Natural Language Generation (NLG). Both types of systems are rife with challenges. The combination of NLU and NLG systems in fact embody the ultimate Turing test – where the human is able to directly communicate with the computer in natural language without the human being able to detect that it is not interacting with a computer. For the present discussion, we will focus the discussion on NLU systems, since they focus on extracting information from unstructured data such as embodied in biomedical literature.

NLU systems are generally built on a combination of linguistic heuristics that approximate human interpretation of concept recognition, grammar, and ultimately meaning connoted from text. At a high-level, there are three major aspects of NLU: (1) Lexical Analysis – identification of named concepts that can be matched to a dictionary of terms; (2) Syntactic Analysis – identification of syntax used to encode grammar in context of identified terms; and, (3) Semantic Analysis – identification of concepts represented by identified terms. NLU systems have been developed that focus on either one or a combination of these major areas. The inherent variety that is afforded through the power of natural language is also what continues to support the need for advanced research in the development of NLU systems.

The challenges faced by NLU systems not withstanding, the potential to leverage automated routines for extracting information from volumes of text addresses a key issue in the leveraging of potentially available knowledge. The recent exposition of artificial intelligence supported by NLU systems is the Watson system developed by

IBM [19], which harnesses available knowledge including that are encoded into natural language formats. It is important to acknowledge the difficulty in NLU and the interpretation of human discourse – it is not uncommon for meaning to be conveyed through idioms or indirect references. Other challenges are faced by NLU systems, such as resolving abbreviations, connecting concepts across statements; and disambiguation of identified entities, form the inspiration for research. To initiate the process of knowledge discovery, complete NLU functionality that equates to human understanding may not be required. The ability for entity recognition and semantic reconciliation of identified concepts has evolved to the point that a number of publicly available systems can be used with confidence within research environments. Two commonly used systems in biomedicine are the MetaMap system from the National Library of Medicine [20] and Annotator from the National Center for Biomedical Ontology (NCBO) [21].

The vast majority of bibliome mining approaches and resources in biomedicine are geared towards researchers. There is some energy in developing "real-time decision support" that would provide some active support for clinicians; however, most decision support applications are based on passive decision support. In contrast to active decision support systems, where important knowledge inferences are made in real time to clinicians through interactive interfaces, passive decision support systems are based on searching already curated (either manually or through the use of bibliome mining algorithms). An increasingly popular exemplar of this type of passive decision support is the "infobutton," which is increasing being integrated into modern electronic health record systems [22]. Infobuttons work through providing a guided interface to information resources, such as MEDLINE for biomedical literature. In addition to a number of systems that have been designed by dedicated researchers for analyzing specific types of natural language (e.g., for analyzing clinical texts, there are a number of well-described systems like MedLEE [23], MPLUS/ONYX [24, 25], or MedSynDiKATe [26]) there is a continued need to develop NLU systems to extract usable information from the range of natural language sources (including those that are more general in nature, like the ARC system that was designed to be 90 % good for 90 % of information extraction tasks [27]). Historically, NLU systems were designed as specific solutions (commonly written in logic programming languages such as Prolog). In recent years, common programming frameworks have facilitated the ability to develop NLU systems that can be more community driven. The two most prevalent systems are the General Architecture of Text Engineering (GATE [28]) and Unstructured Information Management Architecture (UIMA [29]). By being community driven, it is possible to combine features and functionality into new systems that can meet specific information extraction needs as well as design new techniques that can be shared and enhanced by the community. Active decision support systems that leverage bibliome mining techniques are often termed "question-answer" tools, where a clinician (or anyone with a biomedical question) can present a question in natural language and then the response is based on bibliome mining of all available corpora [30]. The most well-known examples of these types of question-answer system includes the aforementioned IBM Watson as well as the publicly accessible WolframAlpha [31]

search engine. There is some indication that both IBM and Wolfram are interested in applying their technologies for biomedicine (especially with respect to providing more insight into the costs of healthcare, but also for identifying meaningful patterns associated with disease), and there has been some progress within the research community. The IBM Watson system, in fact, was built largely using the UIMA framework and is a testament to the ability and potential power of community driven frameworks. It is also possible to integrate existing NLU systems such as MetaMap into frameworks such as GATE or UIMA, thus enabling one to leverage well-known NLU systems with new techniques.

5.2.3 Role of Metadata and Indexing

In addition to knowledge that may be embedded in text, there are additional sources of information that can be used for knowledge discovery. These come generally in the form of "metadata". Metadata are simply defined as "data about data." In a well-curated system, digital objects are associated with a range of metadata. The most significant benefit of indexing initiatives, such as those led by the National Library of Medicine for indexing MEDLINE [32], is the generation and application of additional metadata for enabling information retrieval systems to meet information needs. The aforementioned MeSH descriptors that are applied through a systematic review of content by subject matter experts and librarians enables one to retrieve citations on a given topic (or combination of topics, or even exclusion of certain topics) with high reliability. Thus, while indexing does not reflect every possible topic, it does provide an accurate high-level aggregation of data objects according to some systematic process. In the case of using MeSH descriptors for organizing MEDLINE content, one can navigate a large corpus of biomedical literature according to more than 27,000 descriptors. Metadata can be generic in form and function, such as to enable discovery of the objects that are organized into a collection.

In contemporary context, metadata are applied to data objects in a systematic manner. A popular metadata format applied to general digital data objects is Dublin Core (DC) [33]. DC is designed for describing a wide array of digital objects, such as those discoverable on the Internet. Within biomedicine, the largest repository of publicly available biomedical literature (MEDLINE) is associated with nearly 90 metadata types that are formally described in a Document Type Definition (DTD) schema. DTDs are written in a formal syntax (written in XML) that is used to describe the set of metadata types that can be associated with a particular digital object. Through DTDs, digital objects become "machine readable," which promotes the potential for computational approaches for discovery of new knowledge. Beyond DC and DTDs, additional metadata standards have emerged that allow for description of scientific research objects. One of particular note is the Investigation, Study and Assay tools (ISA-tools) [34]. The ISA-tools metadata standard can be used to organize and publish the artifacts associated with a particular study.

Metadata in and of themselves are of limited value if they are not universally adopted. Indeed, the challenge of big data is largely defined by the enormity of data and the general inability to decipher not only what the data might mean, but also what the data actually are. This is the essential role that metadata plays. The process of applying metadata to individual datum or sets of data is referred to as "indexing." Canonically, the process of indexing involved the use of tags that could be used for quick reference about the contents of a particular data object. Indexing can be divided into two major tasks: (1) organization of key metadata that are supplied by the data object creator; and (2) application of additional metadata that are specific to the purpose of how the data objects might be organized within an information retrieval system.

The process of indexing thus provides structured information that would otherwise be considered unstructured. In the case of biomedical literature, the largest benefit of indexing is that it provides a necessary functionality to retrieve appropriate information at the point and time of need. For example, if one wanted to identify all literature published by a given author, a well-indexed system would allow for query by the author's name that is identified for each contained data object. Of course, there are inherent challenges with this particular type of query, since one's name may not necessarily be unique.

Metadata and its associated indexing process offer something that natural language does not: quick thumbnail descriptions of collections of data objects. On the other hand, metadata does not necessarily reflect the full view of what might be contained within a given data object. Thus, one might consider metadata to allow for rapid, highly reliable retrieval of related data objects but more involved methodologies are required to shed light on the specifics of what are contained in the data objects themselves. The large volumes of data that are generated does suggest that there may be some merit to leveraging computational techniques such as NLU to facilitate the indexing process. Indeed, the National Library of Medicine has been researching this very challenge through its Medical Text Indexer (MTI) initiative [32]. The aforementioned MetaMap system is in fact a major artifact of this initiative and has been shown to perform at similar levels as human indexers. Such initiatives reflect a major paradigm shift that forms the basis for the key challenge in the era of big data: one that is concerned less with the *generation* of new data, but instead one that is focused on *how to identify* relevant data to meet a set of needs.

5.2.4 Modeling Techniques

Through the development and use of techniques such as NLU and metadata indexing, it is possible to explore large volumes of text using systematic techniques. Beyond the day-to-day utility of indexed information that can be readily accessed at the time of need (e.g., to identify the most relevant literature by a physician in need of the latest literature about the efficacy of a particular treatment regimen), there is a significant potential benefit to computable data that can be used to infer new

knowledge. Perhaps the most well known example of the value of extracting information from biomedical literature to demonstrate the potential to identify new knowledge is demonstrated in a study done by Swanson in 1986 [35]. In this study, Swanson extracted key concepts associated with fish oil and Raynaud's disease. By identifying literature that suggested that Raynaud's disease was associated with an increase in blood viscosity as well as literature that described the reducing effect on blood viscosity by fish oil, Swanson was able to use the transitive property to suggest the hypothesis that fish oil may be used as a treatment for Raynaud's disease. The subsequent validation of this hypothesis through a clinical trial [36] has provided the underpinning inspiration for how new knowledge might be discovered from biomedical text that is systematically organized with resources like MEDLINE. However, the most important part of this study was that it fully leveraged a process of systematically analyzing relevant documents in a way that key facts could be identified and later combined using logic relationships.

The promise of bibliomining is embodied by Swanson's discovery, and has since provided the inspiration for developing algorithmic approaches that aspire to recreate human intuition for identifying potential relationships. The Swanson study demonstrates the potential to identify new knowledge, but also highlights the importance of developing systematic techniques to extract information from biomedical literature. Since the original study, a computer-mediated system called ARROWSMITH was developed and enables one to identify potential linkages between two sets of MEDLINE searches [37, 38]. The ARROWSMITH system is built on the principle that common words or phrases that occur in two sets of documents may be used to identify potentially interesting linkages and thus suggest testable hypotheses. The challenge with language, however, is that in the description of archetypal concepts a variety of terms may be used. This is where NLU systems are essential for the mediation of what was *written* versus what was *meant*. Meaning (or "semantics") is the underpinning challenge of NLU and metadata indexing systems.

The general principles that underpin the ARROWSMITH system can be described through what is referred to as "modeling" algorithms. The essence of these modeling approaches is that identified concepts are placed into a mathematical construct that enables the identification of relationships between the concepts. Relationships are of two general types: (1) direct – where concepts are found to explicitly occur with each other in a specified context (e.g., in the same document); or, (2) indirect – where concepts are related based on inferred relationships that are based on some logical formalism (e.g., in the case of Swanson's study, through a bridging concept that enabled the transitive property to be used to relate otherwise unrelated concepts). Depending on the particular representation approach used, various weights can be applied to each relationship. There are a number of ways that weights can be calculated, including those that are based on direct frequency or weighted frequency. Direct frequency approaches are based on a simple tabulation of how often a particular relationship occurs; weighted frequency approaches are based on tabulation of how often a given relationship occurs, normalized according to how common the relationship is in the universe of all possible relationships.

Weighted frequency approaches are generally more reliable, since they account for the relative importance of a given relationship within a particular context.

A common weighted frequency approach is TF*IDF, where the Term Frequency (TF; the frequency of a given term or relationship occurring in a given document) is normalized according to the Inverse Document Frequency (IDF; the frequency of the given term or relationship occurring in the entire collection of documents). TF*IDF was first demonstrated by Salton in his System for Mechanical Analysis and Retrieval of Text (SMART) information retrieval system [39]. The incorporation of TF*IDF into the SMART system was the first demonstration of the potential utility of modeling techniques for representing concepts and their relative relationships to each other. In formal terminology, TF*IDF is an example of a "Vector Space Modeling" technique that enables for one to identify the closeness of two potentially related concepts in a mathematical (algebraic) space. For the SMART system, the concepts of interest were the documents themselves, with the goal being to identify potentially related documents to one another. This can be generalized to identify potential relationships between concepts based on relative relationships to one another. Recent studies have also shown how genetic information can also be incorporated within a vector space modeling approach to identify potential relatedness between concepts (e.g., between genetically related diseases [40] or between potential medicinal plants and therapeutic applications [41]).

The development of modeling techniques for discerning potential relationships between concepts of interest continues to be an active area of research. Computational approaches continue to be needed and enhanced that can accommodate not only the heterogeneity of how data are represented in the plethora of potential knowledge sources, but also accommodate the volume of data that are being generated as a product of a highly accelerated data generation process. These challenges contribute to those that are even more generally seen with the leveraging of big data for the purposes of real-time knowledge generation (as further described in Chap. 7).

5.2.5 Plausibility of Discovered Knowledge: Evaluation

The mere development of approaches for identification of potentially important concepts or potential relationships through modeling techniques is of little value if one cannot ascertain the value of the potentially identified knowledge. Rigorous evaluation is thus essential for the establishment of trust for knowledge discovery systems. Evaluation of bibliome mining can be done in one of two ways: (1) *Ad hoc* review by experts; or (2) Relative to a pre-defined gold standard. There are relative merits and challenges to each approach, but the general principle is that potential results of an algorithm need to be quantified in some way so that one can ascertain the reliability of the predictions.

Since the overall principle behind developing computational approaches for identifying new knowledge is meant to actually reflect human intuition for discovering new knowledge, a common approach for evaluation involves the leveraging of

human experts. This *ad hoc* review process involves the identification of individuals that have relative expertise to determine whether the results produced by a system are meaningful. Meaningfulness can be specified either as a binary decision (e.g., yes/no) or along a graded scale (e.g., 1[best]-5[worst]). To accommodate for inherent biases that may be introduced as a consequence of human subjectivity, it is considered good practice to have two or more reviewers. The relative agreement between experts can then be quantified using a statistical test, such as Cohen's Kappa [42] (best for cases where there are two experts) or Fleiss' Kappa [43] (which works for scenarios that involve more than two experts). In cases where the results to be evaluated may be intractable, it is common practice to analyze a statistically significant sample (which may be determined either as a defined proportion of the entire result set or as an even sampling of result types to be evaluated). The main advantage to *ad hoc* review evaluation is that one does not require *a priori* knowledge of what might constitute a meaningful result. On the other hand, the challenge of determining the value of an *ad hoc* review is that it is a largely subjective determination and can be biased based on the expertise of the reviewers.

The use of a "gold standard" provides an objective benchmark against which results from a knowledge discovery system can be compared. A gold standard is made of verified results that are to be expected from an accurate knowledge discovery system. Results from a system are then categorized as either: (1) True Positive (TP) – those results that match an expected result; (2) False Positive (FP) – those results that are reported as relevant from a system, but not found in the gold standard; (3) True Negative (TN) – those results that are not expected and also not reported by the system; and, (4) False Negative (FN) – those results that are not reported by the system but are expected. Building on these categorizations, additional statistics are used to quantify the system relative to the gold standard. The two most commonly used are: (1) Sensitivity (Sn) – which assesses the system's ability to detect expected results, calculated as TP/(TP + FN); and, (2) Specificity (Sp) – which assesses the correctness of the results returned by the system, calculated as TN/(TN + FP). Statistically, Sn and Sp respectively quantify the Type 1 (incorrect rejection of a true null hypothesis) and Type 2 (incorrect rejection of a false null hypothesis) errors.

Benchmarking relative to a gold standard offers an objective assessment of system performance; however, the challenge with gold standards are related to the completeness and appropriateness of a gold standard for a given context. Appreciating that a gold standard may not actually be complete or contain all possible solutions, they may also be referred to as a "reference standard." Another related shortcoming with a gold or reference standards is the general inability to completely enumerate what should be a true negative for a given system. This is especially the case in bibliome mining context. Of course, the ideal situation is one where the complete set of results can be compared to a gold standard and then evaluated according to Sensitivity and Specificity. However, the reality is that it is often difficult to determine the true negative rate or even completely specify what should not be expected within a gold standard. To address this, an additional statistic is used, called the Positive Predictive Value (PPV) or Precision (Pr), which

is calculated as TP/(TP + FP). Precision is commonly paired with Recall, which is calculated in the same way as Sensitivity. Precision and Recall can be combined into a single statistic that is the harmonic mean, called the F1-score or F-measure: (Pr × Sn)/(Pr + Sn).

Regardless of what approach might be used for evaluation, it is important to be fully cognizant of the limitations of each approach. That is, evaluation is relative to a given gold standard or expertise used in an *ad hoc* assessment. Nonetheless, it is essential to perform evaluation of knowledge discovery systems, especially in the context of bibliome mining. One valid criticism of gold standard based evaluation is that it may not accurately assess the value of a bibliome mining system to assess *new* knowledge. Furthermore, one might consider that evaluation against a gold standard only validates that a system is able to identify what is already known. Thus, in the context of bibliome mining, *ad hoc* evaluations are more commonplace. This is because the evaluation is based on a true comparison between machine inferred knowledge and human expertise. There are certainly limitations to this approach (e.g., having consistent evaluations across experts); however, such limitations can be addressed in part by metrics such as the aforementioned Cohen's or Fleiss' Kappa statistics. It is important to continuously evaluate biblioming systems and provide benchmark evaluations based on statistically meaningful samples.

5.3 Bibliome Mining to Support a Learning Healthcare System

Within biomedicine, bibliome mining is an area of ongoing research. Many bibliome mining systems have been developed with the ultimate goal of identifying putative hypotheses that can be used to inform clinical decisions. The nuances and complications with how data and thus knowledge are embedded within biomedical literature continue to challenge the research community. However, there is great potential for identifying potentially useful knowledge that may be actionable using the bibliome mining systems that have been developed to date. The relevance of knowledge that may be embedded within biomedical literature may indeed be the underpinning support for the identification of innovations and their subsequent evaluations in the context of a learning healthcare system. As described in Chap. 1, the promise of a learning healthcare system is that it is one that knowledge associated with healthcare are seamlessly fed-forward (to identify new knowledge) and fed-back (to quantify the effect of using identified knowledge). Bibliome mining may very well be an essential process that enables this paradigm shift from the current, static environment of knowledge sharing. For example, considering the previously described discovery of the possible treatment of Raynaud's syndrome using cod liver oil from study of literature (by Swanson), it is essential to consider how to disseminate such knowledge into clinical practice and also evaluate the effect at the population level.

5.3.1 Imputing Wisdom from Available Data

The tsunami of data that are generated across the spectrum of biomedicine underlie the belief that the wisdom of the masses will lead to a revolution in biomedical research and offer unprecedented improvements in quality of patient care [44–46]. Advances in computation, both from efficiency and algorithmic innovations, have even suggested that machine-based knowledge extraction systems will help meet the demand for increased health care professionals. It is important to temper these types of beliefs with the reality of how data are currently available. Indeed, the computational advances that have arose in the last 20 years have led to unprecedented ability to analyze more data than ever believed possible. However, the volume at which data are being generated greatly exceeds the potential to leverage them in a meaningful and, perhaps more importantly in the context of healthcare, timely manner.

The use of bibliome mining techniques are thus needed for two critical and interdependent activities: (1) to partition the growing landscape of biomedical data into relevant versus irrelevant for a given context; and, (2) from within relevant corpora, identify associations that either by themselves or in combination with other data give useful insights that may have not been otherwise obvious. These activities should both be driven by the common goal to identify the most relevant data, at the most relevant time, and in the most efficient manner. The general challenges that characterize the difficulty in leveraging big data (described in Chap. 7) are generalizable to data described within a collection of biomedical literature.

The supporting biomedical infrastructure thus requires a repository of knowledge that can underpin future innovations. Especially in the era of big data and high throughput data generation, it is essential to be able to identify meaningful knowledge from just highly occurring concepts. And, herein lies the greatest challenge: deciphering useful knowledge from simply frequently occurring correlations that are a result of aberrations in how the data are generated or recorded. In order for knowledge to be considered wisdom, significant curation effort is required. No matter how much data are generated, or how much purported knowledge is imputed, the amount of usable storage of wisdom will always be finite. This has significant implications. It is therefore important to appreciate that not all knowledge need be catalogued as wisdom, and yet still be accepted as useful for a given scenario.

Mining the bibliome thus requires a realistic understanding of what problem is aiming to be solved. The needs for bench scientists, clinical practitioners, and community participants may be met by a common set of biomedical literature; however, the detailed approach for mining the appropriate level of knowledge will undoubtedly require differing computational or algorithmic approaches. For example, a bench scientist may be satisfied with the generation of testable hypotheses, whereas a clinical practitioner would require some statistical support for a proposed hypothesis, or whereas a community participant would require some understanding of broader implications of a hypothesis. It is essential to reflect that bibliome mining may offer some glimpse at new knowledge, but many mining exercising will only

identify that there are gaps in current knowledge that prevent the generation of new hypotheses.

The potential limited short-term utility of bibliome mining is also its greatest asset in the context of a learning healthcare system. The point of a learning healthcare system is not that it is all knowing, but instead that it is capable of *learning*. Thus, the ability to identify gaps in current knowledge or evaluate new knowledge based on previously accepted reference standards (that may have evolved from accepted gold standards) is the biggest potential contribution of bibliome mining. Similarly, the process of bibliome mining within a particular context (e.g., what is the current wisdom about type 2 diabetes mellitus?) is an essential aspect of level-setting the present status of the healthcare system. The next major task of a bibliome mining process is then to identify potential opportunities (e.g., by addressing the question: "what are common versus uncommon medications that have been described in the context of clinical trials associated with the top three common comorbidities associated with type 2 diabetes mellitus?").

Within learning healthcare system, bibliome mining plays an essential supporting role. The raw data that are generated (e.g., standardized clinical outcomes) are the primary source of evaluation. Bibliome mining would allow one to develop benchmarks to identify how innovations that are implemented within a healthcare environment are changing (either positively or negatively) versus the existing state. However, the most significant limitation of bibliome mining is that it is not, in and of itself, a process to identify primary data. Primary level evaluations need to be done within the learning healthcare system constraints. Nonetheless, bibliome mining offers an ability to provide general understanding of the implications of applied innovations whilst offering a sobering "reality check" of how primary data are being interpreted – either by those within the specific learning healthcare environment (e.g., health care delivery researchers) or by those in a larger context (e.g., epidemiologists).

5.3.2 Leveraging Data as Actionable Knowledge

Within any biomedical context that involves mining for new knowledge, the most sought after type of knowledge is that which is "actionable." Referencing the earlier mentioned Baconian method, a major goal of bibliome mining is the reduction of data interpretations into what can be distilled to definable hypotheses that can be subsequently subjected to the standard Scientific Method. This is the essence of actionable knowledge – that which provides the underpinning for some action that can be tested and evaluated using accepted methodologies. With the increased flow of data from across the biomedical spectrum, the potential to identify coalesced data that can be used to form the basis of testable hypotheses is unprecedented. However, the great volume, variety, and velocity of the data being generated pose significant challenges in ascertaining the potential value towards identifying knowledge. Of particular note is the challenge of data quality. Just because the volume of data may

be great, if the majority of the data are undecipherable then they are of limited value. Thus, within the array of data sources that may be the source of knowledge within a learning healthcare system it is important to consider the quality and reliability of data.

The challenges with data should not preclude the pursuit of eliciting knowledge. Indeed, the methodologies that have been or will be developed to transform individual datum points into potentially new knowledge will undoubtedly enable a paradigm shift both in knowledge discovery and knowledge sharing. To bolster this paradigm shift, the identification of evaluation methodologies (and accepting their potential shortcomings) will be crucial. Within the context of a learning healthcare system, the development of new knowledge (i.e., testable hypotheses that can lead to actionable events) will require the constant evaluation in light of real-world health contexts. By accepting the shortcomings of a given methodological approach for eliciting new knowledge, but identifying the potential advantages, the healthcare system can learn and thus advance towards a step of identifying new wisdom that can be used to inform subsequent decisions.

5.3.3 Making Biomedical Data Consumable for Populations

Beyond its roles in identifying new knowledge through computantional inferencing, the process of bibliome mining can have more general implications. For consumers, the role of bibliome mining has two major facets: (1) to support their providers through the identification of known facts and potentially new knowledge that can lead to new therapies; and, (2) to provide insights into the realm of known information and offer a means to identify potential suggestions to providers. Together, these two facets reflect the essential role that bibliome mining plays in light of the volumes of biomedical data that need to be deciphered. Bibliome mining is necessary by those directly working in biomedicine – from researchers to clinicians – as well as those who benefit from biomedical innovations – most notably, consumers.

As ubiquitous monitoring (often as part of the overall "quantified self" movement) becomes increasingly the norm, the leveraging of bibliome mining techniques will become more relevant for the general public. For example, for an individual who uses a tracker device (e.g., a Fitbit [47]) as well as had their genome profiled (e.g., using 23andMe [48]) may benefit from an understanding of what their personal data mean in light of published reports. For individual genes or disease conditions, services like 23andMe has offered some insights to the meaning of results in light of available literature (which has been subjected to a combination of biobliome mining techniques and manual curation; at the time of this writing – March 2014– this service has been suspended in response to an FDA warning letter). Prospectively, the need for bibliome mining that can address general queries that span multiple resources (e.g., activity *and* genomic data) may help unveil meaningful insights that can promote more overall positive health. The ultimate utility of quantified self technologies or knowledge inferred from biomedical literature for the general public

still remains to be determined; however, the process of bibliome mining will be essential in the proof or refutation of the importance of such movements.

5.4 Implications for Stakeholders

Again, as was the case in the preceding chapters, we will briefly revisit the stakeholders and activities as described in Chap. 2. Each class of such individuals can and should benefit from robust and systematic approaches to mining the bibliome in support of TI. Specific examples of these opportunities for advancing a vision of knowledge-based healthcare include the following:

Evidence and Policy Generators

- Researchers can **inform the design and execution of protocols and other programmatic initiatives based upon a systematic and comprehensive understanding of the current scientific knowledge base**, as represented in the domain literature.
- Similarly, researchers can **discover new knowledge within integrated collections of domain literature and associated data sets, using** *in silico* **hypothesis discovery methods** (a topic that will be addressed in detail in Chap. 8).
- Policy-makers are able to **identify, consume, and otherwise employ scientific literature that represents that state-of-the-art in terms of domain knowledge**, through the use of rigorous information retrieval and delivery mechanisms.

Providers and Healthcare Organizations

- Individual clinicians can **support point-of-care decision making through the coupling of available data and findings with appropriate and application-specific domain knowledge** as found in various bibliographic sources, thus enabling just-in-time evidence-based-practice.
- Healthcare delivery organizations can **promote and/or enforce evidence-based practice through the targeted and tailored delivery of contextually appropriate information** at various junctures throughout the care delivery process (e.g., "infobuttons" and equivalent constructs).

Patients and Their Communities

- Patients are able to **obtain and digest appropriate domain knowledge from the latest scientific evidence** such that they are able to both ensure that the healthcare information they consume is of high quality, and that they are able to act upon that information to enhance health or wellness.
- Community members can **understand and apply the evidence as found in domain-specific literature** to inform activities such as advocacy and community-based participatory research paradigms.

5.5 Conclusion: Mining the Bibliome of the Future

There is no end in sight with technological improvements that enable an increase in available data at an exponential rate that far exceeds human ability for interpretation, understanding, or meaningful use. Within health care, biomedical literature is an essential mechanism for transfer of knowledge, which is the result of transforming raw data into meaningful information, as wisdom that can guide research inquiry, clinical practice, and community understanding. Biomedical literature will increasingly become the essential bridge between the volumes of data and their interpretations. As the medium for literature increasingly shifts from analog (paper) to digital formats, the ability to leverage bibliome mining approaches will correspondingly improve. Similarly, the ultimate utility of biobliome mining techniques for identifying actionable knowledge will depend on appropriate evaluation and understanding of limitations of inferred knowledge. There will undoubtedly be the need for continued improvements in bibliome mining techniques, which collectively will need to shift from research-only environments to contexts for support clinical care decisions (either by providers or patients). Nonetheless, the prospect of leveraging computational approaches to extract and identify potentially novel knowledge from biomedical literature offers some hope in the harnessing the value of next generation biomedical data.

The process of knowledge discovery and cataloguing of wisdom is inherent in the overall decision support lifecycle. A grand challenge remains the generalization of this decision support lifecycle in the context of a real-world health care system with targeted challenges (e.g., identifying ways to reduce costs while keeping the quality of care or clinical outcomes the same). The most significant role of the bibliome, therefore, is to be the catalogue of wisdom onto which new knowledge discovery approaches are used. Furthermore, the process of bibliome mining will help unveil new potential hypotheses that can be demonstrated or tested within the context of a real-world healthcare system.

As advances in technologies like natural language understanding continue, along with improvements in literature and metadata representation, the potential role of the bibliome will become an essential element central to any functioning and self-assessing healthcare system. This is not entirely different from the *ad hoc* manner in which clinicians may base clinical decisions or identify potential treatment regimens – by scanning literature such as systematically catalogued in resources like MEDLINE, a clinician can identify case studies that describe a particular patient population and possible treatment options along with expected outcomes. By leveraging computational methods to further curate biomedical literature, it is conceivable that new knowledge may be identified that might have otherwise been overlooked. This is because in contrast to a clinician seeking a particular answer for a particular scenario, bibliome mining presents the full array of potential knowledge to consider for a variety of scenarios. In this way, the application of bibliome mining in the context of a learning healthcare system may identify potential hypotheses that advance the overall system in a way that may not have been otherwise considered.

Of course, this is not to suggest that traditional, clinical-based query of the literature will be replaced. Instead, bibliome mining techniques can further enhance knowledge seeking tasks.

Thus, while the growth of biomedical data will continue to expand in terms of volume as well as scope, the systematic cataloging and interpretation as available in biomedical literature (both as traditional scientific inquiry driven publications and purely data description publications) will be essential for the ultimate leveraging of data for purposes beyond those that were used for their initial generation. For learning healthcare systems to fully advance using all the knowledge available, they need to leverage not only those data that were generated to support a given healthcare environment but also those data that may have been generated for some other purpose. The key role of bibliome mining is thus to identify how those data that are not primarily associated with a given learning healthcare query may be relevant.

Discussion Points

- Describe the pivotal role that biomedical literature has for both hypothesis generation and hypothesis testing.
- What are the advantages/disadvantages of indexed systems like MEDLINE?
- Do bibliome mining techniques only rediscover what is already known? How might one evaluate novelty of findings from bibliome mining?
- What are practical considerations when using bibliome mining for: (1) Research? [or other T1 contexts]; (2) Clinical practice? [or other T2 contexts]; or (3) Community Guidance? [or other T3+ contexts]

References

1. Ackoff R. From data to wisdom. J Appl Syst Anal. 1989;16:3–9.
2. Sarkar I. Methods in biomedical informatics: a pragmatic approach. Boston: Academic; 2013.
3. Danciu I, Cowan JD, Basford M, Wang X, Saip A, Osgood S, et al. Secondary use of clinical data: the Vanderbilt approach. J Biomed Inform. 2014.
4. Prokosch HU, Ganslandt T. Perspectives for medical informatics. Reusing the electronic medical record for clinical research. Methods Inf Med. 2009;48(1):38–44.
5. Sharing clinical research data: workshop summary. The National Academies Collection: Reports funded by National Institutes of Health. Washington, DC; 2013.
6. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. 2004;32(Database issue):D267–70.
7. Whetzel PL, Noy NF, Shah NH, Alexander PR, Nyulas C, Tudorache T, et al. BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. Nucleic Acids Res. 2011;39(Web Server issue):W541–5.
8. Wiesenauer M, Johner C, Rohrig R. Secondary use of clinical data in healthcare providers – an overview on research, regulatory and ethical requirements. Stud Health Technol Inform. 2012;180:614–8.
9. Collen MF. Computer medical databases: the first six decades (1950–2010). London/New York: Springer; 2012. xix, 288 p.

10. Grivell L. Mining the bibliome: searching for a needle in a haystack? New computing tools are needed to effectively scan the growing amount of scientific literature for useful information. EMBO Rep. 2002;3(3):200–3.
11. The tree of life blog by Jonathan Eisen [Mar 6, 2014]. Available from: http://phylogenomics.blogspot.com/2010/03/bibliome-wikipedia-free-encyclopedia.html.
12. Scientific data [Mar 6, 2014]. Available from: http://www.nature.com/scientificdata/.
13. Muller H, Michoux N, Bandon D, Geissbuhler A. A review of content-based image retrieval systems in medical applications-clinical benefits and future directions. Int J Med Inform. 2004;73(1):1–23.
14. Lam HY, Marenco L, Clark T, Gao Y, Kinoshita J, Shepherd G, et al. AlzPharm: integration of neurodegeneration data using RDF. BMC Bioinforma. 2007;8 Suppl 3:S4.
15. Sandor A, de Waard A. Identifying claimed knowledge updates in biomedical research articles. Proceedings of the Workshop on Detecting Structure in Scholarly Discourse, Jeju Island, Korea. 2012. p. 10–7.
16. Ciccarese P, Wu E, Wong G, Ocana M, Kinoshita J, Ruttenberg A, et al. The SWAN biomedical discourse ontology. J Biomed Inform. 2008;41(5):739–51.
17. Beck J. NISO Z39.96 The Journal Article Tag Suite (JATS): what happened to the NLM DTDs? J Electron Publ. 2011;14(1). http://dx.doi.org/10.3998/3336451.0014.106
18. Cohen KB, Demner-Fushman D. Biomedical natural language processing. Amsterdam: John Benjamins Publishing Company; 2013. pages cm. p.
19. Ferucci D, Brown E, Chu-Carroll J, Fan J, Gondek D, Kalyanpur AA, et al. Building Watson: an overview of the DeepQA Project. AI Mag. 2010;31(3):59–79.
20. Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. J Am Med Inform Assoc. 2010;17(3):229–36.
21. Jonquet C, Shah NH, Musen MA. The open biomedical annotator. Summit Transl Bioinforma. 2009;2009:56–60.
22. Cimino JJ. Infobuttons: anticipatory passive decision support. AMIA Annu Symp Proc. 2008:1203–4
23. Friedman C. A broad-coverage natural language processing system. AMIA Annu Symp Proc. 2000:270–4.
24. Dublin S, Baldwin E, Walker RL, Christensen LM, Haug PJ, Jackson ML, et al. Natural language processing to identify pneumonia from radiology reports. Pharmacoepidemiol Drug Saf. 2013;22(8):834–41.
25. Christensen LM, Haug PJ, Fiszman M. MPLUS: a probabilistic medical language understanding system. In: Proceedings of the workshop on Natural Language Processing in the Biomedical Domain, Philadelphia, PA. 2002. p. 29–36.
26. Hahn U, Romacker M, Schulz S. MEDSYNDIKATE–a natural language system for the extraction of medical information from findings reports. Int J Med Inform. 2002;67(1–3): 63–74.
27. D'Avolio LW, Nguyen TM, Farwell WR, Chen Y, Fitzmeyer F, Harris OM, et al. Evaluation of a generalizable approach to clinical information retrieval using the automated retrieval console (ARC). J Am Med Inform Assoc. 2010;17(4):375–82.
28. Cunningham H, Maynard D, Bontcheva K, Tablan V. GATE: an architecture for development of Robust HLT applications. In: ACL '02 Proceedings of the 40th annual meeting on Association for Computational Linguistics, Stroudsburg, PA; 2002. p. 168–75.
29. Ferrucci D, Lally A. UIMA: an architectural approach to unstructured information processing in the corporate research environment. Nat Lang Eng. 2004;10(3–4):327–48.
30. Athenikos SJ, Han H. Biomedical question answering: a survey. Comput Methods Programs Biomed. 2010;99(1):1–24.
31. WolframAlpha [Mar 6, 2014]. Available from: http://www.wolframalpha.com/.
32. Aronson AR, Mork JG, Gay CW, Humphrey SM, Rogers WJ. The NLM indexing initiative's medical text indexer. Stud Health Technol Inform. 2004;107(Pt 1):268–72.
33. Weibel S. The Dublin core: a simple content description model for electronic resources. Bull Am Soc Inf Sci Technol. 1997;24(1):9–11.

34. Rocca-Serra P, Brandizi M, Maguire E, Sklyar N, Taylor C, Begley K, et al. ISA software suite: supporting standards-compliant experimental annotation and enabling curation at the community level. Bioinformatics. 2010;26(18):2354–6.
35. Swanson DR. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. Perspect Biol Med. 1986;30(1):7–18.
36. DiGiacomo RA, Kremer JM, Shah DM. Fish-oil dietary supplementation in patients with Raynaud's phenomenon: a double-blind, controlled, prospective study. Am J Med. 1989;86(2): 158–64.
37. Smalheiser NR, Swanson DR. Using ARROWSMITH: a computer-assisted approach to formulating and assessing scientific hypotheses. Comput Methods Programs Biomed. 1998; 57(3):149–53.
38. Arrowsmith [Mar 6, 2014]. Available from: http://arrowsmith.psych.uic.edu/arrowsmith_uic/.
39. Salton G, McGill MJ. Introduction to modern information retrieval. New York: McGraw-Hill; 1983. xv, 448 p.
40. Sarkar IN. A vector space model approach to identify genetically related diseases. J Am Med Inform Assoc. 2012;19(2):249–54.
41. Sharma V, Sarkar IN. Leveraging concept-based approaches to identify potential phyto-therapies. J Biomed Inform. 2013;46(4):602–14.
42. Carletta J. Assessing agreement on classification tasks: the Kappa statistic. Comput Linguis. 1996;22(2):249–54.
43. Fleiss JL. Measuring nominal scale agreement among many raters. Psychol Bull. 1971;76(5): 378–82.
44. Kwon SW. Surviving in the era of "Big Data". Blood Res. 2013;48(3):167–8.
45. Howe D, Costanzo M, Fey P, Gojobori T, Hannick L, Hide W, et al. Big data: the future of biocuration. Nature. 2008;455(7209):47–50.
46. Baldwin G. Small fish, big data pond. Health Data Manag. 2009;17(9):48.
47. Fitbit [Mar 6, 2014]. Available from: https://www.fitbit.com/.
48. 23andMe [Mar 6, 2014]. Available from: https://www.23andme.com/.

Additional Reading

Collen MF. Computer medical databases: the first six decades (1950–2010). London: Springer; 2012. xix, 288 p.
Danciu I, Cowan JD, Basford M, Wang X, Saip A, Osgood S, et al. Secondary use of clinical data: the Vanderbilt approach. J Biomed Inform. 2014. (in press) http://dx.doi.org/10.1016/j.jbi.2014.02.003
Friedman C. A broad-coverage natural language processing system. Proc AMIA Symp. 2000:270–4.
Grivell L. Mining the bibliome: searching for a needle in a haystack? New computing tools are needed to effectively scan the growing amount of scientific literature for useful information. EMBO Rep. 2002;3(3):200–3.
Howe D, Costanzo M, Fey P, Gojobori T, Hannick L, Hide W, et al. Big data: the future of biocuration. Nature. 2008;455(7209):47–50.
Prokosch HU, Ganslandt T. Perspectives for medical informatics. Reusing the electronic medical record for clinical research. Methods Inf Med. 2009;48(1):38–44.
Salton G, McGill MJ. Introduction to modern information retrieval. New York: McGraw-Hill; 1983. xv, 448 p.
Sarkar I. Methods in biomedical informatics: a pragmatic approach. Boston: Academic; 2013.

Part III
Applications of Translational Informatics

Chapter 6
Driving Clinical and Translational Research Using Biomedical Informatics

Philip R.O. Payne and Peter J. Embi

By the End of This Chapter, Readers Should Be Able to

- Describe study types and designs commonly encountered in the clinical and translational research domain;
- Understand the information needs of stakeholders involved in clinical and translational research, and how those needs can be satisfied using Biomedical Informatics theories and methods;
- Understand how to critically evaluate study designs and information needs in order to optimize the adoption/adaptation of Biomedical Informatics tools or technologies; and
- Synthesize the challenges and opportunities in the Biomedical Informatics domain as they pertain to clinical or translational research, and that may serve to inform new lines of basic and applied research and innovation.

P.R.O. Payne, PhD, FACMI (✉)
Department of Biomedical Informatics, The Ohio State University Wexner Medical Center, Columbus, OH, USA
e-mail: philip.payne@osumc.edu

P.J. Embi, MD, MS, FACP, FACMI
Departments of Biomedical Informatics and Internal Medicine, The Ohio State University, Columbus, OH, USA
e-mail: peter.embi@osumc.edu

P.R.O. Payne, P.J. Embi (eds.), *Translational Informatics: Realizing the Promise of Knowledge-Driven Healthcare*, Health Informatics,
DOI 10.1007/978-1-4471-4646-9_6,

6.1 Introduction

Clinical and Translational Research (CTR) is a virtuous cycle during which diagnostic or therapeutic discoveries generated in the laboratory settings are validated for safety and efficacy via pre-clinical studies (often involving model organisms), and subsequently evaluated in terms of safety, efficacy, and clinical outcomes in humans via systematic and rigorous clinical studies. If the preceding clinical studies result in a positive finding, the knowledge generated therein is then disseminated for application at either the point-of-care or population levels. Furthermore, in this same virtuous cycle, observations of "real world" phenomena at the individual and population levels can be used to inform new hypotheses for laboratory study, thus bringing the cycle back to its origins. Underlying the CTR cycle is a complementary substrate of Biomedical Informatics (BMI) theories and methods, which per the working central dogma for the field as introduced in Chap. 1, are concerned with the translation of raw data into information and ultimately knowledge. This overall construct is illustrated in Fig. 6.1.

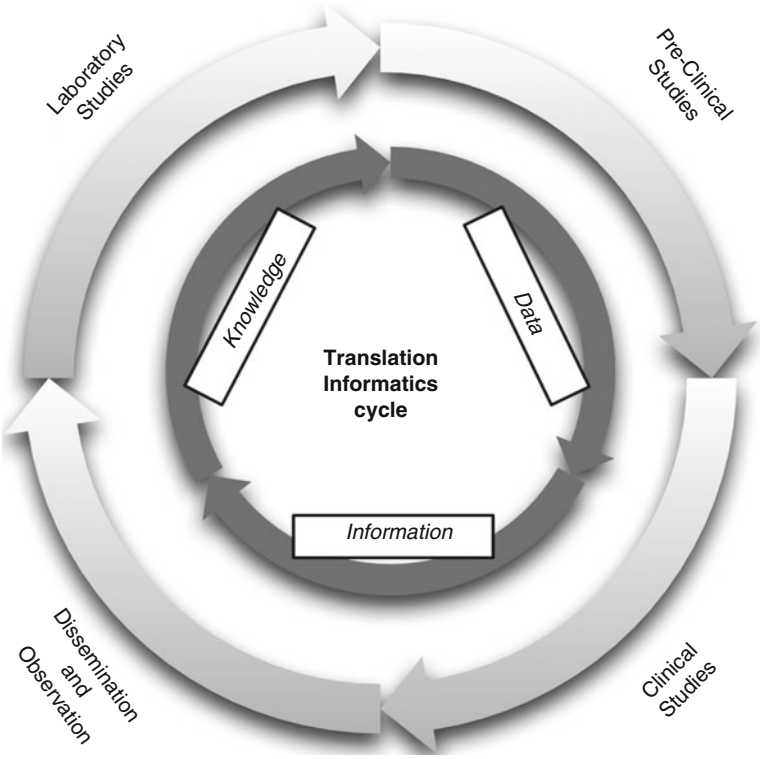

Fig. 6.1 Overview of the CTR cycle and underlying Biomedical Informatics substrate upon which it operates

As described in a number of recently published reports, the conduct of CTR is challenging due to a variety of technical and socio-cultural factors. In one seminal report sponsored by the Institute of Medicine (IOM) and authored by Sung and colleagues, a set of critical "barriers" that exist in the CTR cycle were enumerated, namely [1–5]:

- A "**T1**" **barrier** between the knowledge generated in laboratory studies and the design/conduct of clinical studies informed by that knowledge;
- A "**T2**" **barrier** between the knowledge generated in clinical studies and the dissemination/adoption of that knowledge in clinical care setting; and
- Multiple barriers, including those labeled as "**T3**" **and beyond**, that serve to impede the translation of knowledge across and between settings and communities after such evidence has been generated and initially disseminated via traditional research processes.

Underlying these barriers are numerous informatics-relevant challenges, such as those associated with: (1) providing efficient and high quality data collection instruments; (2) facilitating the integration and "normalization" of complex and heterogeneous data sets; (3) analyzing multi-dimensional data and information in a manner that leverages the best available quantitative practices and pre-existing knowledge bases; and (4) representing and disseminating knowledge in a manner that ensures that those products are both transportable and understandable (by humans and computers alike). These challenges are, in large part, the result of a combination of technical, cultural, and organizational impediments, with examples of such issues enumerated in Table 6.1.

6.2 A Primer on Clinical and Translational Research

In the following discussion, we briefly review the basic definitions, constructs, and frameworks that underlie most, if not all, CTR projects. Specifically, we will first describe the basic components that make up a traditional clinical trial or study, and then describe how such models can be extended or enhanced to yield more translational investigative paradigms.

6.2.1 Clinical Studies

The National Institute for General Medical Sciences (NIGMS) defines a clinical trial as a "*scientific study in which physician-researchers study the effects of potential medicines on people*; *usually conducted in three phases* (*I*, *II*, *and III*) *that determine safety*, *whether the treatment works*, *and if it's better than current therapies*, *respectively*" [6]. Of note, a potential limitation to this definition is that it only references medical treatment modalities. In reality, clinical trials can incorporate additional therapeutic and non-therapeutic approaches, including surgery, medical devices, and tissue banking. In addition, there are various types of clinical studies

Table 6.1 Exemplar factors that contribute to major informatics challenges encountered in the CTR domain

Category	Factor	Description
1. Technical	Availability of appropriate technologies	In many instances, the specific technologies needed to support or enable a given study design do not exist, and must be created for the purposes of the research project. Examples include tools/methods for capturing data from specialized instruments or devices, or the ability to collect data from patients in atypical settings
	Ability to instrument Healthcare IT (HIT) platforms	When research program require the collection of data at the point-of-care, it is often desirable to instrument HIT platforms, such as EHRs, to obtain such data. However, constraints placed by vendors and/or operational IT policies may prevent or impede the customization of HIT platforms to enable such "secondary" data capture and reuse
	Capability to collect, store, manage, and analyze "big data"	As introduced in Chap. 1 and detailed in Chap. 7, "Big Data" is becoming increasingly common place in Biomedicine. However, the lack of appropriate and necessary computational resources (e.g., data storage, network bandwidth, processor capacity) as well as applicable analytical methods often impedes the analysis of such "Big Data"
2. Cultural	Predisposition towards reductionism	Traditional approaches to science often emphasize a reductionist approach to hypothesis formulation and testing. However, complex CTR study types often require a systems level approach to their formulation and conduct. Therefore, the ability to break with cultural norms relative to reductionism continues to be both important and difficult
	Enabling systems thinking	In a corollary manner to the preceding factor, enabling research teams to address driving biological or clinical problems at a systems level requires the creation of cultural environments to foster such "out of the box" and team-based thinking
	Providing support for team science	As research problems become increasing complex, the need for the formation and support of multidisciplinary teams becomes more and more important. Unfortunately, cultural norms related to the assignment of scholarly credit and academic career trajectories, as well as the business needs of private sector stakeholders, can create disincentives to participate in such multidisciplinary and team-oriented projects
3. Organizational	Implementation of appropriate structures	The preceding reductionist factors, often combined with environments that do not support/enable team science, are usually symptoms of organizational structures that are incompatible with systems- and team-based approaches to science. Therefore, organizational structures need to be realigned in order to address the preceding cultural factors. An example of this type of factor may be the disciplinary and geographic distribution of departments and centers in a given Academic Health Center (AHC). Unfortunately, effecting such change is politically and culturally difficult in complex organizations with lengthy operational histories and established cultures
	Engagement of leadership	Without the appropriate engagement and support of leadership, it is unlikely that CTR teams will be able to gain access to the resources and facilities needed to support/enable their research activities. Therefore, the full and regular engagement of key leaders in the CTR "fabric" becomes very important to such initiatives. Such engagement requires that leaders "value" CTR
	Access to sufficient resources	In a corollary manner to the preceding factor, CTR teams also require access to novel or complex infrastructure and resources in order to address their scientific aims. Often, these resources span traditional organizational boundaries. As such, appropriate leadership engagement and resourcing of CTR projects, particularly given complex funding and regulatory environments, becomes critically important

that do not involve direct and study-specific interventions (such as observational or comparative effectiveness studies), and instead rely on data generated during the course of standard-of-care activities. Therefore, for the purposes of this book, we will define a clinical study (which can included clinical trials as well as non-interventional investigations) as follows:

A scientific study in which investigators evaluate one or more aspects of the efficacy, safety, cost, and/or performance of a diagnostic or therapeutic approach to a given disease state, or observe a disease state in order to better understand its biological, clinical, and population level implications.

As implied previously, clinical studies are part of a broader research paradigm, often referred to as translational research [7]. A common metaphor for this process is that of bringing a novel therapeutic approach from "bench to bedside".

Given their role as the "gold standard" for clinical studies, we now focus on the design and conduct of clinical trials. Such trials can be divided into two primary types: (1) observational studies and (2) randomized controlled clinical trials. Observational studies are generally the result of reoccurring clinical phenomena observed by investigators during the delivery of conventional care. These observations are then evaluated, generalized and reported upon [8]. In contrast, randomized controlled trials are specifically designed studies that aim to answer pre-defined scientific questions and minimize bias in study results [9]. The overall design of randomized controlled trials can be further divided into three major classes [8]:

- **Open Trials**: Study investigators and participants know the type of treatment being provided.
- **Single-blind**: Study investigators know the type of treatment being provided, but participants do not.
- **Double-blind**: Neither study investigators nor participants know what type of treatment is being provided. A set of records to allow for un-blinding during subsequent data analysis is maintained by a trusted third party.

A key component of the randomized controlled trial model is the concept of *control*, which refers to the comparison of the performance of an intervention in a test group to the performance of a control group that does not receive the aforementioned intervention. These test and control groups can be further sub-divided based on variations in the aims of the study. The process of assigning trial participants to test or control groups is usually done in a random manner during a process known as randomization. This random assignment is performed in order to reduce potential biases in the study outcomes. The groups to which participants are assigned are often called *treatment* or *study* arms.

The conduct of clinical trials can be further defined by study phases, where each phase corresponds to a discrete scientific aim of the trial. A complete clinical trial ideally consists of four phases [8]:

- **Phase I**: Investigators evaluate the intervention in a small group of participants in order to assess overall safety. This safety assessment includes dosing levels in

the case of medical trials, and potential side effects or adverse effects of the therapy.

- **Phase II**: Investigators evaluate the intervention in a larger group of participants in order to assess the efficacy of the intervention in the targeted disease state. During this phase, assessment of overall safety is continued.
- **Phase III**: Investigators evaluate the intervention in an even larger group of participants and compare its performance to a reference standard, which is usually the current standard of care for the targeted disease state. In general, this is the final study phase to be performed before seeking regulatory approval for the intervention as is required depending on its type/nature.
- **Phase IV**: Investigators study the performance and safety of the intervention after it has been approved and marketed. This type of study is performed in order to detect long-term outcomes and effects of the intervention. It is often called "post-market surveillance."

The conduct of a Phase I, II or III clinical trial can be thought of in an operational sense as consisting of three primary stages: (1) screening, (2) active monitoring and (3) follow-up [10] (Fig. 6.2). During the three stages, a specific temporal series of processes is executed. First, potential participants must be screened to determine if they meet the inclusion and exclusion criteria for the study (e.g., specific demographic and/or clinical parameters required for subjects to be eligible for the study). Once a potential participant has been deemed eligible for the study, they are provided with an informed consent document, which must be signed prior to proceeding with the enrollment process. 'Enrollment' in the context of clinical trials means officially registering as a study participant, and is normally associated with the assignment of a study-specific identifier. Once a person agrees to become a participant, they are enrolled, and in the case of studies with multiple study groups or arms, randomized into one of the study arms. In second stage of the trial, known as active monitoring, the participant receives the intervention indicated by their study arm and is actively monitored to enable the collection of study-specific data. This intervention and active monitoring process is often iterative, involving multiple cycles of intervention delivery and active monitoring. Finally, the follow-up phase begins once a participant has completed the active monitoring stage of a study. During this stage, subjects are contacted on a specified temporal basis in order to collect additional data of interest, such as long-term treatment effects, disease status or survival status [8].

A clinical trial is usually described using a document called a *protocol*, which contains background information, scientific goals, aims, hypotheses and research questions to be addressed by the trial. In addition, the protocol describes policies, procedures, and data collection or analysis requirements. A summary of tasks and events that must occur during the active monitoring phase of a trial, known as the study schema or calendar, is often included in the protocol document.

The quality of data produced by a clinical trial is assessed using multi-dimensional metrics, which take into account the design, execution, analysis and dissemination of the study results. The quality of a clinical trial is also judged with respect to the significance or relevance of the reported study results within a clinical context [11].

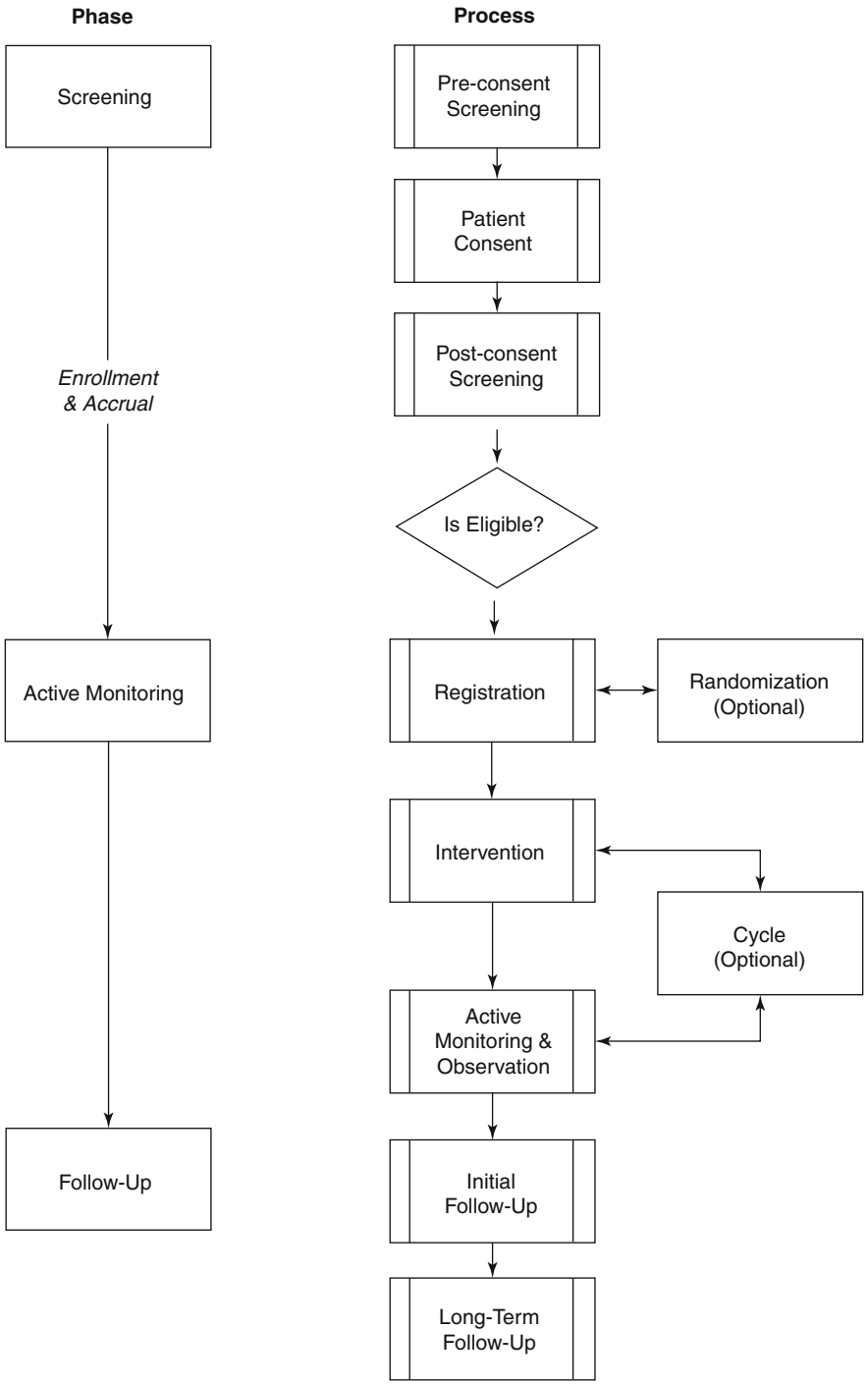

Fig. 6.2 Phase I-III clinical trial processes. Such processes can generally be divided into the screening, active monitoring, and follow-up phases

One key metric used to assess clinical trial quality is *validity*, which can be defined both internally and externally. Internal validity is defined as the minimization of potential biases during the design and execution of the trial, while external validity is the ability to generalize study results into clinical care [11]. The primary source of internal validity problems are either the Hawthorne Effect (e.g., a scenario in which the act of measuring a pheromone of interest serves to change the targeted process or behavior) [12], or study bias, which Juni et al. have defined as falling into four primary types [11]:

- **Selection bias** occurs when participants are enrolled or randomized in a study in preferential manner.
- **Performance bias**, often found in open or single-blind trial designs, occurs when therapies are provided to study participants in a preferential manner.
- **Detection bias**, usually found in open or single-blind trial designs, occurs when knowledge regarding to which arm a participant has been assigned is allowed to affect the interpretation of study outcomes.
- **Attrition bias** occurs when data is censored or otherwise removed from study analyses due to the attrition of participants.

6.2.2 Extending the Clinical Study Paradigm to Achieve Translational Aims

In order to extend the clinical study model described in Sect. 6.2.1 to achieve broader and translational aims, it is usually necessary to extend and enhance the clinical study design. Such extension and enhancement include the provision of correlative aims and scientific activities both prior to the design and conduct of Phase I clinical studies, as well as the creation of feedback mechanisms from Phase IV and/or observational studies to inform novel hypotheses for testing in laboratory based settings. Such additions to the model introduced in Sect. 6.2.1 are illustrated in Fig. 6.3 and correspond to the conduct of "***Basic and Pre-Clinical Research***" upstream of clinical studies, and the conduct of "***Pragmatic Research***" downstream of clinical studies, as described below.

6.2.2.1 Basic and Pre-clinical Research

In this precursor phase to the design and conduct of clinical studies, laboratory-based investigators explore the structural and functional bases for a given disease state in order to identify causative or modifiable factors that may serve to inform diagnostic and therapeutic strategies (referred to as clinical end-points in the remainder of this discussion). For those findings that show the potential to lead to new or improved clinical end-points, a process of pre-clinical research is then undertaken, often using model organisms as a "test bed", so as to determine the safety and

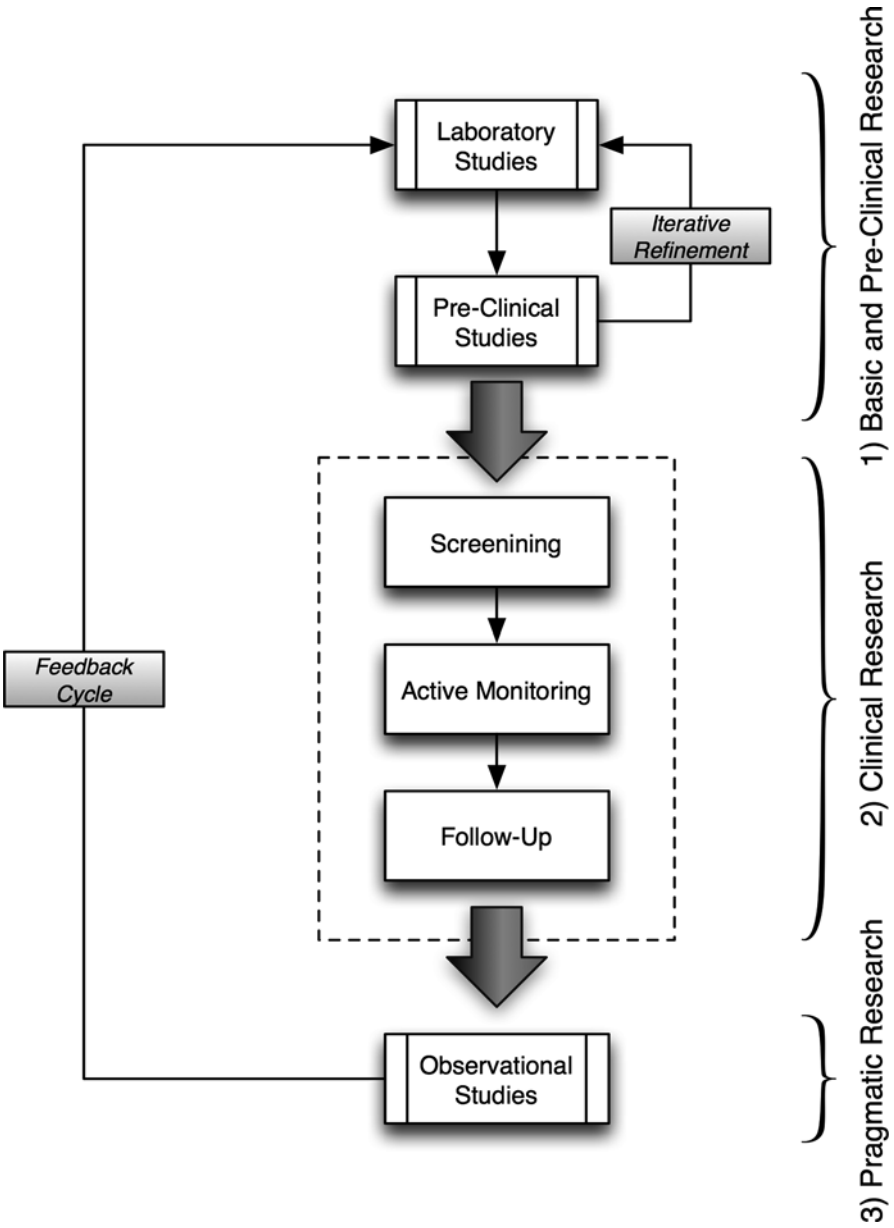

Fig. 6.3 Overview of upstream and downstream extensions and enhancements to traditional clinical studies intended to support translational aims

efficacy of the specific diagnostic or therapeutic strategy. In many cases, pre-clinical studies lead to new basic science questions that must be answered prior to moving forward with the translation of the clinical end-point. This leads to an iterative refinement cycle. Once such iterative refinement between basic science and pre-clinical research comes to fruition, the resulting diagnostic or therapeutic strategy is translated into early clinical studies per the processes and activities described earlier in this chapter.

6.2.2.2 Pragmatic Research

In the follow-on phase to clinical research, pragmatic scientific questions are posed and answered through the assessment of data generated either: (1) during the course of standard-of-care activities; or (2) public-domain data sets resulting from historical clinical studies. These questions can include basic associations between the various dimensions of diagnostic or treatment modalities and the outcomes experienced by patients (including health status, quality of life, or cost). When such associations are found to be present and of interest, they may be further explored to determine what hypothesis can be formulated as to their biological or mechanistic bases, thus generating research questions that can be "fed back" into the translational cycle and used by basic scientists to inform new lines of laboratory based investigation. Of note, in some instances, such pragmatic research, if yielding purely clinical hypothesis, can lead to direct feedback to the clinical research process without the necessity for basic science and pre-clinical research.

6.3 The Role of Informatics in Clinical and Translational Research

The benefits made possible by using BMI theories and methods to address the information needs associated with the CTR paradigm have been described frequently in the literature [1, 4, 7, 13]. In general, the use of BMI theories and methods in this context can be aligned with one or more of the following problems area:

6.3.1 The Collection and Management of Heterogeneous and Multi-dimensional Data Sets

With the increasing availability of high-throughput data sources, such as electronic health records (EHRs) and clinical research management systems (CRMS) or Electronic Data Capture (EDC) tools, as well as 'omics' instrumentation, the size and complexity of data sets that researchers must collect, store and retrieve on a

regular basis is growing at a rapid and almost unheard of rate [13–16]. At the same time, the data management practices currently used in research environments rely on the use of conventional database or file-based management approaches that are ill-suited to such "big data" sets [14, 16, 17]. Therefore, the use of integrative and scalable information management platforms is critical to reducing the data management burden associated with such multi-dimensional data, thus allowing researchers and their staff to focus on fundamental scientific problems, rather than practical computing needs [5, 7, 14, 18]. In addition, with the growth of scenarios in which investigators need to link such high-throughput bio-molecular and phenotypic data together in meaningful ways so as to better understand potential relationships between them, it is also imperative that the semantics of such data be well understood [17, 19, 20]. Such semantic interoperability between data (either within a given data set or across data sets) requires the use of knowledge engineering approaches to map among various representational schemas and codification regimes for source data sets [17, 20]. When taken as a whole, the types of motivating questions one might encounter relative to the collection and management of heterogeneous or multi-dimensional data sets can include:

1. *What are the optimal tools to allow me to collect or re-use data for my research project as it is generated* via *either clinical encounters or research specific interactions with participants or populations*?
2. *How can I store large collections of research data in ways that make it timely and easy to both index and retrieve depending on downstream data analysis needs*?
3. *How can I "normalize" the coding schemas or data structures for multiple source data sets so that I can then analyze the interrelationships between variables of interest contained within those resources*?

6.3.2 Using Knowledge-Anchored Methods to Discover and Test Hypotheses Concerning Linkages Between Phenotypic and Bio-molecular Variables

Current approaches to hypothesis generation and testing primarily tend to rely upon the intuition of an individual investigator or their team [13, 21]. As such, these research questions tend to be limited in scope relative to the knowledge and experience of those individuals, and not necessarily representative of the full scope of applicable scientific knowledge or inquiry. Beyond this primary limitation, it is also important to note that such a human-centered approach is really only practicable when the scale and scope of scientific data being considered is commensurate with basic human cognitive capabilities. However, as data sets expand to reach "big data" proportions (minimally in terms of size and speed at which data are generated), such an approach becomes rapidly intractable and highly limiting [17, 19]. At the same time, significant knowledge that could be used to assist in the formulation of

hypotheses relevant to a given data set are often dispersed across a myriad of resources such as the domain literature, formal ontologies, and public data sets or models [19]. At the current time, tools and applications that allow researchers to access and extract knowledge from domain-specific sources, and then use those resulting knowledge extracts to inform "high throughput" hypothesis generation, remain relative immature [17, 19]. As a result, significant additional effort is needed to design and validate such tools and provide them for regular use by the scientific community. Again, as was the case with the preceding problem area, and when taken as a whole, the types of motivating questions one might encounter relative to using knowledge-anchored methods to discover and test hypotheses concerning linkages between phenotypic and bio-molecular variables in large-scale or "big data" constructs can include:

1. *What are all of the research questions I could ask regarding my data collection*?
2. *Based upon the contents of the current biomedical literature, are there interesting associations between data elements in my research data set that I should be exploring*?
3. *Can I augment a research data set with linked, open data so that I can test complex or otherwise intractable hypotheses*?

6.3.3 The Provision of Systematic and Extensible Data-Analytic Pipelining Platforms

Often, BMI tools and methods are employed in the CTR setting in order to provide for systematic data-analytic "pipelining" platforms that are capable of supporting the definition and reuse of data analysis workflows incorporating multiple source data sets, intermediate data analysis steps and products, and output types [22, 23]. The value of such data analysis pipelines are many, including: (1) support for the rapid execution of complex data analysis plans that would otherwise require time- and resource-intensive manual multi-step processes to transact, manipulate, and analyze data sets; and (2) enable the collection of information concerning the data analysis methods being used. In the case of the latter benefit, such information can be utilized to better understand the outcomes of such analyses, and to ensure reproducible results and high data quality through the documentation of all intermediate analytical processes and products [22, 23]. While these type of tools remain somewhat early in their development, their potential benefits are already being demonstrated in the computational biology, bioinformatics, and translational bioinformatics domains, where they have been used to enable the high-throughput and reproducible analyses of large amounts of multi-dimensional bio-molecular instrumentation data [22, 24–26]. Emergent efforts are similarly exploring their applicability to the integrative analysis of clinical phenotype data in combination with such bio-molecular data, in order to achieve translational end-points. Repeating the assessments in the preceding problem areas, and when taken as a whole, the types of

motivating questions one might encounter relative to using systematic and extensible data-analytic pipelining platforms can include:

1. *How can I ensure that my data analysis plan is both efficient and reproducible*?
2. *Are there ways to reduce the workload associated with the repetitive analysis of large-scale and multi-dimensional data sets*?
3. *Can I capture intermediate data products associated with my data analysis plans so that I can perform quality assurance regarding such evaluative processes*?

6.3.4 Dissemination and Exchange of Knowledge Generated Via Research Activities

It is a well-known phenomenon that the time period required to move a basic science discovery into clinical research, and ultimately clinical or population-level practice can span in excess of one or two decades [1, 4, 14, 27]. As noted previously in this chapter, numerous studies have identified the dissemination or exchange of information between various research and operational settings as one of the most pressing issues contributing to such long research, development and implementation lifecycles [14]. A wide variety of BMI tools and methods have been developed that are intended to overcome these barriers, such as web-based communication and collaboration tools, knowledge representation standards and platforms, public data and literature registries/databases and associated query and reporting tools, and evidence-based practice tools such as guideline delivery systems and clinical decision support systems [4, 7, 16, 28]. However, current research and development concerning the implementation and utilization of these types of informatics platforms tends to be focused on distinct domains or settings, rather than conceptualizing and integrating them across the full CTR spectrum. Furthermore, there are a plethora of socio-cultural challenges, including human factors, workflow limitations, and historical or cultural norms of both research and clinical activities, which serve to impede the deployment and use of such BMI approaches and technologies. As such, the area of data, information, and knowledge exchange across traditional organizational and disciplinary boundaries remains an open and vigorous area of BMI research and development. Finally, and again repeating the assessments in the previously described problem areas, and when taken as a whole, the types of motivating questions one might encounter relative to the dissemination and exchange of knowledge generated via research activities can include:

1. *How do I communicate my laboratory findings in a way that will support or enable the rapid design of pre-clinical and or clinical studies informed by the knowledge associated with such findings*?
2. *Are there optimal ways to encode the data, information, and knowledge generated during clinical studies so as to accelerate their dissemination to clinical practioners*?
3. *What is the best way to ensure that new clinical evidence or guidelines are rapidly adopted across a broad clinical or population-level spectrum*?

6.4 An Organizing Framework for Informatics Capable of Driving Clinical and Translational Research

At the most basic level, one can conceptualize of the role of BMI methods tools in terms of driving CTR and addressing the problem areas identified in Sect. 6.3 as falling into four complementary categories, as illustrated in Fig. 6.4.

The first major category of BMI theories and methods that can support and enable CTR are focused on the "**Active Research**" phase of such projects, wherein investigators and their teams need to collect and manage heterogeneous and multi-dimensional data. Examples of the types of tools and technologies used to support this category of information need include database management systems; electronic data capture (EDC) systems; clinical trials/research management systems (CTMS/CRMS); or Electronic Health Record (EHR) systems. Underlying the design and optimal use of such tools and technologies are theoretical and methodological frameworks related to data modeling, semantics, software engineering, and human computer interaction (HCI).

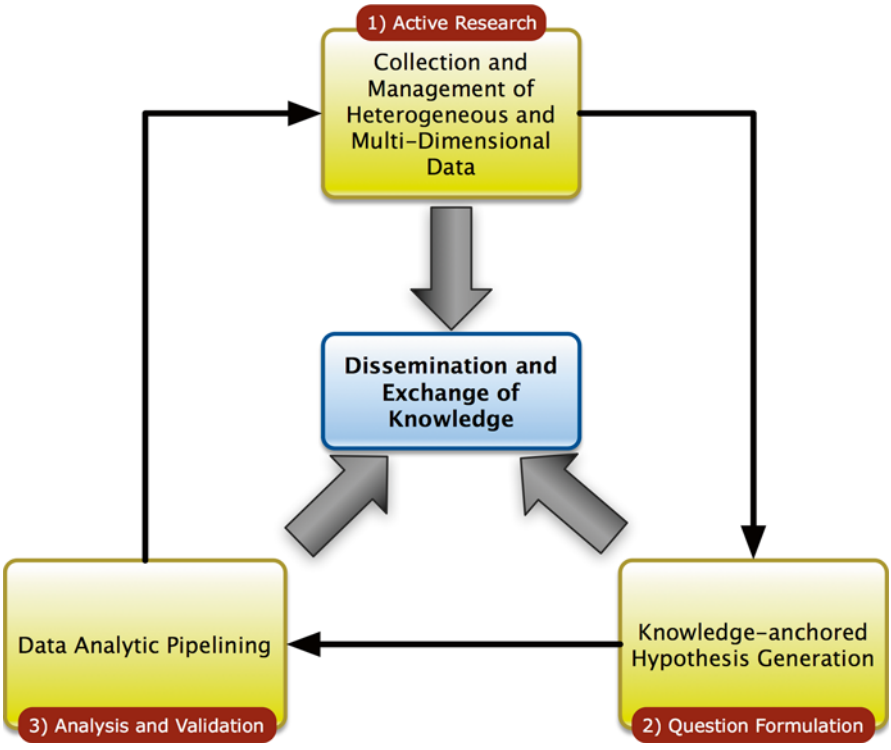

Fig. 6.4 Overview of Biomedical Informatics framework by which methods and tools support and enable the information needs spanning the broad CTR spectrum

The second major category of BMI theories and methods that can support and enable CTR are focused on the "**Question Formulation**" phase of such projects, wherein investigators and their teams need to formulate data-centric research questions for subsequent analysis. Examples of the types of tools and technologies used to support this category of information need include: (1) database query tools; data mining and machine learning "workbenches"; (2) data visualization engines; and (3) knowledge-based *in silico* hypothesis generation systems. Underlying the design and optimal use of such tools and technologies are theoretical and methodological frameworks related to data modeling, information theory, artificial intelligence, knowledge engineering, and data visualization.

The third major category of BMI theories and methods that can support and enable CTR are focused on the "**Analysis and Validation**" phase of such projects, wherein investigators and their teams need to test and validate the results of the research questions generated in the prior phase. Examples of the types of tools and technologies used to support this category of information need include: (1) database query tools; (2) statistical analysis packages; (3) data-centric scripting languages; and (4) visualization tools. Underlying the design and optimal use of such tools and technologies are theoretical and methodological frameworks related to the quantitative data sciences, as well as information theory and HCI.

A fourth and cross-cutting type of BMI theories and methods that can support and enable CTR are focused on "**Dissemination and Exchange of Knowledge**", wherein intermediate and final knowledge products generated via all of the preceding phases are delivered to stakeholders in a variety of human and computer readable formats. Examples of the types of tools and technologies used to support this category of information need include: (1) database management systems, (2) electronic data interchange and sharing infrastructures, and (3) knowledge editing environments. Underlying the design and optimal use of such tools and technologies are theoretical and methodological frameworks related to knowledge engineering, data standards, software architecture, and human computer interaction.

6.5 Discussion

As noted at the outset of this chapter, the design, execution, analysis, and dissemination of results generated via CTR projects are complex and information-intensive endeavors. The ability to efficiently and effectively conduct CTR requires the availability of comprehensive and systematic data, information, and knowledge management tools and methods. Furthermore, the importance of such platforms and techniques is greatly amplified when project involve geographically or temporally distributed teams, as well as when the scientific aims of given project involve the collection of multi-dimensional and heterogeneous data sets (for example, when a study involves the collection and integrative analysis of patient-derived clinical and bio-molecular phenotypes). To this end, we have introduced not only a general set of categories that serve to define the motivating information needs incumbent to

CTR, but also a framework for organizing the BMI theories and methods capable of meeting those needs. These constructs ultimately provide a model for evaluating, understanding, and planning for the BMI requirements of CTR projects, as well as informing the ability to critically evaluate such plans and their implementation.

6.6 Implications for Stakeholders

When viewed in the context of the stakeholder and activities introduced in Chap. 2, the advancement of CTS using BMI theories methods has a large number of potential benefits. Specific examples of these opportunities include the following:

Evidence and Policy Generators

- **Researchers are increasingly approaching complex and systems-level problems via the formation and operation of multidisciplinary teams**. These teams can and will require the close coupling of BMI theories and methods with fundamental CTR design principles in order to adequately address such needs.
- **Policy generators need to be able to better "connect the dots" between basic science research and the clinical application of the findings generated therein**. The integration of BMI and CTR approaches enables the rapid and measurable "cycling" of new evidence between such domains, thus enabling an argument for clinical "actionability" that helps to drive critical policy and funding decisions.

Providers and Healthcare Organizations

- As personalized medicine paradigms become increasingly common, there is **a need for "translational" knowledge that bridges bio-molecular and clinical phenotypes in a way that is useful to "front line" care providers**. The coordinated use of BMI and CTR methods provides an "engine" by which this type of cross-domain and clinically actionable knowledge can be created and delivered to the point-of-care.
- In order to **realize the "triple threat" of improved quality, safety, and cost of care, provider organizations need highly tailored and contextualized information that serves to inform care delivery**. The use of combined BMI and CTR methods provides a basis for creating the knowledge base that is requisite to such decision-making needs.

Patients and Their Communities

- **Patients and their advocates can become critical participants in the CTR cycle** by serving as study participants and/or contributing data from beyond traditional organizational boundaries, thus enhance the scope and reach of such investigatory activities.
- Finally, **communities-at-large can become integral members of the "research fabric"** through participatory and other information gathering methods, thus ensuring that CTR project are targeted upon topics of interest to such groups of constituents.

6.7 Conclusions

The timely and efficient pursue of CTR is central to creating an evidence base that can in turn enable the delivery of knowledge-driven healthcare, wherein the combination of quality, safety, and cost are all optimized. As can be seen in this chapter and the referenced works, the application of BMI theories and methods has significant and wide-ranging potential in terms of achieving such benefit. As such, the tight integration BMI and CTS is of the utmost importance in terms of achieving the TI vision set forth in this book.

Discussion Points

- What types of information needs characterize the CTR environment? How are these requirements different that information needs encountered in the discrete sub-disciplines that make up the overall CTR spectrum?
- How do major BMI method and theories map to the aforementioned information needs? In what cases does such a mapping elucidate gaps in knowledge or practice that represent open areas of research and innovation?
- What barriers prevent the effective and timely application of BMI theories and methods to address the needs of CTR teams?
- How does the increasing use of "big data" serve to influence or effect the preceding factors?
- How do socio-cultural factors play a role in the effective integration of BMI theories and methods with CTR related activities?

References

1. Sung NS, Crowley Jr WF, Genel M, Salber P, Sandy L, Sherwood LM, et al. Central challenges facing the national clinical research enterprise. JAMA. 2003;289(10):1278–87. PubMed PMID: 12633190, Epub 2003/03/14. eng.
2. Payne PR, Embi PJ, Sen CK. Translational informatics: enabling high-throughput research paradigms. Physiol Genomics. 2009;39(3):131–40. PubMed PMID: 19737991, Pubmed Central PMCID: 2789669, Epub 2009/09/10. eng.
3. 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. PubMed PMID: 19261934. Epub 2009/03/06. eng.
4. Chung TK, Kukafka R, Johnson SB. Reengineering clinical research with informatics. J Investig Med. 2006;54(6):327–33. PubMed PMID: 17134616. eng.
5. Ash JS, Anderson NR, Tarczy-Hornoch P. People and organizational issues in research systems implementation. J Am Med Inform Assoc. 2008;15(3):283–9. PubMed PMID: 18308986. ENG.
6. NIGMS. Definition of a clinical trial. Bethesda: National Institute of General Medical Sciences; 2006 [cited 2006 2/25/2006].
7. 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. PubMed PMID: 15974245.
8. Spilker B. Guide to clinical trials. New York: Raven; 1991. xxv, 1156 p.
9. Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. N Engl J Med. 2000;342(25):1887–92. PubMed PMID: 10861325.

10. caBIG. Study Calendar 2.0. In: v2.pdf S, editor. Adobe Acrobat. 2.0 ed. Bethesda: National Cancer Institute; 2006.
11. Juni P, Altman DG, Egger M. Systematic reviews in health care: assessing the quality of controlled clinical trials. BMJ. 2001;323(7303):42–6. PubMed PMID: 11440947.
12. Adair JG. The Hawthorne effect: a reconsideration of the methodological artifact. J Appl Psychol. 1984;69(2):334.
13. Butte AJ. Medicine. The ultimate model organism. Science. 2008;320(5874):325–7. PubMed PMID: 1842092, Pubmed Central PMCID: 2749009, Epub 2008/04/19. eng.
14. Fridsma DB, Evans J, Hastak S, Mead CN. The BRIDG project: a technical report. J Am Med Inform Assoc. 2008;15(2):130–7. PubMed PMID: 18096907. eng.
15. Kaiser J. U.S. budget 2009. NIH hopes for more mileage from roadmap. Science. 2008; 319(5864):716. PubMed PMID: 18258872. eng.
16. Kush RD, Helton E, Rockhold FW, Hardison CD. Electronic health records, medical research, and the Tower of Babel. N Engl J Med. 2008;358(16):1738–40. PubMed PMID: 18420507. eng.
17. Ruttenberg A, Clark T, Bug W, Samwald M, Bodenreider O, Chen H, et al. Advancing translational research with the Semantic Web. BMC Bioinforma. 2007;8 Suppl 3:S2. PubMed PMID: 17493285. eng.
18. Maojo V, García-Remesal M, Billhardt H, Alonso-Calvo R, Pérez-Rey D, Martín-Sánchez F. Designing new methodologies for integrating biomedical information in clinical trials. Methods Inf Med. 2006;45(2):180–5. PubMed PMID: 16538285. eng.
19. Payne PR, Mendonca EA, Johnson SB, Starren JB. Conceptual knowledge acquisition in biomedicine: a methodological review. J Biomed Inform. 2007;40(5):582–602. PubMed PMID: 17482521.
20. Richesson RL, Krischer J. Data standards in clinical research: gaps, overlaps, challenges and future directions. J Am Med Inform Assoc. 2007;14(6):687–96. PubMed PMID: 17712081. eng.
21. Nature Glossary: Nature Publishing Group; 2008 [cited 2008 Nov 15]. Available from: http://www.nature.com.
22. van Bemmel JH, van Mulligen EM, Mons B, van Wijk M, Kors JA, van der Lei J. Databases for knowledge discovery. Examples from biomedicine and health care. Int J Med Inform. 2006;75(3–4):257–67. PubMed PMID: 16198618. eng.
23. Oster S, Langella S, Hastings S, Ervin D, Madduri R, Phillips J, et al. caGrid 1.0: an enterprise grid infrastructure for biomedical research. J Am Med Inform Assoc. 2008;15(2):138–49. PubMed PMID: 18096909. ENG.
24. Howard K. The bioinformatics gold rush. Sci Am. 2000;283(1):58–63.
25. van Beek JH. Channeling the data flood: handling large-scale biomolecular measurements in silico. Proc IEEE. 2006;94(4):692–709.
26. Warr WA. Scientific workflow systems: pipeline pilot and KNIME. J Comput Aided Mol Des. 2012:1–4.
27. Zerhouni EA. Translational and clinical science–time for a new vision. N Engl J Med. 2005;353(15):1621–3. PubMed PMID: 16221788.
28. Sim I. Trial registration for public trust: making the case for medical devices. J Gen Intern Med. 2008;23 Suppl 1:64–8. PubMed PMID: 18095047. eng.

Additional Reading

Ash JS, Anderson N, Tarczy-Hornoch P. People and Organization issues in research systems implementation. J Am Med Inform Assoc. 2008;15(3):283–9.
Bernstam EV, Hersh WR, Johnson SB, et al. Synergies and distinctions between computational disciplines in biomedical research: perspective from the clinical and translational science award programs. Acad Med. 2009;84(7):964–70.

Buetow KH. Cyberinfrastructure: empowering a “third way” in biomedical research. Science. 2005;308(5723):821–4.
Chung TK, Kukafka R, Johnson SB. Reengineering clinical research with informatics. J Investig Med. 2006;54(6):327–33.
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.
Payne PR, et al. Breaking the translational barriers: the value of integrating biomedical informatics and translational research. J Investig Med. 2005;53(4):192–200.
Payne PR, Embi PJ, Sen CK. Translational informatics: enabling high throughput research paradigms. Physiol Genomics. 2009;39(3):131–40.
Sung NS, Crowley Jr WF, Genel M, et al. Central challenges facing the national clinical research enterprise. JAMA. 2003;289(10):1278–87.

Chapter 7
Using Big Data

Nigam H. Shah

By the End of this Chapter, Readers Should be Able to

- Define what is big data
- Understand how big data change biomedical science
- Identify the key questions to ask when analyzing big data
- Understand the limitations to be aware of when reasoning with or analyzing big data.

7.1 Introduction

"Big Data" is term that is difficult to fully define. As a concept, Big Data have existed since there was ever a notion of data that required mechanical devices for processing or analyzing. Indeed, the first notion of Big Data may very well be attributed to the motivation of Charles Babbage's differential and analytic engines, which were originally contemplated to develop nautical charts. In recent years, Big Data are defined as those that data that are challenging relative to available computational processing [1, 2]. Generally speaking, data sets that are so large and complex that traditional data processing applications are unable to process them are referred to as Big Data. Of course, what is traditional is highly dependent on the field of use; and what is Big Data in healthcare, might be considered routine (or even "small") in high energy physics research. Therefore, it is necessary to understand the boundary at which a field would term a dataset as "Big" and to understand the associated change in mindset that Big Data analysis entails [3].

N.H. Shah, MBBS, PhD
Department of Medicine (Biomedical Informatics),
Stanford University, 1265 Welch Road, Stanford, CA 94305, USA
e-mail: nigam@stanford.edu

P.R.O. Payne, P.J. Embi (eds.), *Translational Informatics: Realizing the Promise of Knowledge-Driven Healthcare*, Health Informatics,
DOI 10.1007/978-1-4471-4646-9_7,

As argued in the Wired magazine article "The End of Theory," we need a change in our way of thinking about science when faced with very large datasets [4]. Typical statistical approaches breakdown with sufficiently large data and the challenge becomes to measure and to interpret the data instead of hypothesis-driven experimentation [3, 4]. One of the supporting case studies in the Wired magazine article is a tool that predicts crop yields[1] down to the level of individual fields by sifting through over 50 terabytes of data. In the legal[2] domain, big data analysis is replacing armies of lawyers with tireless computer algorithms for task of *discovery*, i.e., producing relevant materials for a case by examining millions of emails, memorandums, and other documents to extract relevant information.[3]

It is time for biomedicine to embrace the trend because historical barriers to the adoption of EHR systems are giving way to new Federal incentives, resulting in the collection of medical data at an unprecedented scale [5]. The Health Information Technology for Economic and Clinical Health Act (HITECH) legislation calls for meaningful use objectives[4] to measure progress, to sustain early adoption, and to provide accountability. In parallel, new kinds of datasets, such as next-generation sequencing datasets and personalized omic profiles [6, 7], are creating large amounts of data on individual patients that cannot be analyzed or reviewed by a doctor in a 15 min office visit. New approaches to capture, store, analyze and interpret such massive datasets are urgently needed.

7.2 The Kinds of Big Data in Medicine and Analyses They Enable

The discussion of Big Data in translational informatics frequently connotes next-generation sequencing data [8–10]. However, this is beginning to change: the use of large datasets of various kinds increased dramatically in recent years. 'Big Data' is an increasingly comprehensive term, including both large amounts of *molecular measurements* on a person (e.g., next-generation sequencing) as well as small amounts of *routine measurements* on a large number of people (e.g., clinical notes, lab measurements, claims data and adverse event reports).

Imagine how scientific inquiry, and the ability of our healthcare system to 'learn' [5], would be different if we collect and share access to lots of data—both genomic and "routine." How will the kinds of questions we ask change when we cross a certain data-threshold? [3, 11]. Outside of healthcare and biomedicine, a small amount of data about millions of individuals is already being collected and mined by Web

[1] http://www.wired.com/science/discoveries/magazine/16-07/pb_feeding.

[2] http://www.wired.com/science/discoveries/magazine/16-07/pb_lawsuit.

[3] http://www.nytimes.com/2011/03/05/science/05legal.html.

[4] http://edocket.access.gpo.gov/2010/pdf/2010-17207.pdf.

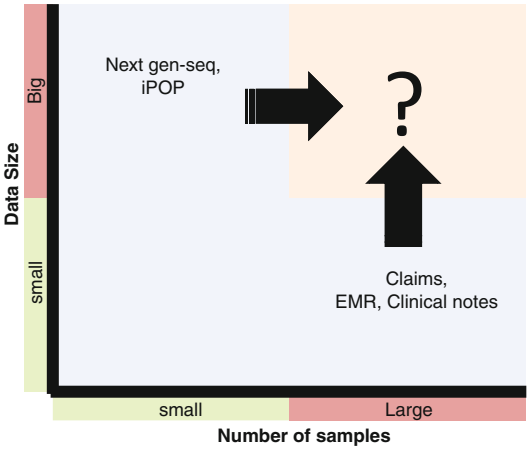

Fig. 7.1 Simply put, data can be big in the amount of measurements on an individual (e.g., next generation sequencing) or can be big in the number of individuals on whom there are some measurements (e.g., clinical notes, laboratory measurements, claims). Naturally, it is exciting to imagine what happens when we reach the upper right quadrant

companies (e.g., a typical social network profile, when exported is a couple of gigabytes) resulting in a gold rush around analyzing this "digital exhaust"[5]

The idea of using data for enhancing health and well-being is popular in groups such as the *Quantified Self* collaborative and other self-tracking groups.[6] Given the rising popularity of such efforts and the increasingly sophisticated collection of phenotypic data enables "mass phenotyping," which is the collection and integration of massive amounts of diverse phenotypical information (continuous or categorical variables) in order to discover latent patterns and correlate those patterns with health and wellbeing [12].

In thinking about Big Data in healthcare—genomic, medical, environmental or personal phenotypic—it is essential to think about the dimensions along which the data are big. For example, genomic data are big in size but relatively small in numbers of samples; whereas claims data are small in size but are available for over 100 million individuals. Thinking along these two axes forms a sector-map (Fig. 7.1), which aids in thinking about potential analyses and computational solutions to use.

Finally, both disease and its treatment are processes that unfold over time. Hence, it is essential to understand the nature and temporal density of any dataset that is used. Continuous time-traces such as those collected by an electrocardiogram monitor are very different from billing data, which are collected only when a person gets sick and interacts with the health system. Similarly genomic data are usually a one-time measurement and are rarely re-collected over time, except in highly specialized situations such as studying the response of a tumor to specific anti-neoplastic drugs.

Depending on the axis along which the data are dense (samples, variables, and time in Fig. 7.2), different methods apply and lead to different insights. For example, relatively simple methods based on recognizing mentions of drugs, diseases,

[5] http://www.vlab.org/article.html?aid=304.

[6] http://quantifiedself.com/about/.

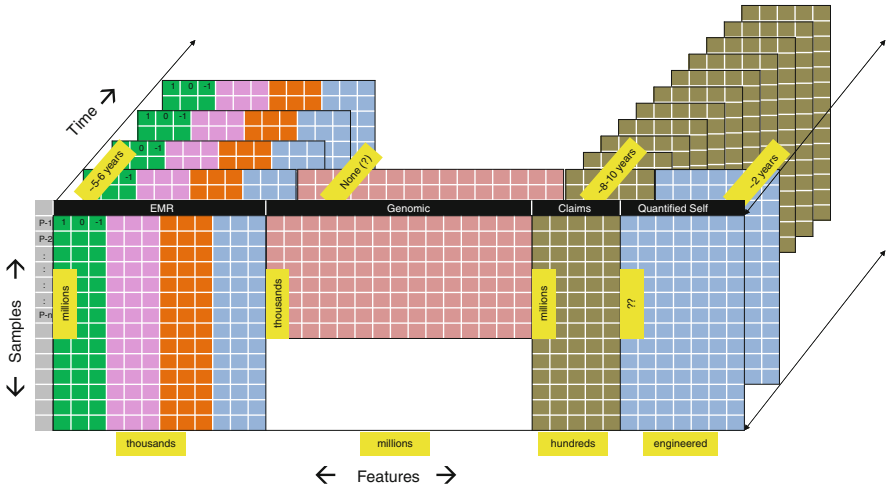

Fig. 7.2 Data can be "big" along three dimensions; the number of samples (*rows*), the number of features (*columns*), and temporal coverage (frequency at which data are collected, and how far back the data go). Different sources have different shape along these three axes. For example, genomic data have millions of features (e.g., SNPs) but are available for a small number of people (currently in the thousands). Medical billing and claims data have fewer features (e.g., ICD-9 codes, laboratory codes) but are available for millions of individuals over a much longer period of time

and adverse events in electronic medical records could identify six out of nine drugs recalled in the past decade, roughly 2 years ahead, when incorporating the time dimension into the analysis [13].

When we cross a certain data-threshold the kinds of questions we can ask changes, consider the example by Frankovich et al. [14]. Where the existing literature and evidence based guidelines were insufficient to guide the clinical care of a patient, Frankovich et al. applied trend analysis to the EMR data from 98 patients to "learn" a data-driven guideline on how to provide care for a 13-year-old girl with systemic lupus erythematosus (SLE) [14]. In terms of size, the Electronic Health Record data from 98 patients are certainly not "big" as would be the case with full genome sequences from 98 individuals. However, such approaches, which analyze data that are already routinely collected, are particularly valuable when a formal guideline is not available or feasible from a practical standpoint—we refer to such approaches as practice-based evidence [15].

As another example, Leeper et al. show that it is possible to examine the outcomes of decisions made by doctors during clinical practice to identify patterns of care that are optimal—generating evidence based on the collective practice of experts. The authors examined the validity of a black-box warning on an effective drug (Cilostazol) for peripheral vascular disease. Cilostazol is the same kind of drug as Milrinone, which in a study in 1991 was found to be associated with sudden cardiac death when used to treat heart failure; as a result, Cilostazol "inherited" a

black-box warning contraindicating its use in any patients with any kind of heart failure. Using over 5,892 patient years of data, they observed no association between Cilostazol use and major adverse cardiovascular events. The findings recapitulate a prospective clinical study, which had less than half the patient years of data and failed to reach a statistically significant conclusion. In addition, the data-mining study was able to profile a subset of patients with CHF who were prescribed Cilostazol despite its black box warning and examine its safety in this high-risk group of patients—something that could not be done prospectively, given the black-box warning on the drug.

Analyses using "Big Data" go beyond learning biomedical insights as outlined in the examples above. The personalization of cancer chemotherapy by examining the genomic variations that drive specific tumor subtypes and the response of those subtypes to specific chemotherapy drugs is already being used to personalize treatment of breast cancer [16][7] and is likely to be among the first clinical success stories about the application of data science to identify alternative treatment strategies.

As discussed in the chapter on personalized medicine, consider the example of personalizing treatment based on genomic sequencing, where Dr. Howard Jacob's team pinpointed a new casual mutation for the treatment of a 6 year-old boy with an extreme form of inflammatory bowel disease [17, 18]. The authors diagnosed an X-linked inhibitor of apoptosis deficiency, based on which, they decided to perform an allogeneic hematopoietic progenitor cell transplant. This case report demonstrates the power of using genomic data to arrive at a molecular diagnosis in an individual patient in the setting of a novel disease.

On the operational side—which is related to the practice and delivery of healthcare—the use of Big Data has gained a lot of momentum. Predictive-analytic approaches, such as those designed to predict readmissions [19], are increasingly gaining traction[8] and have direct implications for reducing cost, as well as maintaining economic viability of health care delivery systems in the light of new regulation. The main drivers for using Big Data on the operational side are [20]:

- The Patient Protection and Affordable Care Act, and the creation of accountable care organizations (ACOs), which require that healthcare systems have a higher degree of "business intelligence".
- Inefficiencies, fraud, and waste; where big data can play a crucial role in performance improvements.
- Adoption of open-data policies by the U.S. Department of Health and Human Services, which spark innovation and increase transparency in health care.

As more data are made available, and health systems are increasingly under pressure to provide the same or better quality care at less cost, it is natural to use data to increase efficiency, design less costly care workflows, increase intervention

[7] http://breakthroughs.cityofhope.org/molecular-subtyping-chemotherapy/5946/.

[8] http://www.beckershospitalreview.com/healthcare-information-technology/4-steps-to-leveraging-qbig-dataq-to-reduce-hospital-readmissions.html.

targeting, and to preserve as well as improve quality. In some sense, the direct application of Big Data analytics for healthcare is late to the game and is making a transition that several other industries—such as airlines, weather prediction, and actuarial sciences—made several years ago.

7.3 Discussion: Hot Topics and Future Directions

As the use of data in health care grows, there are several sources of data that need to be synthesized and integrated. The key sources to watch are:

- Clinical: Electronic Health Records
- Genomic data
- Research and clinical trial data
- Patient reported data (via personal health records, surveys or mobile apps)
- Social media data (Twitter, Facebook, Google Plus)
- Billing, claims and financial data
- Sensors collecting data on people and their environment
- Public health surveillance and health care utilization data (AHRQ; DHHS)

Approaches that go across data-sources and that attempt predictive data-mining—such as predicting falls before they happen (via carpet sensors), readmissions [19], suicides [21], depression [22], abuse [23] —are fruitful directions to apply Big Data analyses.

Along with the increase in data availability, there is increased participation, ownership and stewardship in using data by the "engaged patient". Instead of being a by-stander in one's own care, patients are empowering themselves with raw data (e.g., Quantified self) and other individual's experiences (e.g., www.patientslikeme.com) as well as the wisdom of other patients (e.g., on forums such as www.smartpatients.com).

The idea of using user generated content for enhancing health is exemplified by the *Quantified Self* collaborative, which lists over 500 modalities of collecting raw data from an individual for self-tracking.[9] Given the popularity of such efforts, the application of Big Data analysis for mass phenotyping to discover patterns and correlate those patterns with health and well-being, is bound to increase [12].

Finally, it is important to realize that just because vast amounts of data are available we are not guaranteed to find better insights. The results based on Big Data will only be as good as the analysis methods employed, and there is therefore an urgent need for new formal science methods as advocated by the Data Science movement[10] to help with Big Data interpretation.

[9] http://quantifiedself.com/about/.

[10] http://bigdatablog.emc.com/2012/11/09/openchorus-project-the-dawn-of-the-data-science-movement/.

7.4 Implications for Stake Holders

As we have done in prior chapters, we can review the implications of "Big Data" as it applies to the broad biomedical and health sciences domains by organizing such thoughts round the major stakeholder groups we have already enumerated. Specifically, we can determine that:

Evidence and Policy Generators

- **Although the use of Big Data holds great promise, sustained progress requires that cultural and ethical issues related to patient privacy and data ownership be addressed.** Consider a simplified health care setup with a medical group, a hospital and an insurer. Usually the hospital pays for and installs an electronic health record system, the practitioners from the medical group enter the data, which in turn is generated from patients during the provision of care; the care is paid for by the insurer. Even in this simplified set up who owns the data? Most stake-holders agree that use of data for improving quality of care is acceptable use. However, most other uses require some manner of consent, oversight or both. The issues get more complicated when the data under consideration can also affect *someone elses's* privacy, as is the case with genetic data where disclosing an individual's data also discloses some of their parent's and sibling's data. As Larson argues, we need to build trust in the power of Big Data to serve the public good [24].

Providers and Healthcare Organizations

- **With the shift from a fee-for-service to a pay-for-performance model, providers have to interact in a zero-sum game to figure out what practices work, which approaches are most efficient, how to track each participants' contribution to a patient's care, and how to keep patients engaged in their own care** [25, 26]. Efforts toward close coordination and patient outreach will benefit immensely.

Patients and Their Communities

- **The primary stake-holders in the decision of, and the impact of using Big Data analyses are the patients**, and by extensions, the researchers, the healthcare providers, the payers, and the regulators that collectively impact the healthcare delivered to such individuals. The "providers" group in this relationship is the most heterogeneous group encompassing individuals (e.g., physicians or nurse practitioners), small group practices, hospitals, and (indirectly) device and pharmaceutical companies.

For all of the preceding stakeholders, it is crucial to have clarity on the goal of using Big Data analyses. The goals fall into two broad groups: (1) enhancing the practice of medicine; and (2) advancing biomedical science. It is quite common to confuse these two goals and promise new cures in a few years after a research publication. While it is true that both biomedical science and clinical practice stand to

benefit from using Big Data, it is important for stakeholders to appreciate the pace at which the benefits are realized unfolds on vastly different time scales. Research is inherently exploratory, with false starts and frequent dead ends. Using Big Data can speed the process along, but it is possible that the basic nature of the activity (i.e., the high failure rate) does not change. At its core, the practice and delivery of health care, is an operational problem and Big Data approaches may have their first impact in optimizing the conveyer belts of healthcare. It is imperative that scientists do not oversell their research use cases that advance the science and that the enterprise of health care delivery pay attention to the use of Big Data to improve the practice of medicine.

7.5 Conclusion

In summary, Big Data, which is widely used in retail industries as a means to understand consumers' purchasing habits and preferences, is increasingly being used in health care—both for advancing medical science as well as improving the delivery of healthcare. In order to achieve the desired goals from the analysis of Big Data, it is imperative for stakeholders to be clear on the end goals and the sources of data that need to be integrated. Privacy and data ownership are key issues that should not be ignored. The axis along which data are voluminous—samples, features, or time—affects the kind of analyses that are possible and is a key question to ask along with the goal of the analysis. This is further complicated by the rapid rate at which data are available (velocity), the heterogeneity of the data (variety), and the ability to trust the inferences made from the data (veracity).

It is important for all stakeholders to recognize that research is inherently exploratory, while the practice and delivery of care is an operational exercise. Big Data approaches may have their lasting impact in optimizing the delivery of health care. Finally, data will be king; therefore initiatives such as the "blue button", which empower patients in controlling the generation and the access to their healthcare data, are developments that stake-holders should monitor.

Discussion Points

- Review the dimension on which data can be big: Samples, Variables, and Time. Discuss what kind of data are big along which axes.
- Discuss how the applicable analytic methods change as datasets differ along the dimensions of: number of samples, number of variables, and number of measurements over time.
- Big Data analysis can be approached in a problem-first or methods-first manner. Each approach results in different solutions to the question at hand. Discuss the trade-offs.
- What problems would you solve with Big Data analysis?

References

1. Zikopoulos P. Understanding big data: analytics for enterprise class Hadoop and streaming data. New York: McGraw-Hill; 2012.
2. Franks B. Taming the big data tidal wave: finding opportunities in huge data streams with advanced analytics. Hoboken: Wiley; 2012.
3. Halevy A, Norvig P, Pereira F. The unreasonable effectiveness of data. Intell Syst IEEE. 2009;24(2):8–12.
4. Anderson C. The end of theory: the data deluge makes the scientific method obsolete. Wired Mag. 2008;16–7. http://archive.wired.com/wired/issue/16-07
5. Friedman C, Wong A, Blumenthal D. Achieving a nationwide learning health system. Sci Transl Med. 2010;2(57):57cm29.
6. Li-Pook-Than J, Snyder M. iPOP goes the world: integrated personalized Omics profiling and the road toward improved health care. Chem Biol. 2013;20(5):660–6.
7. Chen R, Mias GI, Li-Pook-Than J, Jiang L, Lam HY, Miriami E, et al. Personal omics profiling reveals dynamic molecular and medical phenotypes. Cell. 2012;148(6):1293–307.
8. Schadt EE, Linderman MD, Sorenson J, Lee L, Nolan GP. Cloud and heterogeneous computing solutions exist today for the emerging big data problems in biology. Nat Rev Genet. 2011;12(3):224.
9. Ashley EA, Butte AJ, Wheeler MT, Chen R, Klein TE, Dewey FE, et al. Clinical assessment incorporating a personal genome. Lancet. 2010;375(9725):1525–35.
10. Samani NJ, Tomaszewski M, Schunkert H. The personal genome–the future of personalised medicine? Lancet. 2010;375(9725):1497–8.
11. Hays J, Efros AA. Scene completion using millions of photographs. Commun ACM. 2008;51(10):87–94.
12. Shah NH. Translational bioinformatics embraces big data. Yearb Med Inform. 2012;7(1):130–4.
13. LePendu P, Iyer SV, Bauer-Mehren A, Harpaz R, Mortensen JM, Podchiyska T, et al. Pharmacovigilance using clinical notes. Clin Pharmacol Ther. 2013;93(6):547–55.
14. Frankovich J, Longhurst CA, Sutherland SM. Evidence-based medicine in the EMR era. N Engl J Med. 2011;365(19):1758–9.
15. Leeper NJ, Bauer-Mehren A, Iyer SV, Lependu P, Olson C, Shah NH. Practice-based evidence: profiling the safety of cilostazol by text-mining of clinical notes. PLoS One. 2013;8(5):e63499.
16. Gluck S, de Snoo F, Peeters J, Stork-Sloots L, Somlo G. Molecular subtyping of early-stage breast cancer identifies a group of patients who do not benefit from neoadjuvant chemotherapy. Breast Cancer Res Treat. 2013;139(3):759–67.
17. Worthey EA, Mayer AN, Syverson GD, Helbling D, Bonacci BB, Decker B, et al. Making a definitive diagnosis: successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease. Genet Med. 2011;13(3):255–62.
18. Mayer AN, Dimmock DP, Arca MJ, Bick DP, Verbsky JW, Worthey EA, et al. A timely arrival for genomic medicine. Genet Med. 2011;13(3):195–6.
19. de Lissovoy G. Big data meets the electronic medical record: a commentary on "identifying patients at increased risk for unplanned readmission". Med Care. 2013;51(9):759–60.
20. Ranck J. Sector RoadMap: health care and big data in 2012. Giga Omni Media, Inc., 2012.
21. Pestian J, Matykiewicz P, Cohen K, Grupp-Phelan J, Richey L, Meyers G, et al. Suicidal thought markers: a controlled trial examining the language of suicidal adolescents. In: 46th American Association of Suicidology Annual Conference, Austin, 2013.
22. Kennedy SH, Downar J, Evans KR, Feilotter H, Lam RW, MacQueen GM, et al. The Canadian Biomarker Integration Network in Depression (CAN-BIND). Pharm Des. 2012;18(36):5976–89.
23. Reis BY, Kohane IS, Mandl KD. Longitudinal histories as predictors of future diagnoses of domestic abuse: modelling study. BMJ. 2009;339:b3677.

24. Larson EB. Building trust in the power of "big data" research to serve the public good. JAMA. 2013;309(23):2443–4.
25. Moses 3rd H, Matheson DH, Dorsey ER, George BP, Sadoff D, Yoshimura S. The anatomy of health care in the United States. JAMA. 2013;310(18):1947–63.
26. Moore KD, Eyestone K, Coddington DC. The big deal about big data. Healthc Financ Manag. 2013;67(8):60–6. 8.

Additional Reading

Data Hero Awards 2011. http://www.greenplum.com/sites/default/files/EMC_Greenplum_2011_DataHeroBook.pdf.

Kohane IS, Drazen JM, Campion EW. A glimpse of the next 100 years in medicine. N Engl J Med. 2012;367:2538–9.

Larson EB. Building trust in the power of "big data" research to serve the public good. JAMA. 2013;309(23):2443–4.

Swan M. The Quantified Self: Fundamental Disruption in Big Data Science and Biological Discovery. Big Data. 2013;1(2):85–99. doi:10.1089/big.2012.0002.

Chapter 8
In Silico Hypothesis Discovery

Philip R.O. Payne

By the End of This Chapter, Readers Should Be Able to:

Understand the role of conceptual knowledge collections in terms of informing the design and use of reasoning systems for the purpose of in silico hypothesis discovery

- Select appropriate evaluation methodologies that can be used the assess the performance of in silico hypothesis discovery tools and platforms
- Identify open research questions related to the future of high-throughput hypothesis generation and the impact of such innovations on current and future scientific and healthcare delivery paradigms.

8.1 Introduction

As noted in the preceding chapters, the fundamental methods needed to conduct basic science, and clinical and translational research are very complex, involving a multitude of actors, workflows and data types. For example, the translational research paradigm focuses on cyclical flow of data, information and knowledge between laboratory researchers, clinical investigators and clinical or public health practitioners, and is predicated on systems-level approaches that involve diverse information needs, sources and management requirements [1]. A variety of reports and scholarly works have enumerated challenges that may prevent the effective conduct of translational research. As introduced in Chap. 1, one such challenge is commonly known as the "T1 block" and is concerned with issues that impact the ability to move data, information and knowledge between basic science and clinical research settings. Similarly, a second challenge, often known as the "T2 block", focuses upon impediments affecting the movement of data, information and knowledge between

P.R.O. Payne, PhD, FACMI
Department of Biomedical Informatics,
The Ohio State University Wexner Medical Center, Columbus, OH, USA
e-mail: philip.payne@osumc.edu

P.R.O. Payne, P.J. Embi (eds.), *Translational Informatics: Realizing the Promise of Knowledge-Driven Healthcare*, Health Informatics,
DOI 10.1007/978-1-4471-4646-9_8,

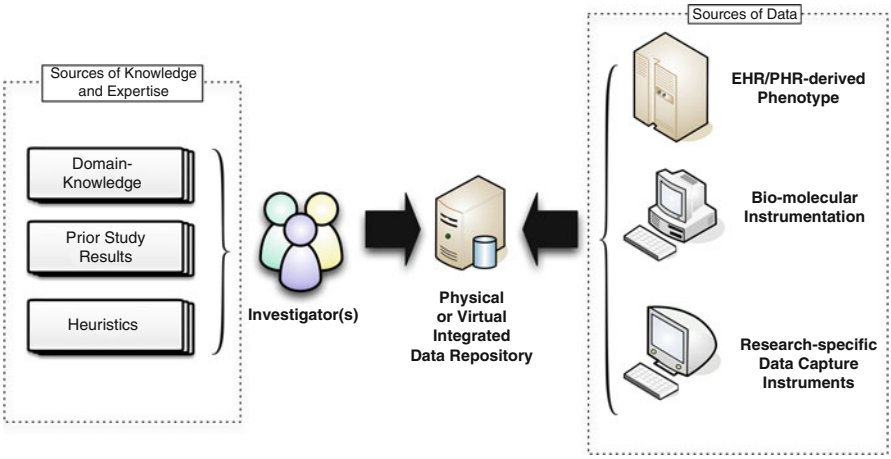

Fig. 8.1 Overview of traditional, investigator-driven approach to asking and answering questions regarding complex and large scale data sets. In this model, the investigator (or research team) serves as the primary integration of various sources of knowledge and expertise, formulating and asking questions concerning available data using a combination of their domain knowledge, experiential knowledge from prior studies, and heuristics that they may have formulated relative to an application domain

the clinical research environment and clinical or public health practice [2]. For both of these categories of challenges, the methods required to address them are extremely reliant on the provision of tools and methods that can facilitate the collection, formalization, analysis and dissemination of large-scale and integrative data sets [3]. The potential impact of informatics-based approaches in terms of addressing such information needs has been well established; yet those same tools and methods remain largely under-utilized by the research and practice communities [4–12].

Within this broad context, one major area of concern is the way in which we formulate and test hypotheses relative to "big" biomedical data. This concern is amplified by the fact that the volume, velocity and variability of biomedical data continue to expand at a rapid rate. This growth is in large part a function of the proliferation of computerized sources of biomedical data, such as Electronic Health Records (EHRs), Personal Health Records (PHRs), Clinical Trial or Research Management Systems (CTMS/CRMS), high-throughput bio-molecular instrumentation, and ubiquitous sensor technologies. While computational methods continue to be devised and applied to support or enable the capture, storage and transaction of these data sets, there has not been a corresponding focus on improvements in the ways in which we ask and answer important questions utilizing this data. In fact, the traditional, reductionist approach to intuitive hypothesis generation based on the expertise or insights of an individual or small number of investigators remains the norm (Fig. 8.1). However, this approach is highly linear, and limited by the cognitive capacities of such investigators or teams, leading to an underutilization of available and costly to assemble data sets.

In effect, we continue to create and maintain bigger and more complex data sets at great expense, while we ask and answer small numbers of questions regarding the

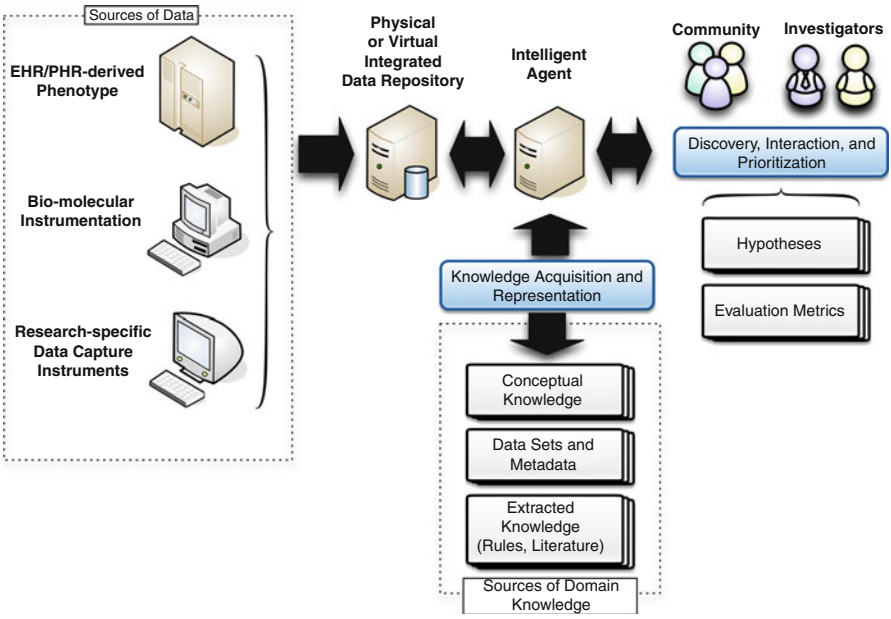

Fig. 8.2 Alternative, high-throughput approach to asking and answering questions regarding "big data" resources, using in silico hypothesis discovery methods. In this model, intelligent computational agents draw upon a variety of domain knowledge collections, using formally represented variants of those collections, in order to identify potential relationships of interest between elements or collections of elements in an integrated repository. These relationships are then presented, along with corresponding evaluation metrics that serve to characterize their potential accuracy and novelty, to both investigators and their teams as well as broader groups of interested community members, who can then discover, interact with, and prioritize such hypotheses concerning data-level interactions for subsequent investigation

contents of those data sets using methods that are not far removed from those used around the time of the dawn of modern science [13]. This concerning juxtaposition is the driver for an emerging body of research that seeks to couple high-throughput data generation with new and similarly high-throughput hypothesis generation techniques, which can at a high level be referred to as ***in silico hypothesis discovery methods*** (Fig. 8.2).

Such high-throughput approaches to asking and answering questions corresponding to "big data" resources are essential to the synthesis of novel biomedical knowledge, such as that required to support personalized medicine paradigms. Such precision approaches to wellness promotion and care delivery aim to improve quality, outcomes and cost of care [2, 3, 14–16]. Acting upon this vision of high-throughput in silico hypothesis discovery requires:

1. An understanding of the design and appropriate use of domain-specific conceptual knowledge collections;
2. The application of intelligent agents that are informed by such knowledge collections and based upon formal computational methods; and
3. The evaluation of ensuing hypothesis using appropriate metrics and measures.

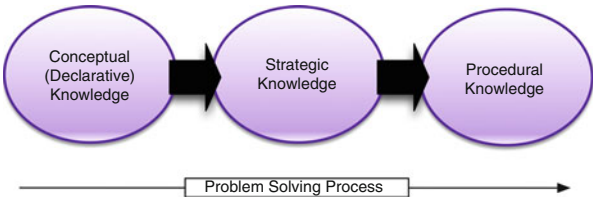

Fig. 8.3 Spectrum of knowledge types, spanning from conceptual to strategic to procedural knowledge, where conceptual knowledge is the most abstract form of understanding a domain, and procedural knowledge is the most application- or problem-oriented understanding of a given need or task

In the following sections, we will explore critical aspects of all three of the aforementioned foundational dimensions that underpin the design and use of in silico hypothesis discovery tools and platforms.

8.2 Conceptual Knowledge in Biomedicine

Conceptual knowledge has been defined in the computational, psychology, and education literature as being comprised of a combination of atomic units of information *and* the meaningful relationships between those units. The same literature goes on to define two additional types of complementary knowledge, known as procedural and strategic knowledge respectively. *Procedural knowledge* is a process-oriented understanding of a given problem domain [17–20], effectively concerned with the methods and approaches used to solve a given problem or address a task. *Strategic knowledge* is that which is used by individuals in order to translate conceptual knowledge into procedural knowledge [19] (Fig. 8.3).

Of note, these definitions are based upon a wide-ranging collection of empirical research on learning and problem-solving in complex scientific and quantitative domains such as mathematics and engineering [18, 20]. The cognitive science literature provides a very similar and confirmatory differentiation of knowledge types, making the distinction between procedural and declarative knowledge. Declarative knowledge in this context is synonymous with conceptual knowledge as defined previously [21].

Conceptual knowledge collections in the biomedical domain include a variety of constructs such as ontologies, controlled terminologies, semantic networks and database schemas. A common theme when considering the existing state-of-the-art relative to the design and use of conceptual knowledge collections in the biomedical domain is the need for systematic and rigorous processes for representing conceptual knowledge in a computable form. It is also important to note when considering the need for such knowledge representation best practices that conceptual knowledge collections rarely exist in isolation. Instead, they usually occur within structures that contain multiple types of knowledge. For example, a modern clinical decision support system (CDSS) might include: (1) a database of potential find-

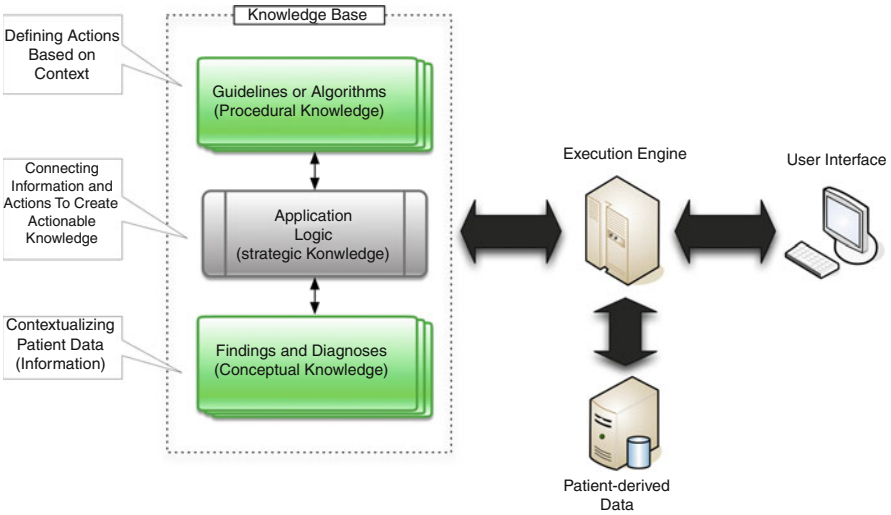

Fig. 8.4 Overview of a prototypical CDSS platform, incorporating conceptual, procedural, and strategic knowledge types in order to generate actionable knowledge from patient-derived data

ings, diagnoses and the relationships between them (*conceptual knowledge*); (2) a set of guidelines or algorithms used to reason upon the preceding database (*procedural knowledge*); and (3) a formal definition of the logic used to operationalize the preceding two knowledge collections (*strategic knowledge*) (Fig. 8.4).

It is only when these three types of knowledge are combined that it is possible to realize a functional decision support system [22]. Given the close similarities between such CDSS and the previously introduced framework for in silico hypothesis discovery methods or tools (as is illustrated in Fig. 8.2), this phenomenon is important to keep in mind for the remainder of this chapter.

8.2.1 Knowledge Engineering

The core theories and methods that underlie the ability to systematically and rigorously represent conceptual knowledge inform a set of application-level techniques known as knowledge engineering (KE). The KE process (Fig. 8.5) incorporates four major steps:

1. Acquisition of knowledge (KA)
2. Representation of that knowledge (KR) in a computable form
3. Implementation or refinement of intelligent agents (e.g., applications that use formally represented knowledge to reason upon data sets and generate results of interest to end-users) or applications
4. Verification and validation of the output of those knowledge-based agents or applications against one or more reference standards.

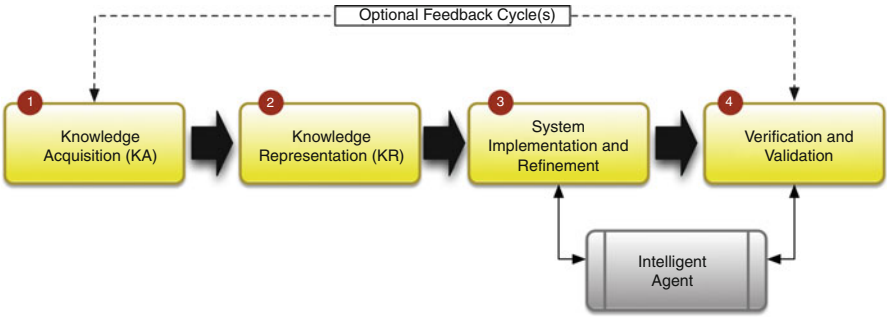

Fig. 8.5 Overview of the Knowledge Engineering (*KE*) process, consisting of knowledge acquisition (*KA*), knowledge representation (*KR*), system implementation and refinement, and the verification and validation of those systems (numbered per the steps enumerated in Sect. 8.2.1). Of note, there is an optional feedback mechanism from the verification and validation results back to the initial KA component, which helps to inform subsequent KA activities and the refinement of existing knowledge bases

With regards to the final step of the KE process (verification and validation), the reference standards used to evaluate the performance of an intelligent agent can include expert performance measures, requirements acquired before designing the knowledge-based system and/or requirements that were realized upon implementation of the knowledge-based system. In this context, verification is the process of ensuring that the knowledge-based system meets the initial requirements of the potential end-user community. In comparison, validation is the process of ensuring that the knowledge-based system meets the realized requirements of the end-user community once a knowledge-based system has been implemented [23].

8.2.2 Theoretical Frameworks for KE

Underlying the KE process is a set of theories concerning the ability to acquire and represent knowledge in a computable format, which is known as the physical symbol hypothesis. First proposed by Newell and Simon [24], and expanded upon by Compton and Jansen [25], the physical symbol hypothesis argues that knowledge consists of both symbols of reality, and relationships between those symbols. This definition of knowledge thus allows for the creation of "physical symbol systems" (e.g., conceptual knowledge collections), which are defined as:

> …a set of entities, called symbols, which are physical patterns that can occur as components of another type of entity called an expression (or symbol structure). Thus, a symbol structure is composed of a number of instances (or tokens) of symbols related in some physical way (such as one token being next to another). At any instant of time the system will contain a collection of these symbol structures. [26]

In a similar manner, it has been argued within the KE literature that the psychological constructs used by experts can be used as the basis for informing the design and composition of conceptual knowledge collections [27]. This argument is based on a framework for expertise transfer known as Kelly's Personal Construct Theory (PCT). PCT defines humans as "anticipatory systems", where individuals create templates, or constructs that allow them to recognize situations or patterns in the "information world" surrounding them. These templates are then used to anticipate the outcome of a potential action given knowledge of similar previous experiences [28]. Kelly views all people as "personal scientists" who make sense of the world around them through the use of a hypothetico-deductive reasoning system. The details of PCT help to explain how experts create and use such constructs. Specifically, Kelly's fundamental postulate is that "*a person's processes are psychologically channelized by the way in which he anticipated events*" [28]. This is complemented by the theory's first corollary, which is summarized by his statement that [28]:

> Man looks at his world through transparent templates which he creates and then attempts to fit over the realities of which the world is composed… Constructs are used for predictions of things to come… The construct is a basis for making a distinction… not a class of objects, or an abstraction of a class, but a dichotomous reference axis.

Building upon these basic concepts, Kelly goes on to state in his Dichotomy Corollary that "*a person's construction system is composed of a finite number of dichotomous constructs*" [28]. Finally, the parallel nature of personal constructs and conceptual knowledge is illustrated in Kelly's Organization Corollary, which states, "*each person characteristically evolves, for his convenience of anticipating events, a construction system embracing ordinal relationships between constructs*" [27, 28].

When taken as a whole, the two preceding theoretical frameworks provide the basic premises for arguing that:

1. Domain experts (e.g., humans) use personal constructs that roughly approximate those constructs that define formal knowledge (e.g., conceptual, strategic, and procedural knowledge), so as to make sense of the "information world" surrounding them;
2. Formal knowledge can be represented in a computationally tractable format, based upon the physical symbol hypothesis, and again, such symbolic systems closely approximate the definitions of conceptual knowledge; and
3. Knowledge engineering methods, and in particular, knowledge acquisition techniques, provide a set of tools for the elicitation and representation (in computable formats) of domain expert knowledge, helping to bridge the two preceding and complementary postulates.

Thus, it is possible to systematically and rigorously collect, formalize, and represent domain knowledge in a manner such that computers can reason upon those knowledge collections in a high throughput manner, thus replicating expert hypothesis generation processes in a way that is not constrained by innate human cognitive limitations and/or potential biases. Such a conclusion "opens the door"

for an exploration of ensuing in silico hypothesis discovery methods, as will be introduced in Sect. 8.3. Additionally, Payne et al. [29] provide a more comprehensive review of the theories, frameworks, and methods that make up the biomedical KE domain.

8.3 Design and Use of Intelligent Agents for In Silico Hypothesis Generation

While there exist a broad variety of methods that can be used for the purposes of in silico hypothesis discovery, spanning a spectrum from machine learning and data mining to iterative human-computer interaction in order to discovery high level patterns within complex data sets, for the purposes of this chapter, we will focus on a specific and exemplar type of methodology known as *knowledge discovery in databases* (KDD). This specific method has been selected in order to highlight the generalizable features of a much broad class of knowledge-based software and intelligent agents that can be used for in silico hypothesis generation. At a high level KDD is concerned with the utilization of intelligent agents, which are software applications that are designed to replicate human problem solving through the leverage of conceptual knowledge collections as an integral part of their architecture and function. In KDD, intelligent agents are used specifically to derive knowledge from the contents of databases, including database metadata. The use of domain-specific conceptual knowledge collections, such as ontologies, is central to the KDD induction process since commonly used database modeling approaches do not incorporate semantic knowledge corresponding to the database contents. This overall approach is the basis for a specific KDD methodology known as *constructive induction* (CI). In CI, data elements defined by a database schema are mapped to concepts defined by one or more ontologies or equivalent conceptual knowledge collections. Subsequently, the relationships included in the mapped ontologies are used to induce semantically meaningful relationships between the mapped data elements. The induction process generates what are known as "facts" concerning the contents of the database, which are defined in terms of data elements and semantic relationships that significantly link those elements together (Fig. 8.6).

These "facts" (which are a type of conceptual knowledge) can then be used to support higher level reasoning about the data defined by the targeted database schema. It is important to note that such "facts" can exploit the transitive closure principles associated with the graph-like representation of most ontologies, and therefore may include intermediate concepts that do not map to a database element but serve to create a semantically related concept triplet or high-order relationship that begins and terminates with concepts that do map to database elements.

The implementation of an intelligent agent that utilizes the preceding CI methodology often follows the multi-step process illustrated in Fig. 8.7 (which each phase numbered to reflect the following description) and outlined below:

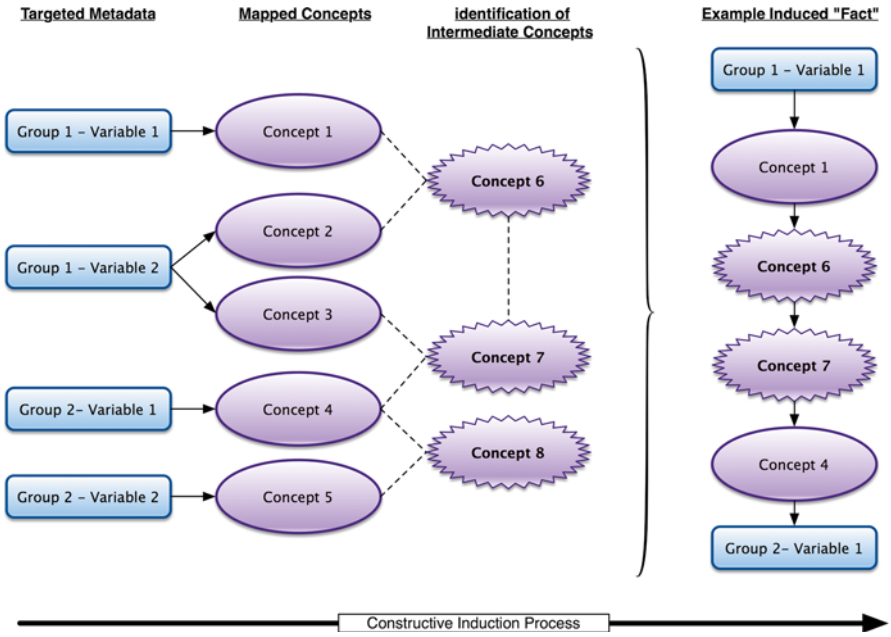

Fig. 8.6 Overview of constructive induction process whereby mapping between database elements as described via their metadata and corresponding ontology concepts are used to induce new "facts" concerning the contents of the database. In this general case, concepts 6–8, which is included in the ontology but does not map to the database construct, is used as an intermediate concept to define a concept triplet or higher order construct involving multiple intermediate entities that begins and terminates with data elements that map to concepts in the ontology construct

- **Phase 1 – Metadata to Conceptual Knowledge Entity Mapping**: In the first phase of implementing a CI-based agent, the metadata that serves to define a knowledge source of interest (e.g., a data dictionary or equivalent description of the contents of a data set or sets) must be mapped using either manual or automated processes to the entities that comprise one or more conceptual knowledge collections (e.g., syntactic or semantic matching of metadata definitions to entities in a terminology, ontology, or equivalent construct). This process usually results in one-to-many mappings, in which each metadata items corresponds to more than one conceptual knowledge entity. For example, if mapping a clinical data set with the specific variable corresponding to a "White Blood Cell Count", depending on the mapping approach being used and the intent of the KE initiative, that variable could be linked to multiple ontology-anchored concepts, such as the molecular entity "White Blood Cell", the laboratory procedure "White Blood Cell Count", as well as the clinical findings of "White Blood Cell Count Normal", "White Blood Cell Count High", and "White Blood Cell Low." This process generates a "knowledge map" that resolves individual variables of interest in the metadata being utilized to a corresponding set of atomic conceptual knowledge entities.

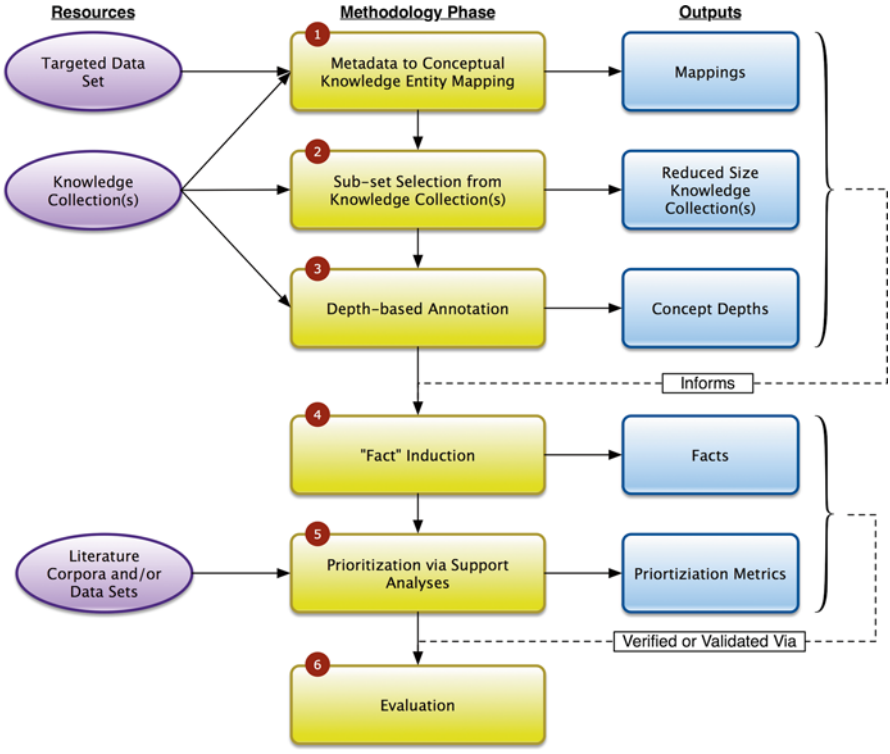

Fig. 8.7 Overview of major steps, resources, and outputs associated with the design and use of a CI-based agent

- **Phase 2 – Subset Selection from Knowledge Collection(s)**: Given that many conceptual knowledge collections contain thousands, if not hundreds of thousands, of distinct atomic entities and corresponding hierarchical or semantic relationships, such constructs can present computational challenges, such as the tractability, computational cost, or timeliness of computational tasks applied to such knowledge collections, which can be addressed through a process known as *search space reduction*. Effectively, once Phase 1 (Metadata to Conceptual Knowledge Entity Mapping) is complete, we can select a subset of those conceptual knowledge collections that directly correspond to: (1) the atomic elements mapped to the targeted metadata; (2) the hierarchical and/or semantic relationships that serve to link those atoms together; and (3) any additional atoms necessary to complete the linking paths identified via [2]. This allows refinement of the initial knowledge collections to one that is constrained to the problem-solving task at hand.
- **Phase 3 – Depth-based Annotation**: Once we have reduced the overall search space (Phase 2), an additional computational challenge must be addressed, concerned with the granularity of concepts being used for reasoning purposes. If we extend our prior example of "White Blood Cell Count" and its mapping

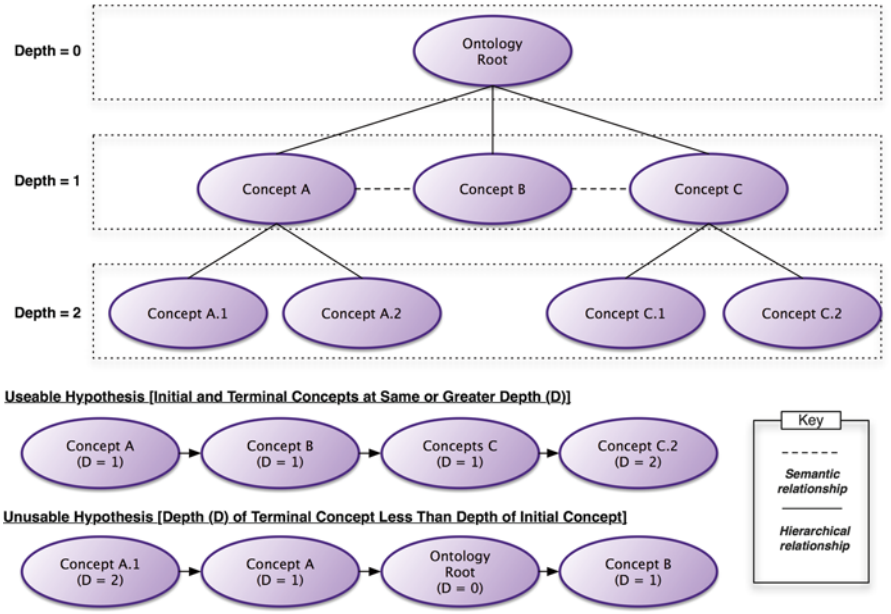

Fig. 8.8 Illustration of depth-based annotation and its implications for the induction of useable vs. unusable (e.g., overly general) "facts" or hypotheses

to an ontology-anchored concept of the laboratory procedure that has that same name, such a mapping could be used, when traversing the atomic units of information and relationships that comprise an ontology, to assert a relationship between "White Blood Cell Count" and the broad category of "Laboratory Procedures", which then in turn allows for the resolution of relationships with every other known laboratory procedures subsumed by that concept. This would be a factually accurate relationship to assert, but one that is functionally useless for hypothesis discovery, as it is overly broad and general. Why is this the case? Simply put, the concepts of "White Blood Cell Count" and "Laboratory Procedure" are not of an equivalent level of granularity (e.g., the former is much more specific than the latter). One approach that can serve as a surrogate for concept granularity in the source ontologies employed by a CI-based agent is the relative depth from the ontology root of those concepts (Fig. 8.8). Using such measurements, we can then constrain "fact induction" (Phase 4) to include only relationships between conceptual entities that exist at a similar or deeper depth from the ontology root and therefore can be expected to express useful and not overly generic hypothetical relationships. Doing so, however, requires us to first calculate the depth to the ontology root (or roots) for every conceptual entity selected in Phase 2 of this process, usually using the shortest such path as the preferred measurement when there exist more than one path from the concept to the root of the source ontology or equivalent conceptual knowledge construct.

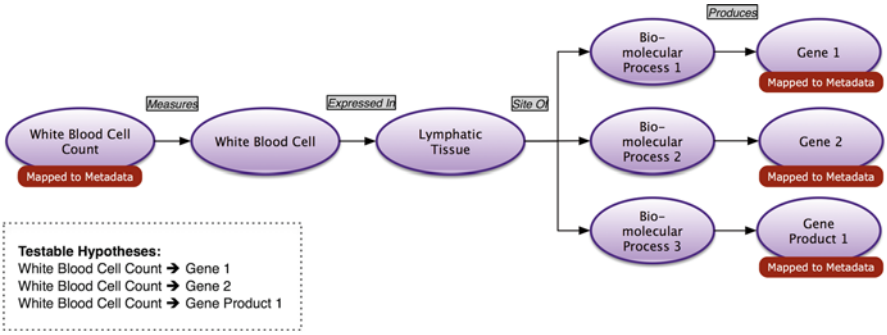

Fig. 8.9 Example of "fact" induction for prototypical example, in this case, creating testable hypotheses linking the initial and terminal concepts via multiple intermediate concepts and relationships (please note, this example assumes satisfaction of the depth based granularity controls associated with Phase 3 of the overall CI process)

- **Phase 4 – "Fact" Induction**: Once we have completed Phase 1–3, we can begin the "fact" induction process. In this phase we begin with a collection of variables contained within the target metadata of interest. For example, we could select all of the clinical measurements available that might serve to characterize how a patient would response to a therapy (such as laboratory findings or disease-specific performance or functional status indicators). Now, beginning with those variables, we can select a second set of variables that might serve as biomarkers of interest for predicting such treatment outcome, for example, indicators of genomic expression. Then, using the conceptual knowledge collection(s) that are mapped and sub-selected due to their connections to such variables, we can begin to explore the graph like representation of that knowledge to identify pathways that may link together variables in those two respective target "sets", being mindful of the granularity controls introduced in Phase 3. It is important to note in this process that such pathways are often "higher order" and can include multiple "intermediate" concepts and relationships that serve to link together an initial and terminal concept. For example, using our favorite case of "White Blood Cell Count", we might find that it is linked to the molecular entity "White Blood Cell" via a relationship labeled as "measures", and that "White Blood Cell" in turn has a relationship labeled as "expressed in" that connects it to the entity "Lymphatic Tissue." Subsequently "Lymphatic Tissue" could be linked via multiple "site of" relationships to a variety of bio-molecular processes that in turn may have relationships to certain genes or gene products that serve to measure the function or outcomes of those processes. Thus, we can then assert a "fact" that may infer a testable hypothesis linking our initial and terminal concepts and that could be tested using information contained in the source data sets(s) characterized by the metadata first identified in Phase 1 of this process (as illustrated in Fig. 8.9).
- **Phase 5 – Prioritization via Support Analyses**: In this nearly final step, it is often necessary to prioritize the hypotheses (or "facts") generated in Phase 4,

using some sort of quantifiable metric. This is necessary as CI-based agents can often generate thousands of hypotheses when reasoning over even a hundred or more initial and terminal variables. It is unlikely that human beings will take the time and expense (or have the energy and focus) to review and test all possible hypotheses. In response to this need, we often go back to the source data or alternatively, look at published literature and the knowledge that can be extracted from that literature (for example, the statistical distributions or co-occurrence of two variables of interest in the data or literature respectively) to calculate a support metric. Such support metrics tell us how common or uncommon those data or concepts are, and can be used to judge either the likelihood of the hypothesis being testable and/or novel. Then, depending on our use case, we can apply such metrics to prioritize or rank hypotheses for exploration and testing.

- **Phase 6 – Evaluation**: Finally (and perhaps most importantly), we must evaluate the output of CI-based agents using a variety of verification and validation methodologies. Such evaluations must incorporate multiple dimensions, include the factual accuracy or validity of system output, its likelihood in terms of informing novel hypotheses, and its overall utility as judged by the targeted end users. Further details on specific approaches to addressing this particular need are provided in Sect. 8.4.

8.4 Evaluating the Output of In Silico Hypothesis Generation Tools and Methods

The verification and validation of conceptual knowledge collections and the results of intelligent agents that leverage such knowledge to reason over data sets is ideally approached as an iterative and multi-method evaluation approach. First and foremost, when designing and applying such evaluation plans, it is very important to recognize and understand what types of process or outcomes measures are being targeted. Attaining such an understanding, in the context of intelligent agent design, requires us to differentiate between verification and validation. To summarize the definitions provided earlier, *verification* is the evaluation of whether an intelligent agent meets the perceived requirements of end-users, and *validation* is the evaluation of whether that same agent meets the realized (i.e., "real-world") requirements of the end-users. The only difference between these techniques is that during verification, results are compared to initial design requirements, whereas during validation the results are compared to the requirements for the system that are realized after its implementation.

Examples of verification and validation criteria include the degree of interrelatedness of the relationships discovered by the intelligent agent, the logical consistency of those relationships, and multiple-source or expert agreement with the results generated therein. Often, the degree of interrelatedness between relationships generated by an intelligent agent for hypothesis discovery purposes is used as a measure of its "quality", with such "quality" being defined by the degree to

		Nomenclature	
		Same	**Different**
Distinctions	**Same**	**Consensus** *Experts use the same nomenclature and distinctions to describe a conceptual entity.*	**Correspondance** *Experts use different nomenclature but the same distinctions to describe a conceptual entity.*
	Different	**Conflict** *Experts use the same nomenclature but different distinctions to describe a conceptual entity.*	**Contrast** *Experts use different nomenclature and distinctions to describe a conceptual entity.*

Table 8.1 Differentiation of types of agreement in multi-expert KA studies. In this model, the use of the "same" nomenclature or distinctions refers to the sources or experts using semantically similar or compatible means of describing or classifying concepts in a domain. Similarly, the use of "different" nomenclature or distinctions refers to the sources or experts using semantically dissimilar or incompatible means of describing or classifying concepts in a domain

which possible relationships between entities are enumerated or otherwise defined within the underlying knowledge collections. The logical, or axiomatic consistency of the relationships that comprise a hypothesis is often used as a measure of the accuracy of the output of the agent, again as defined by the correspondence of axioms that may be derived from the source knowledge collection(s) with the hierarchical and semantic assertions that make up such conceptual knowledge. Finally, multiple-source or expert agreement is most commonly used to validate the utility or impact of the output of the intelligent agent in "real world" application-oriented scenarios. This later set of measures is a critical criterion when attempting to measure the likely utility or impact of results generated by an intelligent agent. Unfortunately, there is not a single approach for measuring multiple-source, or expert agreement – since most evaluation methods corresponding to this type of metric involve the engagement of multiple (human) subject matter experts (SMEs). Instead, metrics must be chosen based upon variables such as data type as well as the number and types of knowledge sources being used. Most importantly, such analyses must be formulated in a manner consistent with the relative importance of four different types of agreement: (1) consensus; (2) correspondence; (3) conflict; and (4) contrast. Definitions of each of these types of agreement are provided in Table 8.1. A detailed discussion of the techniques that may be applied to measure agreement can be found in the reviews provided by Hripcsak et al. [30, 31].

At the highest level, the specific methods that can be used to satisfy the types of evaluation measures introduced above can be organized into a taxonomy consisting of the following major categories: heuristic, quantitative, information theoretic, graph theoretic and logical (Fig. 8.10). Brief descriptions of the techniques included in each category are provided below:

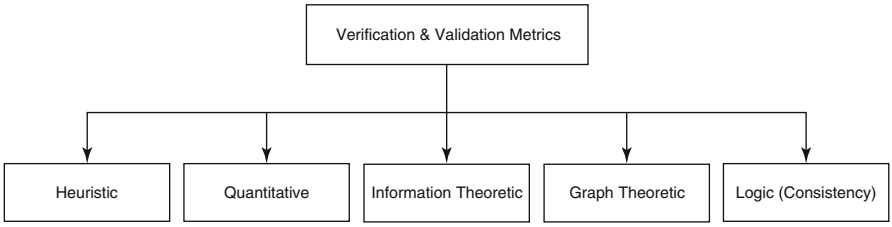

Fig. 8.10 Taxonomy of verification and validation metrics for the results generated by intelligent agents that leverage conceptual knowledge collections

8.4.1 Heuristic Methods

Heuristic metrics are probably the most common approach to verifying or validating the output of intelligent agents such as in silico hypothesis discovery tools. In this case, we use the term heuristic to refer to "rules of thumb" or more formally, rules that are informed by the expertise or commonly held knowledge of human SMEs. The advantages of using heuristics are the ability to incorporate domain-specific knowledge or conventions, and their simplicity (i.e., knowledge engineers or experts manually review the knowledge collection to determine if the contents are consistent with the heuristics). However, since such measures are difficult to automate or scale to larger data sets, such heuristic techniques are limited in their tractability when applied to "big data" contexts. Furthermore, heuristically comparing "quality" across multiple hypotheses or underlying knowledge collections is difficult, as a result of the relative and qualitative nature of the evaluation. Specific heuristic criteria for verifying or validating the output of intelligent agents have previously been proposed by Gruber [32] and include the following factors:

- Clarity
- Coherence
- Extendibility
- Minimal encoding bias
- Minimal deviation from ontological commitment, where ontological commitment refers to the situation were all observable actions of a knowledge-based system utilizing the given ontology are consistent with the relationships and definitions contained within that ontology.

8.4.2 Quantitative Methods

Quantitative methods of evaluating the results generated by intelligent agents are best suited for measuring both multi-source agreement and the degree of interrelatedness of ensuing hypotheses. Such measures can include simple statistics such as the precision, accuracy and chance-corrected agreement of the multiple sources

used during reasoning processes [31–36]. Using frequency-based measures (e.g., measuring the frequency with which a given entity is related to other entities within the knowledge collection) in addition to simple statistics can allow for the assessment of the degree of interrelatedness of a set of multiple hypotheses [37].

8.4.3 Information Theoretic Methods

Information theoretic methods are most commonly applied to measure multi-source agreement in an aggregate collection of multiple hypotheses. The use of information theory to evaluate the agreement between multiple sources is based on the argument that if such agreement exists, it will be manifested as repetitive patterns within the resulting information constructs. To utilize this verification and validation approach, the relationships between units of knowledge that make up each constituent hypothesis must be represented as a numerical matrix, where each cell contains a numerical indication of the strength of the relationship between the two units of knowledge identified by the corresponding row and column indices. Given such a matrix, repeating patterns can be quantified based on their effect on information content or complexity. Matrix complexity is determined by calculating the number of repeating patterns within the matrix less the contribution of the overall environment within which the matrix is constructed. The probability of each repeating pattern detected in the actual matrix occurring randomly or as a result of the environmental contribution can be computed by generating multiple random matrices. As matrix complexity decreases, the degree of multi-source agreement increases [35]. This type of evaluation method is summarized in Fig. 8.11, and further detail can be found in the work reported on by Kudikyala et al. [35].

8.4.4 Graph Theoretic Methods

Graph theoretic methods are based on the ability to represent knowledge-based formulations, such as the output of intelligent agents, as graph constructs, where individual units of information or knowledge are represented as nodes, and the relationships between these units as arcs. Such graph representation of knowledge collections has been described in a number of areas, including ontologies [32, 38], taxonomies [39, 40], controlled terminologies [41] and semantic networks [40, 42]. Given a particular graph representation of a hypothesis or set of hypotheses, the degree of interrelatedness of those knowledge-based products can be assessed using a group of graph-theoretic techniques known as class cohesion measures. Such metrics are used to assess the degree of cohesion, a property representative of connectivity within a graph. Specific class cohesion measurement algorithms include the Lack of Cohesion of Methods (LCOM), Configurational-Bias Monte Carlo (CBMC), Improved Configurational-Bias Monte Carlo (ICBMC) and Geometrical

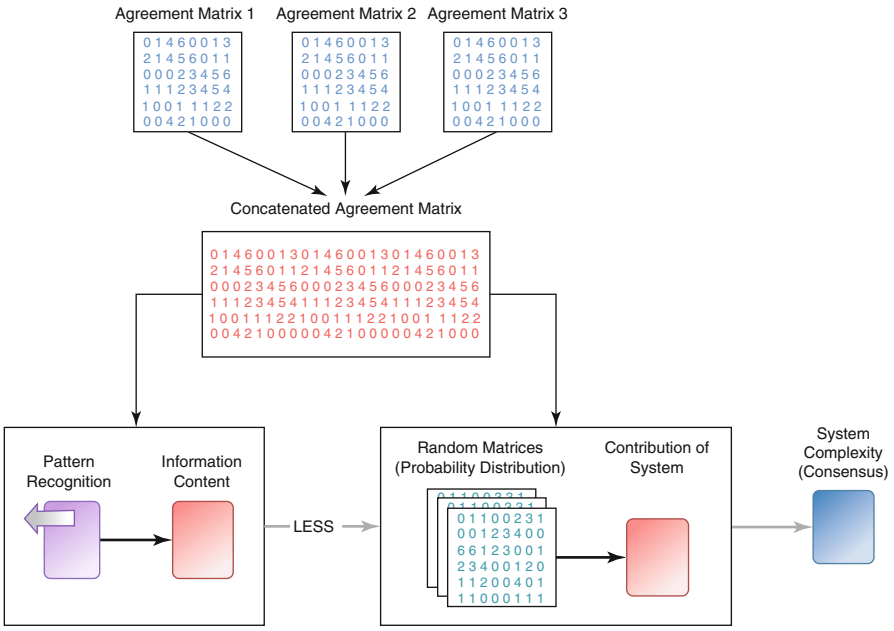

Fig. 8.11 Overview of information theoretic evaluation method for determining the degree of multi-source or expert agreement within a knowledge collection or system

Design Rule Checking (DRC) algorithms [43]. All of these algorithms use some combination of the number of and distance between interrelated vertices within the graph as the basis for determining cohesion. Most cohesive graphs generally possess more interrelated vertices with relatively short edges between them. However, it is important to note that a precise definition of what constitutes "cohesion" in a graph is not necessarily universally agreed upon. Due to this lack of agreement, class cohesion algorithms tend to utilize different measures for cohesion. The applicability of these metrics varies depending on the specific evaluation context. As a result, the selection of an appropriate cohesion measure is highly dependent on the specific nature of the data set and application scenario being evaluated. Further details concerning the theoretical basis and application of graph theory-based cohesion measures can be found in the review provided by Zhou et al. [43].

8.4.5 Logical Methods

The application of logic-based verification and validation techniques for the output of intelligent agents focuses on the detection of axiomatic consistency. These techniques require the extraction of logical axioms from the knowledge collection that has informed such in silico hypothesis discovery operations. Once axioms have been extracted, they are then applied within the targeted domain in order to evaluate

their consistency and performance. In addition, logical methods can be utilized to examine axioms and assess the existence of unnecessary or redundant relationships within the knowledge collection. One of the most common approaches to implementing this type of evaluation is the representation of the individual hypothesis generated by the agent as formal ontological constructs within the Protegé knowledge editor [44]. Once such hypotheses have been represented in Protégé, logical axioms can be extracted and evaluated using the Protegé Axiom Language (PAL) extension [45]. An example of this method can be found in the formal evaluation of the logical consistency of the Gene Ontology (GO) [46] reported by Yeh et al. [45].

8.4.6 Hybrid Methods

As described earlier, hybrid methods for verifying or validating knowledge collections involve the use of techniques belonging to two or more of the classes of measures as described above. An example of such a hybrid method is the novel computational simulation approach to validating the results of multi-expert categorical sorting studies as proposed by Payne and Starren [47]. This approach measures multi-source agreement using a combination of quantitative and graph theoretic methods. Another example of a hybrid technique is the use of hypothesis discovery methods, such as hierarchical clustering [48] to determine the degree of interrelatedness of a knowledge collection. Such evaluative methods combine statistical, heuristic and graph theoretic techniques.

8.5 Implications for Stakeholders

It can be seen that each of the different stakeholders described in Chap. 1 benefits realizing the vision of a Translational Informatics model that enables and facilitates knowledge-driven healthcare. With specific regard to the concepts associated with in silico hypothesis discovery, these benefits are multi-fold, and largely focus upon the accelerated pace and ease with which new diagnostic and therapeutic discoveries can be generated from existing or new data sets. Specific benefits at all of the levels introduced in Chap. 1 include:

8.5.1 Evidence and Policy Generators

- **Investments in the creation of large-scale and multi-dimensional data sets can exhibit much higher returns on investment** owing to the ability to generate a larger number of testable and potentially clinically actionable hypotheses from those resources;

- **Novel evidence and/or policy frameworks can be inferred based upon previously undiscovered patterns or motifs in historical data sets**, thus allowing such knowledge or decision making to be informed by the best possible information.

8.5.2 Providers and Healthcare Organizations

- Providers are able to **engage in the delivery of evidence-based and precision medicine informed by a full spectrum of scientific knowledge** that has been formulated by identifying and testing large numbers of hypotheses against all available data types and resources
- **Healthcare organizations can leverage their investments in EHR technologies and bio-molecular instrumentation** so as to rapidly learn from all patient-centered data being created during the course of normal clinical operations; that is, achieving the vision of a "learning healthcare system" wherein every patient encounter is an opportunity to both create new knowledge and improve care for that patient, their family, and their community.

8.5.3 Patients and Their Communities

- **Patients are able to be part of the "learning healthcare system"** such that they both become an integral component of research processes and benefit from the knowledge generated therein
- **Interested community members can begin to identify novel or interesting associations between disparate data that spans healthcare providers and the world-at-large**, thus becoming part of the research enterprise. For example, community members could use in silico hypothesis discovery tools to identify relationships between healthcare outcomes and socio-demographic factors that could inform advocacy and/or community development activities intended to promote wellness.

8.6 Conclusions

As has been discussed in a variety of ways throughout this book, the ongoing growth and increasing complexity of biomedical data presents a wealth of challenges and opportunities relative to informing a Translational Informatics vision for knowledge-drive healthcare. In this chapter, we have discussed a specific aspect of those challenges and opportunities, concerned with the disconnect between the volume of data being generated in numerous settings and the current state-of-the-art

in terms of hypothesis formulation and testing relative to such resources, which remains extremely basic. As has been illustrated, most if not all hypotheses that are evaluated in the modern scientific setting are generated in a low-throughput manner based upon the intuition or belief systems of an individual or team of investigators. Despite historical precedence for such approaches, they are discordant with the modern, high-throughput data types we regularly encounter, and that are being generated by EHRs, PHRs, sensor technologies and bio-molecular instrumentation (to name a few of innumerable examples). In response to this challenge, we can look to a set of core concepts that underlie alternative and high-throughput approaches that can lead to in silico hypothesis discovery paradigms. These types of methods employ domain-specific conceptual knowledge collections, such as ontologies or knowledge that can be extracted from the domain literature using machine learning or natural language processing methods, in order to reason upon and generate hypothesis corresponding to a data set or data sets in an extremely high throughput manner, usually realized via the implementation of knowledge-based and intelligent software agents. While these types of in silico hypothesis discovery methods remain very early in their development, they also hold great promise in terms of accelerating the pace, breadth and depth of scientific discovery in the "big data" era, and thus represent a critical dimension of the vision for Translational Informatics.

Discussion Points

- What are the major barriers to the generation and testing of hypotheses in large-scale and/or heterogeneous data sets?
- What differentiates procedural, strategic, and conceptual knowledge? How are these knowledge types related across a continuum of operationalization?
- What role can conceptual knowledge collections play in overcoming the preceding barriers?
- As an example of an in silico hypothesis discovery method, what considerations must be addressed when employing Constructive Induction (CI) relative to concept granularity and/or the evaluation of ensuing hypotheses?
- When evaluating the output of knowledge-based intelligent agents used for in silico hypothesis generation, what is the fundamental difference between the verification versus validation of such constructs?

References

1. Zerhouni EA. US biomedical research: basic, translational, and clinical sciences. JAMA. 2005;294(11):1352–8. PubMed PMID: 16174693.
2. Sung NS, Crowley Jr WF, Genel M, Salber P, Sandy L, Sherwood LM, et al. Central challenges facing the national clinical research enterprise. JAMA. 2003;289(10):1278–87. PubMed PMID: 12633190.
3. 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. PubMed PMID: 15974245.

4. Liu H, Motoda H. Feature extraction, construction and selection: a data mining perspective. Norwell: Kluwer Academic Publishers; 1998.
5. Payne PR, Borlawsky TB, Kwok A, Dhaval R, Greaves AW, editors. Ontology-anchored approaches to conceptual knowledge discovery in a multi-dimensional research data repository. In: 2008 AMIA Translational Bioinformatics Summit. San Francisco: American Medical Informatics Association; 2008.
6. Payne PR, Borlawsky TB, Kwok A, Greaves AW. Supporting the design of translational clinical studies through the generation and verification of conceptual knowledge-anchored hypotheses. AMIA Annu Symp Proc. 2008:566–70. PubMed PMID: 18998958. Pubmed Central PMCID: 2656058. Epub 2008/11/13. eng.
7. Payne PR, Borlawsky TB, Rice R, Embi PJ. Evaluating the impact of conceptual knowledge engineering on the design and usability of a clinical and translational science collaboration portal. AMIA Clinical Research Informatics Summit. San Francisco: American Medical Informatics Association; 2010.
8. Payne PR, Embi PJ, Johnson SB, Mendonca EA, Starren JB. Improving the usability of clinical trial participant tracking tools using knowledge-anchored design methodologies. Appl Clin Inform. 2010;1(2).
9. Payne PR, Huang K, Keen-Circle K, Kundu A, Zhang K, Borlawsky TB. Multi-dimensional discovery of biomarker and phenotype complexes. AMIA Translational Bioinformatics Summit. San Francisco: American Medical Informatics Association; 2010.
10. Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. Stat Appl Genet Mol Biol. 2005;4:Article 17. PubMed PMID: 16646834. Epub 2006/05/02. eng.
11. Zhang J, Ding L, Keen-Circle K, Borlawsky TB, Xiang Y, Ozer G, et al. Predicting biomarkers for chronic lymphocytic leukemia using gene co-expression network analyses for ZAP70. AMIA Translational Bioinformatics Summit. San Francisco: American Medical Informatics Association; 2010.
12. Zhang J, Xiang Y, Jin R, Huang K. Using frequent co-expression network to identify gene clusters for breast cancer prognosis. Proc Int Joint Conf Bioinforma Syst Biol Intell Comput. 2009;428–34.
13. Goldstein T. Dawn of modern science: from the ancient Greeks to the Renaissance. New York: Da Capo Press; 1995.
14. Chung TK, Kukafka R, Johnson SB. Reengineering clinical research with informatics. J Investig Med. 2006;54(6):327–33. PubMed PMID: 17134616.
15. 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. PubMed PMID: 19261934. Epub 2009/03/06.
16. Zerhouni EA. Translational and clinical science – time for a new vision. N Engl J Med. 2005;353(15):1621–3. PubMed PMID: 16221788.
17. Glaser R. Education and thinking: the role of knowledge. Am Psychol. 1984;39(2):93–104.
18. Hiebert J. Procedural and conceptual knowledge: the case of mathematics. London: Lawrence Erlbaum Associates; 1986.
19. McCormick R. Conceptual and procedural knowledge. Int J Technol Des Educ. 1997;7:141–59.
20. Scribner S. Knowledge at work. Anthropol Educ Q. 1985;16(3):199–206.
21. Barsalow LW, Simmons WK, Barbey AK, Wilson CD. Grounding conceptual knowledge in modality-specific systems. Trends Cogn Sci. 2003;7(2):84–91.
22. Borlawsky T, Li J, Jalan S, Stern E, Williams R, Lussier YA. Partitioning knowledge bases between advanced notification and clinical decision support systems. AMIA Annu Symp Proc. 2005:901. PubMed PMID: 16779188.
23. Preece A. Evaluating verification and validation methods in knowledge engineering. In industrial knowledge management. 2001:91–104. Springer London
24. Newell A, Simon HA. Computer science as empirical inquiry: symbols and search. In: Haugeland J, editor. Mind Design. Cambridge: MIT Press/Bradfor Books; 1981. p. 35–66.
25. Compton P, Jansen R. A philosophical basis for knowledge acquisition. Knowl Acquis. 1990;2(3):241–57.

26. Newell A, Simon HA, editors. Computer science as empirical inquiry: symbols and search. ACM annual conference. Minneapolis; 1975.
27. Gaines BR, Shaw MLG. Knowledge acquisition tools based on personal construct psychology 1993 [Cited 2005 8/23/2005]. Available from: http://www.repgrid.com/reports/KBS/KER/.
28. Kelly GA. The psychology of personal constructs. 1st ed. New York: Norton; 1955. 2 v. (1218) p.
29. Payne PR, Mendonca EA, Johnson SB, Starren JB. Conceptual knowledge acquisition in biomedicine: a methodological review. J Biomed Inform. 2007;40(5):582–602. PubMed PMID: 17482521.
30. Hripcsak G, Heitjan DF. Measuring agreement in medical informatics reliability studies. J Biomed Inform. 2002;35(2):99–110. PubMed PMID: 12474424.
31. Hripcsak G, Wilcox A. Reference standards, judges, and comparison subjects: roles for experts in evaluating system performance. J Am Med Inform Assoc. 2002;9(1):1–15. PubMed PMID: 11751799.
32. Gruber TR. Toward principles for the design of ontologies used for knowledge sharing. Int J human-computer studies. 1995;43(5):907–28.
33. Hripcsak G, Rothschild AS. Agreement, the f-measure, and reliability in information retrieval. J Am Med Inform Assoc. 2005;12(3):296–8. PubMed PMID: 15684123.
34. Humphreys BL, McCray AT, Cheh ML. Evaluating the coverage of controlled health data terminologies: report on the results of the NLM/AHCPR large scale vocabulary test. J Am Med Inform Assoc. 1997;4:484–500. PubMed PMID: 9391936.
35. Kudikyala UK, Allen EB, Vaughn RB, editors. Measuring consensus during verification and validation of requirements. Proceedings of the tenth IEEE International Software Metrics symposium (METRICS 2004). Chicago; 2004.
36. Morgan MS, Wm. Benjamin Martz, Jr. Group Consensus: do we know it when we see it? In: Proceedings of the Proceedings of the 37th annual Hawaii international conference on System Sciences (HICSS'04) – Track 1 – vol. 1: IEEE Computer Society. Waikoloa: Hawaii; 2004.
37. Brachman RJ, McGuinness DL. Knowledge representation, connectionism and conceptual retrieval. In: Proceedings of the 11th annual international ACM SIGIR conference on Research and Development in Information Retrieval. Grenoble: ACM Press; 1988.
38. Ian N, Adam P. Towards a standard upper ontology. In: Proceedings of the international conference on Formal Ontology in Information Systems – vol. 2001. Ogunquit: ACM Press; 2001.
39. Alan LR, Chris W, Jeremy R, Angus R. Untangling taxonomies and relationships: personal and practical problems in loosely coupled development of large ontologies. In: Proceedings of the international conference on Knowledge Capture. Victoria: ACM Press; 2001.
40. Burgun A, Bodenreider O. Aspects of the taxonomic relation in the biomedical domain. Ogunquit: ACM Press; 2001. p. 222–33.
41. Cimino JJ. From data to knowledge through concept-oriented terminologies: experience with the Medical Entities Dictionary. J Am Med Inform Assoc. 2000;7(3):288–97. PubMed PMID: 10833166.
42. Griffith RL. Three principles of representation for semantic networks. ACM Trans Database Syst. 1982;7(3):417–42.
43. Zhou Y, Lu J, Xu HB. A comparative study of graph theory-based class cohesion measures. SIGSOFT Softw Eng Notes. 2004;29(2):13.
44. Noy NF, Crubezy M, Fergerson RW, Knublauch H, Tu SW, Vendetti J, et al. Protege-2000: an open-source ontology-development and knowledge-acquisition environment. AMIA Annu Symp Proc. 2003:953. PubMed PMID: 14728458.
45. Yeh I, Karp PD, Noy NF, Altman RB. Knowledge acquisition, consistency checking and concurrency control for Gene Ontology (GO). Bioinformatics. 2003;19(2):241–8. PubMed PMID: 12538245.
46. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. Gene Ontol Cosortium Nat Genet. 2000;25(1):25–9. PubMed PMID: 10802651.

47. Payne PR, Starren JB. Quantifying visual similarity in clinical iconic graphics. J Am Med Inform Assoc. 2005;12(3):338–45. PubMed PMID: 15684136.
48. Everitt B, Landau S, Leese M. Cluster analysis. 4th ed. New York: Oxford University Press; 2001. p. 237.

Additional Reading

Everitt B, Landau S, Leese M. Cluster analysis. 1st ed. New York: Oxford University Press; 2001. 2 v. (1218) p.
Glaser R. Education and thinking: the role of knowledge. Am Psychol. 1984;39(2):93–104.
Goldstein T. Dawn of modern science: from the ancient Greeks to the Renaissance. New York: Da Capo Press; 1995.
Hiebert J. Procedural and conceptual knowledge: the case of mathematics. London: Lawrence Erlbaum Associates; 1986.
Hripcsak G, Heitjan DF. Measuring agreement in medical informatics reliability studies. J Biomed Inform. 2002;35(2):99–110.
Hripcsak G, Wilcox A. Reference standards, judges, and comparison subjects: roles for experts in evaluating system performance. J Am Med Inform Assoc. 2002;9(1):1–15.
Liu H, Motoda H. Feature extraction, construction and selection: a data mining perspective. Norwell, MA: Kluwer Academic Publishers; 1998.
Newell A, Simon HA. Computer science as empirical inquiry: symbols and search. In: Haugeland J, editor. Mind design, vol. 1. Cambridge: MIT Press/Bradfor Books; 1981.
Payne PR, Mendonca EA, Johnson SB, Starren JB. Conceptual knowledge acquisition in biomedicine: a methodological review. J Biomed Inform. 2007;40(5):582–602.

Chapter 9
Patient Engagement and Consumerism

Adam B. Wilcox

By the End of this Chapter, Readers Should Be Able to:

- To understand the importance of the patient role in translational informatics
- To identify examples of how patients can be engaged
- To understand risks to patients and how patients can be protected
- To recognize trends and factors in patient engagement
- To identify potential risks to translational informatics from patient-driven activities

9.1 Introduction

Throughout the previous chapters, there have been many components that are directly related to patients. In reality, all of translational informatics is related to patients, since the ultimate goal is to improve the health and healthcare of patients. But some parts of the field are more directly related than others. De-identification algorithms that allow a large set of medical records to have protected health information removed, so that the dataset can be transferred to other scientists who are applying data mining algorithms that can discover associations between groups is ultimately related to patients, since those associations may eventually lead to new treatments. But it feels comparatively more distant to patients than a website that helps them directly to enroll in clinical trials related to their conditions. Other initiatives, such as mining the bibliome (Chap. 5) and Big Data (Chap. 7) feel even more distant. Some translational informatics projects are fully dependent on patients and

A.B. Wilcox, PhD
Department of Medical Informatics, Intermountain Healthcare,
Murray, UT, USA
e-mail: adam.wilcox@imail.org

P.R.O. Payne, P.J. Embi (eds.), *Translational Informatics: Realizing the Promise of Knowledge-Driven Healthcare*, Health Informatics,
DOI 10.1007/978-1-4471-4646-9_9,

consumers being engaged in the project, others just require their consent, and some just require their data.

In this chapter, we focus on the more patient-connected parts of translational informatics and clarify connections to patients that exist. Previous chapters have looked at a specific scientific paradigm. Here, we will instead focus directly on the patient. We begin by describing from the patient perspective the various ways one can be involved in research and affected by translational informatics. We then focus on areas from that perspective dependent or affected on patient engagement, and describe methods that may be used to make the engagement, and therefore the research, more successful. Throughout, we give examples of projects that have improved research and translational informatics by improving patient engagement.

9.2 Patient Participation in Research

Patients participate in the research process in various ways. Among the most obvious is by participating as subjects in research studies. In a prospective study, individuals are followed over a period of time, with data being collected throughout the study for measurements used in statistical tests.

9.2.1 Recruitment

Participation in a prospective trial begins with recruitment where potential subjects are contacted and requested to participate in a trial. Recruitment happens in multiple ways. With some studies, patients may hear about a study through advertisements and then contact the study team. In other studies, patients are contacted directly by a clinician providing care or someone representing the care provider. These represent the most common forms of recruitment.

Ideally, studies with generic inclusion criteria understood by patients or needing to study a broad population would use advertising as a recruitment strategy, while studies with more complicated criteria or with rare conditions that are better understood by clinicians would have providers recruit patients directly. In reality, neither works well. Two challenges to recruitment are complicated or restrictive entry criteria, and clinician participation in recruitment [1]. Unfortunately, these two issues work together – restrictive entry criteria often need clinical judgment or information to identify potential subjects, but identifying this information by clinicians takes time, which is already a scarce resource. The result is that most clinical trials experience recruitment problems [1, 2], including delays, increased costs, or failures of the study [3, 4].

Recently, some initiatives have improved recruitment success by leveraging patient engagement in the recruitment process. Rather than patients being contacted by researchers through advertising or clinicians, patients can proactively indicate

their desire to participate in clinical trials, and are then contacted directly. Other initiatives have proactively outreached to patients to participate generally in trials, creating registries of potential subjects. And still others allow patients to search for potential clinical trials based on their clinical information. For example, TrialX.com is a consumer website where patients can specifically search for clinical trials by entering health information and find matching entry criteria [5]. Notification information can then be sent to investigators if the individual matches the criteria, and wishes to volunteer as a subject. These initiatives are areas where translational informatics has facilitated patient engagement to improve recruitment.

A future direction for clinical trial recruitment and informatics is to facilitate the use of data in personal health records (PHRs) to facilitate patient-driven recruitment. Use of clinical information by patients, either manually entering the data into a search site like TrialX.com, or interpreting the link between data in a PHR and trial entry criteria, is a barrier that can reduce patient engagement such that they may not participate. The initial successes and proliferation of sites like TrialX make it clear that many patients are willing and eager to participate in trials. Recruitment failure is therefore often due to barriers patients face in the process, that makes them either not engage initially or lose interest before actual recruitment. Anything that reduces the burden of patients to enroll in trials will improve the success of patient engagement. By proactively using the information in PHRs, patients could automatically search for qualifying clinical trials [6]. TrialX.com allows for importing data from PHRs, but ideally the trial search function would be incorporated in the PHR.

9.2.2 Consent

After patients are recruited to participate in a study, they must individually give consent to participate, and agree to have data collected and used. In a natural history study, only information is collected – no intervention is given to any subjects as part of the research study. Some of the most famous research studies in medicine have been natural history studies. The Framingham Heart Study, which began with a cohort of over 5,000 individuals in Framingham, MA, has been among the most influential studies in medicine. Much of what we now know as important for treatment and management of cardiovascular health began with this study. The subjects were followed for many years, with multiple data collection points throughout the study. Over time, scientists were able to see correlations between data they had collected on subjects and their eventual health. While natural history studies do not change the treatments given to subjects, they are still intrusive. Data collection can be difficult, time-consuming, and, where it requires invasive medical tests, even painful. If the information needed for a natural history study is already being collected as part of standard care, this data can be collected and used for the study instead, thus decreasing the discomfort for the subject and increasing the efficiency of the study. It can also increase the ease of participation, and lead to easier patient

recruitment. Some institutions have created research data warehouses based only on data collected as part of standard care, existing in the EHRs. These research initiatives have limited the invasiveness of research recruitment down to just consent to use the information. Examples include the many current translational informatics initiatives currently requesting for patients to "donate" their data [7–9].

Even when the data collection is minimized by using existing data from the EHR and not changing the intervention, there is still a burden to the patient for which consent is needed. This burden is a privacy and confidentiality risk. Health care assessments and therapies are personal and sensitive, and the record of those interactions should not be known outside those providing care, receiving care, or paying for care. The Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule and the Health Information Technology for Economic and Clinical (HITECH) Health Act specifically protect patients, by defining what data are protected, defining rules for the use for the protected health information (PHI), and including penalties for the misuse and loss of PHI. Institutional review boards (IRBs) also protect patients in natural history studies against loss of privacy. While protection of confidentiality and privacy does not require engagement of patients directly, it is by intent a very patient-centered component of research.

A method to reduce the privacy and confidentiality risk is to de-identify data by removing identifying information, such that there are not certain indications of the individual with the data. Though full anonymization of clinical data is not possible [10], removal of specific identifying information can reduce the risk of re-identification to reduce the risk to acceptable levels [11]. For some institutions, careful use of data de-identification has allowed the creation of very large research cohorts, that use both data from EHRs linked to biospecimens from discarded samples [12]. By reducing the risk to privacy and confidentiality, they have fully reduced the burden to patients needed to create a biobank.

9.2.3 Data Collection and Integration

Once a patient has consented for use of their EHR data for a natural history, secondary data use study, the patient engagement is basically done. For interventional studies or natural history studies that require primary data collection, patient participation begins at consent. Subjects are required to participate in either intervention or data collection activities multiple times throughout the study, for the initial data, for the final outcomes measurements, and for monitoring during the course of the study. This can also be a burden for the patients, requiring time and inconvenience. Ideally, data needed for research is the same as data needed for care, but often the research and care processes are separate, requiring either different testing schedules or different data elements. To reduce patient inconvenience, some researchers have considered integrating scheduling data for clinical care and research calendars [13]. The result would be that where appropriate, research visits and clinical care visits are coordinated, decreasing the patient burden for research participation.

A side effect of participation in an interventional clinical trial for a patient is that the patient is receiving an intervention that is not part of standard care, and may not be coordinated with other care received. For example, if a patient is recruited to a trial that is not directed by the patient's primary physician, other clinicians will be providing interventional care for that patient independent of their primary physician. Because the clinical trial information system is usually separate from the EHR, information about the treatment received may not be accessed as part of regular clinical care. For example, a patient may receive an intervention as part of a clinical trial that is not indicated in the patient record, and the treating physicians in an emergency department may not know about the intervention unless the patient informs them of the intervention. Integrating information from the research data system to the EHR to notify treating clinicians is thus important to reduce unanticipated interactions between care provided, and improve patient safety.

Patient engagement can directly influence research in the area of data collection, because a result of increased patient engagement in health and healthcare is greater data about patients. Consumer health and social media sites like PatientsLikeMe.com have emerged as more patients are using electronic tools for support. Studies of such electronic health data sources have identified the potential of using these data for research [14, 15]. A benefit of such sources is that participants collect data more frequently than other clinical sources. Extreme data collection by patients with such engagement initiatives as Quantified Self provide opportunities for analysis of a high level of self-data collection, and can also lead to innovations for advancements in increased data collection for the general population [16].

9.2.4 Providing Results

In most clinical trials, there is no patient engagement beyond recruitment, consent and data collection. Once the data are collected, contact with subjects ends, as the research moves to analysis and dissemination phases. Much of the disconnect is due to cost – research funds usually only cover what is needed for the study. The problem is that patients participating in a trial are among the most interested in the results of a study, but standard dissemination methods are (a) slow and (b) directed at the medical community rather than the patient. This can have a dampening effect on the enthusiasm of both subjects and a community for participation in clinical trials, where they are less willing to participate in other trials. It can also create distrust between the patient and research communities. One study of community engagement showed that less than a quarter of studies had meaningful engagement with the community, in either advising the direction of the research, facilitating recruitment, or having findings shared [17]. Unfortunately, the effects of non-engagement have been seen most strongly in minority and underserved populations, the same populations where more research is needed to address health disparities.

Community-based participatory research (CBPR) has emerged as an approach to improve engagement in research at the community level. CBPR is a form of

engagement driven by researchers to actively involve communities in both directing questions for research, and providing results [18]. It has been especially applied to populations where the levels of patient engagement in research is comparatively low, leading to disparities in research applicability that may be related to disparities in health and healthcare. To improve community engagement as a core component of translational research, specific elements of community engagement have been required for the national Clinical and Translational Science Awards (CTSAs) [19]. A critical component of translational informatics is to facilitate CBPR in clinical studies. An example is the WICER project mentioned in Chap. 4. WICER is an informatics infrastructure for comparative effectiveness research, focused on a mostly-immigrant, Hispanic population in New York City. A core component of WICER is the population survey, where data are collected directly from community members. Focus groups within the community were used to refine the questions. Studies using subjects from the WICER cohort have shown an increased level of engagement that other sources, and it has substantially improved the engagement of the community in research studies [20].

A current innovation of WICER is to provide research data back to subjects in a way that helps them understand the information, as a health promotion tool. Focusing on improved health understanding and cognition, researchers are investigating how to represent data in a simplified and non-ambiguous way, so that subjects can see their health relative to other subjects in the study. This innovation may be able to translate research engagement into tools supporting overall health engagement, creating an innovative method of health promotion.

9.3 Implications for Stakeholders

As was introduced in Chap. 2, a variety of stakeholders can and will benefit from the engagement of patients and their communities in the broad TI paradigm. Critical examples of these benefits stratified by stakeholder type include the following:

Evidence and Policy Generators

- Traditional approaches to research often focus solely upon data collected by practitioners and/or researchers in clinical care or other closely controlled settings. In contrast, significant amounts of data and activity that pertain to health and wellness occur or are generated outside of those **settings. By embracing the role of patients and their communities as part of the "evidence generation" team, the benefits of systems-level approaches to complex healthcare problems can be achieved**.
- In a similar manner, **policies that influence healthcare research and delivery are often generated based on data and evidence that does not incorporate patient or community input and related information resources**. As such, these policies can frequently fail due to unanticipated barriers to adoption or

implementation when faced with "real world" scenarios. By addressing such challenges at the outset, the likelihood of success for such policy related activities can be greatly enhanced.

Providers and Healthcare Organizations

- Providers are often forced to make clinical decision based on incomplete evidence and data, for example, not being able to adjust treatment plans based upon a rigorous understanding of dietary or activity patterns that occur outside of the care delivery environment. **By incorporating patients and communities into the data generation pipeline that supports/enables clinical decision making**, **more comprehensive and efficacious clinical decision-making is made possible**.
- Much as is the case at an individual level, healthcare delivery organizations are also increasingly concerned with managing populations to lower risk and improve outcomes, especially when they are financially "at risk" for the wellness of those individuals. **By creating a** "**data fabric**" **incorporating patients and their communities**, **the ability to comprehensive and predictively model such trends in an impactful manner becomes feasible**, **thus resulting in both financial and quality of care benefits to organizations and the patients they serve**.

Patients and Their Communities

- By providing patients with a "voice" in the knowledge-driven healthcare environment, **it is possible to catalyze a cultural change via which those individual become active participants in healthcare delivery**, **rather than passive consumers**. Ample evidence exists showing that such activist patients are more likely to experience positive healthcare outcomes and greater overall wellness.
- Finally, **by making communities part of the same healthcare data** "**dialogue**", **trends that may impact entire populations** (**either positively or negatively**) **can be surfaced and acted upon**, often through community-level advocacy efforts and at much lower costs with improved benefits when compared to acute or episodic care models.

9.4 Conclusion

We have described patient engagement around the process of a clinical trial, from recruitment, consent, data collection, and providing results. At each point, patient engagement can improve the specific research tasks, and improve translational research. We have also identified important innovations in each area that affect or are affected by patient engagement. As the nation moves to a vision of a learning health system, the importance of patient engagement, both as a facilitator of research

and a catalyst for translation, becomes increasingly important. Patient-initiated activities will be particularly interesting in how they can improve research and health of individuals.

Discussion Points

- How do the different stages of a clinical trial where patients may be more or less engaged? What barriers exist at each stage that would lead to patients not engaging?
- For what reasons may patients want or not want to participate in clinical trials. What types of information might affect these reasons?
- What are the possible biases that can occur as more patients become engaged in the research process? How could research studies mitigate against these biases?
- How can the results of a study best be presented to a patient who participates in the trial? How would this be different than results for patients who did not participate?

References

1. Prescott RJ, Counsell CE, Gillespie WJ, Grant AM, Russell IT, Kiauka S, et al. Factors that limit the quality, number and progress of randomised controlled trials. Health Technol Assess Winch Engl. 1999;3:1–143.
2. Drennan KB. Patient recruitment: the costly and growing bottleneck in drug development. Drug Discov Today. 2002;7:167–70.
3. Lovato LC, Hill K, Hertert S, Hunninghake DB, Probstfield JL. Recruitment for controlled clinical trials: literature summary and annotated bibliography. Control Clin Trials. 1997;18:328–52.
4. Peters-Lawrence MH, Bell MC, Hsu LL, Osunkwo I, Seaman P, Blackwood M, et al. Clinical trial implementation and recruitment: Lessons learned from the early closure of a randomized clinical trial. Contemp Clin Trials. 2012;33:291–7.
5. Patel CO, Garg V, Khan SA. What do patients search for when seeking clinical trial information online? AMIA Annu Symp Proc. 2010;2010:597–601.
6. Wilcox A, Natarajan K, Weng C. Using personal health records for automated clinical trials recruitment: the ePaIRing model. Summit Transl Bioinforma. 2009;2009:136–40.
7. Vayena E, Mastroianni A, Kahn J. Caught in the web: informed consent for online health research. Sci Transl Med. 2013;5:173fs6.
8. Do-It-Yourself Medicine | The Scientist Magazine® [Internet]. The Scientist. Available from: http://www.the-scientist.com/?articles.view/articleNo/34433/title/Do-It-Yourself-Medicine/. Cited 18 Mar 2014.
9. PatientsLikeMe Launches Campaign To Promote Health Data Sharing – iHealthBeat [Internet]. Available from: http://www.ihealthbeat.org/articles/2014/3/13/paitentslikeme-launches-campaign-to-promote-health-data-sharing. Cited 19 Mar 2014.
10. Cimino JJ. The false security of blind dates: chrononymization's lack of impact on data privacy of laboratory data. Appl Clin Inform. 2012;3:392–403.
11. Malin B, Benitez K, Masys D. Never too old for anonymity: a statistical standard for demographic data sharing via the HIPAA privacy rule. J Am Med Inform Assoc. 2011;18:3–10.
12. Pulley J, Clayton E, Bernard GR, Roden DM, Masys DR. Principles of human subjects protections applied in an opt-out, de-identified biobank. Clin Transl Sci. 2010;3:42–8.

13. Weng C, Li Y, Berhe S, Boland MR, Gao J, Hruby GW, et al. An Integrated Model for Patient Care and Clinical Trials (IMPACT) to support clinical research visit scheduling workflow for future learning health systems. J Biomed Inform. 2013;46:642–52.
14. Yoon S, Elhadad N, Bakken S. A practical approach for content mining of Tweets. Am J Prev Med. 2013;45:122–9.
15. Bove R, Secor E, Healy BC, Musallam A, Vaughan T, Glanz BI, et al. Evaluation of an online platform for multiple sclerosis research: patient description, validation of severity scale, and exploration of BMI effects on disease course. PLoS ONE. 2013;8:e59707.
16. Know Thyself: Tracking Every Facet of Life, from Sleep to Mood to Pain, 24/7/365 [Internet]. WIRED. Available from: http://www.wired.com/medtech/health/magazine/17-07/lbnp_knowthyself?currentPage=all. Cited 19 Mar 2014.
17. Hood NE, Brewer T, Jackson R, Wewers ME. Survey of community engagement in NIH-funded research. Clin Transl Sci. 2010;3:19–22.
18. Israel BA, Schulz AJ, Parker EA, Becker AB. Review of community-based research: assessing partnership approaches to improve public health. Annu Rev Public Health. 1998;19:173–202.
19. Fagnan LJ, Davis M, Deyo RA, Werner JJ, Stange KC. Linking practice-based research networks and clinical and translational science awards: new opportunities for community engagement by academic health centers. Acad Med J Assoc Am Med Coll. 2010;85:476–83.
20. Bakken S, Suero-Tejada N, Bigger JT, Wilcox A, Boden-Albala B. Weaving a strong trust fabric through community-engaged research: lessons from the WICER project about digital infrastructure for the learning health system. 2014 Jt. Summits Transl Sci. San Francisco; 2014.

Additional Reading

Gary Wolf. Know thyself: tracking every facet of life, from sleep to mood to pain. 24/7/365. WIRED. http://www.wired.com/medtech/health/magazine/17-07/lbnp_knowthyself.

PatientsLikeMe. PatientsLikeMe research. http://news.patientslikeme.com/research.

Rachel Rettner. What is the quantified self? LiveScience. http://www.livescience.com/39185-quantified-self-movement.html.

Wilcox A, Natarajan K, Weng C. Using personal health records for automated clinical trials recruitment: the ePaIRing model. Summit Transl Bioinforma. 2009;2009:136–40.

Part IV
The Future of Translational Informatics and Knowledge-Driven Healthcare

Chapter 10
Future Directions for Translational Informatics

Peter J. Embi and Philip R.O. Payne

By the End of This Chapter, Readers Should Be Able to

- Describe the relevance of translational informatics to the establishment of effective learning health systems
- Discuss the roles of various stakeholders to achieving the vision of knowledge-driven healthcare
- Explain the Evidence Generating Medicine approach and its relevance to the paradigm shift toward a virtuous evidence cycle between research and practice

10.1 Introduction

10.1.1 Revisiting the Vision of TI and Knowledge-Driven Healthcare

As the preceding chapters make clear, the future of healthcare holds great promise for accelerating biomedical discoveries and translating them into practice. Many advances in biomedical informatics in recent years have begun to reap benefits and address some of the fundamental challenges to translational science [1–3]. In many ways, these advances represent early demonstrations of the potential for

P.J. Embi, MD, MS, FACP, FACMI (✉)
Departments of Biomedical Informatics and Internal Medicine,
The Ohio State University, Columbus, OH, USA
e-mail: peter.embi@osumc.edu

P.R.O. Payne, PhD, FACMI
Department of Biomedical Informatics, The Ohio State University Wexner Medical Center,
Columbus, OH, USA

P.R.O. Payne, P.J. Embi (eds.), *Translational Informatics: Realizing the Promise of Knowledge-Driven Healthcare*, Health Informatics,
DOI 10.1007/978-1-4471-4646-9_10,

translational informatics [4]. Nevertheless, the frequently cited 17-year lag for biomedical discoveries to make their way into widespread practice likely persists, and there remains a great need for further progress in translational informatics to drive improvements in science and resultant knowledge-driven healthcare [5].

10.1.1.1 Capitalizing on the Promise of Translation

The past several years have seen substantial ongoing investments and policy interventions designed to encourage the adoption of healthcare information technologies, establish translational science infrastructure, and accelerate both healthcare and research. The goal of such efforts is to capitalize on scientific advances, compare the effectiveness of diagnostic and therapeutic interventions, and ultimately improve the quality and cost-effectiveness of healthcare [6, 7]. Even with existing investments and policies, the success of such initiatives will ultimately rely on additional and fundamental changes to the ways in which healthcare and biomedical research are practiced. Indeed, the very relationship between healthcare delivery and biomedical science still requires change before the ever-increasing amounts of biomedical information can be leveraged to accelerate both science and practice.

10.1.1.2 Creating Learning Healthcare Systems

Fundamental to such advances is the creation of what the Institute of Medicine has called a "learning health system" [8, 9]. Indeed, simultaneous advances in the widespread adoption of interoperable, standards-based EHRs, the infrastructure for conducting translational science, and the movement toward healthcare reform in order to improve quality and contain costs are helping to support the need for systematic "learning" (or science) as part of routine healthcare [10, 11]. While the increased collection and availability of healthcare data facilitated by EHR adoption and so-called "meaningful use" are necessary for creating such a learning health system, they are not sufficient. Many additional components, ranging from further technological advances, to regulatory and policy changes at the governmental level, to fiscal and administrative changes at the organizational level, and cultural shifts among the public will likely be needed [12]. Once created and functional at local, regional and national levels, such a learning health system will be essential to enabling translational informatics.

10.1.1.3 Addressing the Challenges and Opportunities of "Big Data" and Precision Medicine

As the adoption and use of EHRs continues and as that use is coupled with the simultaneous acceleration in the availability of vast amounts of non-EHR-based data about patients (e.g. internet-based social information, information about Internet usage including social media, and increasingly inexpensive and available

ways to test for and store one's genomic information, etc.), great opportunities and challenges will arise. Whatever the current capabilities of our computational infrastructure, there comes a point where the volume, velocity, and/or variety of the data present challenges to existing systems and users [13]. So-called big-data, while thereby challenging to leverage, clearly present great opportunities for advancing science and practice of healthcare.

A case in point involves the central promise of the genomics revolution – that of precision (or genetically guided) medicine. While potentially quite powerful, routine use of genomic information in the clinical setting has yet not come to pass. A recent systematic review of precision medicine workflows cites three critical barriers [14]. First, clinicians are poorly equipped to make sense of genomic data; genomic competency represents a limited part of medical training and guidelines change so rapidly they are often obsolete by the time a clinician has completed training and enters practice. Second, genetic experts who can provide actionable interpretations and who keep abreast of clinically relevant genomic discoveries are a relative few and are often not available to assist. Finally, the growth of both patient-specific genomic data via exome and genome sequencing as well as generated knowledge provides a unique challenge to existing electronic health record (EHR) platforms that were not designed to manage such information. Clearly, it is unreasonable to expect a single clinician to aggregate the needed information from multiple sources, keep abreast of and integrate the knowledge-based information to interpret the available data, and then render a cogent and appropriate diagnosis and treatment plan. Yet, this is how our health system is currently designed to operate. To address these issues and realize the benefits of the big data opportunities as facilitated by translational informatics innovations, fundamental shifts in our healthcare paradigm may be needed.

10.1.2 Current State (Where Are We Today?)

10.1.2.1 Technical Capabilities

Today's technologies, while advanced relative to the past and improving on both the clinical and research fronts [15], remain limited in certain ways related to the capacity and functionality needed for translational informatics. For instance, a key element to realizing a learning health system by enabling the kinds of team-science and rapid translation needed involves data exchange. As the volume and velocity of data expand, significant advances in network capabilities will be needed across the globe. Current efforts to create ubiquitous broadband communication capability is a step in the right direction, though it will likely still not be sufficient for widespread "big data" sharing efforts. As a result, initiatives are now underway to develop improved approaches to this problem [16]. Beyond the purely computational, other key technical capabilities including the development of services to enable federated data sharing, secure transmission of data, and meta-data are

needed. While many standards currently exist for describing healthcare data, robust standards and meta-data that describe the how, what, when, and where of data collection for the types of data essential to translational informatics remain underdeveloped [12, 17].

10.1.2.2 Cultural Norms

Beyond the technological, prevailing cultural norms significantly impact progress in translational informatics. In many ways cultural issues have a greater impact on the development adoption and use of translational informatics capabilities then do the technical and are often far more challenging to overcome. Indeed, as Reed Garnder, an informatics pioneer once stated, "The success of an (informatics) project is perhaps 80 % dependent upon the development of social and political interactions of the developer and 20 % or less on the implementation of the hardware and software technology!" [18]. Today's cultural norms in healthcare and biomedicine impact such translational informatics efforts greatly, largely because they are still very much aligned to support the traditional, non-translational, healthcare approaches. Only when those change over time will translational informatics professionals be able to make more rapid progress with fewer challenges; unfortunately culture change often takes much longer than technological change. After all, computers do what they are "told" whereas people often do not.

10.1.2.3 Organizational Factors

Along the same lines, organizational factors are also of great importance to the success or failure of translational efforts. After all, it is at the organizational level that decisions about relevant policies and funding are often made. Unfortunately, current models for organizational structure and alignment in healthcare are largely designed with only the healthcare delivery and not the translational research agenda in mind. Even at the level of empowered and resourced IT leadership, today's prevailing leaders such as CIOs and CMIOs concern themselves primarily with the clinical and operational missions, often with limited expertise in nor attention paid to the needs of the research mission. Recently, some academic health centers have seen fit to established IT leadership roles specifically focused on advancing the research mission, centralizing and empowering the governance of research informatics/IT alongside clinical and operational IT. Such models should help to facilitate responsiveness and strategic planning focused on the unique and often complex needs of the research community. They should also yield economies of scale both by enabling investments in IT infrastructure that can often serve clinical, operational and research mission areas, and by advancing translation between knowledge and practice [19].

10.1.3 Need to Involve All Stakeholders to Achieve This Vision

In order to fully realize the promise of translational informatics, fundamental changes in technological, socio-cultural, and organizational realities must take place. As such, no single group alone will be able to bring about the changes needed. Instead, such a vision will only be realized when all relevant stakeholders (i.e. patients, practitioners, healthcare organizations, government, industry, etc.) take part and work toward the same goal. There are early signs that this is beginning to happen, spurred by the investments and initiatives alluded to above. Still, a roadmap for the future is essential such that all will be able to follow it.

10.2 Critical Dimensions for a TI and Knowledge-Driven Healthcare "Roadmap"

10.2.1 Aligning Technical Capabilities with Motivating Problems and Designing for an Informatics "Ecosystem"

10.2.1.1 Overview

Fundamental to the advances needed is the development of a framework that informs the basic flow of data, information and knowledge essential to translational informatics. As depicted in Fig. 10.1, the vision for patient care involves the Capture of both clinical data and bio-molecular data about a patient that when combined with domain knowledge and properly conceptualized inform the diagnosis prevention and treatment of the patient. It is this fundamental conceptual workflow that underlies the changes envisioned for knowledge-driven healthcare.

10.2.1.2 Focus on Applications and Infrastructure

One of the challenge facing today's IT and informatics environment is the presence of multiple often disconnected systems that are not able to interchange data based on common standards. In order to enable the advances in vision for translational informatics the future must include in focus on taking current knowledge and applications and integrating them in ways that will infrastructure capable of supporting the mirror you workflows of clinicians and scientists working to advance translational medicine. The presence of comprehensive electronic health records that integrate multiple formerly disparate systems into what functions as a common platform is a good example of the kinds of application development and infrastructure that is needed. However, while electronic health records have

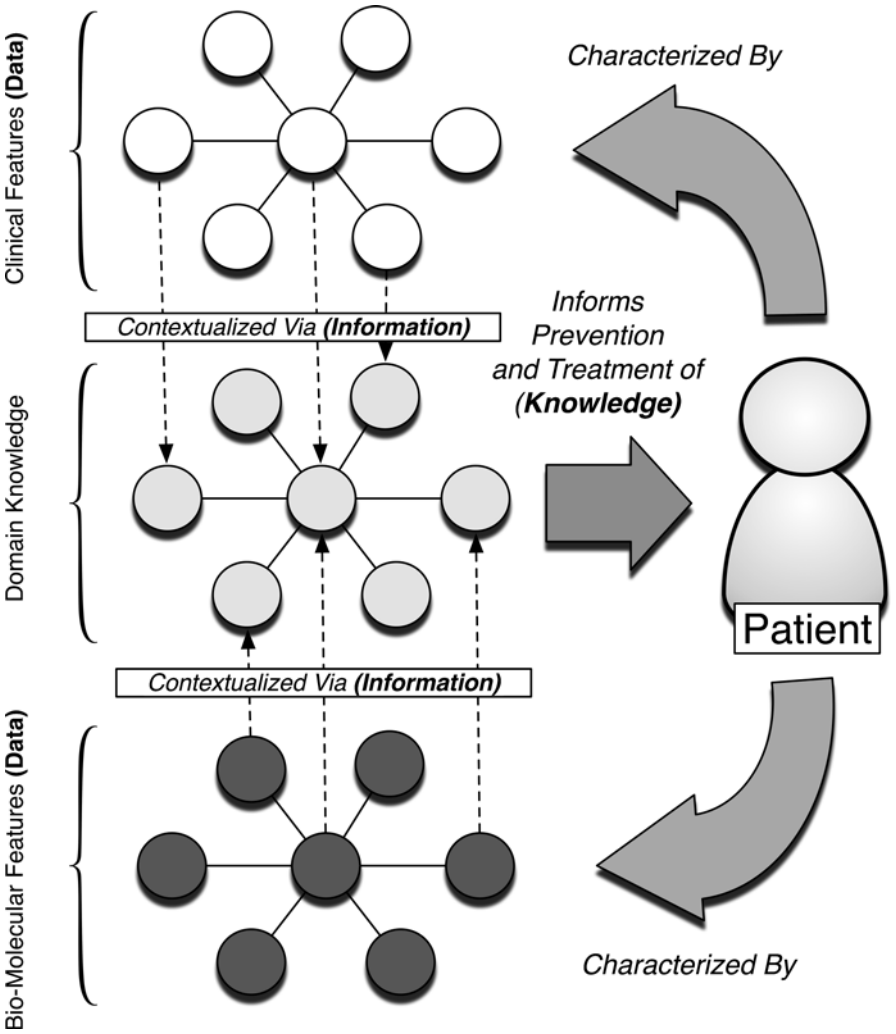

Fig. 10.1 The conceptual framework underlying many of the advances needed to enable translational informatics

accomplished this for the purposes of clinical care, there are advances needed in those platforms in order to integrate them with other translational tools in order to realize the vision articulated herein. Once established, an infrastructure that can readily transact data between component systems in efficient, effective and secure ways will significantly expand the capabilities needed in order to not only achieve meaningful use for clinical care, but also create a learning health system capable of improving quality and advancing science by leveraging data collected first and foremost for clinical purposes.

10.2.1.3 Lowering the Barriers to Adopting/Adapting Data Standards

Despite the presence of many standards available for the various types of data that will need to be exchanged, there are many more that are underdeveloped or simply nonexistent for rapidly emerging types of data essential to translational science. Moreover, the adoption of said standards by various stakeholders including vendors and healthcare organizations has traditionally been challenging, requiring significant and limited expertise and resources. In order to realize the vision of a learning health system and translational informatics capacity, standards to govern the use storage and retrieval of not only clinical but research specific data must be improved such that it is clear which standards are necessary for particular purpose and their adoption and application can be readily operationalized across multiple settings.

10.2.1.4 Improving UX (User Experience)

Any system is only as valuable to an end user as is the interface that user interacts with. Unfortunately, health IT systems have traditionally been plagued by limited attention to the next level user experience, the kind of experience that has become common in consumer applications such as smart phones. Given the complexities inherent in integrating and interpreting data in order to then apply it in real world clinical settings (as is required to translation knowledge into practice), user interface design will become increasingly critical part of the informatics efforts need to establish a translational informatics infrastructure.

10.2.2 Overcoming Socio-cultural Barriers to Team Science, Rapid Cycle Translation, and Systems Thinking

10.2.2.1 Creating Understanding of Open/Active Research Questions and Systems Thinking

In the future, it will be the responsibility of not just researchers but also clinicians and even patients to actively develop and contribute to the research process. In order to enable the kind of activity those who currently do not consider the development and pursuit of research or learning as part of their job will need to be properly incentive to do so. Models for recognizing measuring and incentivizing subjectivity have been proposed, and will need to be implemented in order to create a culture that encourages participant stakeholders' contributions to translational science [20]. In addition, scientists will increasingly be required to consider the clinic around more so than make it today and bring the development of systems thinking from biology to the very practice of translational and precision medicine. Only through such culture shifts in the clinical and biomedical scientific rounds will true advances in team science and translational activities take place.

10.2.2.2 Establishing Career Paths for Translational Professionals and Training a Multi-faceted Workforce (Tailoring Training to Roles and Responsibilities)

A key element that will need to be addressed moving forward is the development and growth of a dedicated workforce trained to enable translational science and practice. Currently professionals working in the transitional field tend to have expertise on one or another and of the translational spectrum. The same is typically true for informatics and IT professionals. In the future those working at the intersection of translational science will need to grow and their training to include formal knowledge in a variety of realms as well as in the methodologies necessary to enable team science. Furthermore, incentives including funding for career paths and opportunities for advancement of promotion will need to be aligned much better than they are today in order to recognize the value of those working to advance translational science and the informatics aspects therein, in particular. Only by creating attractive structures and career paths with visible success stories that will motivate trainees to enter the field will we overcome the current severe shortage and experts needed to achieve the vision laid out herein.

In addition to training dedicated professionals in this area, those who will continue to work on opposite ends of the translational spectrum must also be educated as they will be critical members of the translational workforce. Indeed, there is a need to educate learners at varying levels of intensity based upon their stage of training, their role in the research and informatics/IT enterprise, and their career goals. A description of the varying types of learners and the related types of training that would likely be relevant/of interest to such groups of learners is depicted in Table 10.1. As the chart depicts using different size marks, learners in each category on the left may opt for more or less intensive training, but we have indicated with the large "**X**" those offerings we think most appropriate to each type. Already, multi-week courses and tutorials have been developed to start addressing this need [21].

Table 10.1 Educational program applicability by learner stage/role

	Educational program			
Learner stage/role	Tutorial	Multi-week course	Certificate program	Master's degree
Student/resident, clinicians, faculty, leadership	**X**	x		
Investigators, research staff, or informatician liaisons	x	**X**	x	
Informatician, investigator, or research staff who will use or support Research Informatics		x	**X**	x
Informatician with Research Informatics career focus			x	**X**

X = most applicable; x = possibly applicable

10.2.3 Creating Organizational Settings That Support and Enable TI and Knowledge-Driven Healthcare

10.2.3.1 Creating Durable Homes and Funding Mechanisms

Given the critical importance of translational science at the informatics methods to facilitate it, the creation of dedicated academic units and organizations as well as funding mechanisms focused on the unique aspects of translational informatics will become increasingly important. Well initially embedded in existing organizational structures and funding apparatus, the importance of this work would benefit greatly from the creation of durable and dedicated homes in funding environments that would not be subject to the wins of those who may not understand the nuances and complexities inherent to work.

10.2.3.2 Removing Artificial Barriers Between Research, Practice, and Training

At both the organizational and regulatory levels it is essential that the current outmoded and even artificial barriers between research practice and training that currently impede progress and translational informatics be removed. Moving forward organizational structures and policies must be established that facilitate the free flow of information between the research and practice environments in order to enable the kinds of translation needed to realize the vision laid out here.

10.2.3.3 Catalyzing a Culture of Science and Innovation

By achieving such advances organizational settings will be established that not only allow for but actively support and enable translational informatics and knowledge driven healthcare by creating a culture of science and innovation that exists synergistically with healthcare practice.

10.3 A Paradigm Shift to an Evidence Generating Medicine (EGM) Approach

10.3.1 Overview

It has become painful evident that the existing healthcare paradigm which distinguishes research from clinical care as distinct activities impedes translational science and the related ongoing efforts in translational informatics. Current initiatives demand that we leverage point of care activities information and resources to

generate evidence and improve the quality of healthcare. Unfortunately the current healthcare system is designed to enable the care of a single patient of the time and not to focus on research or systematic learning. Without a fundamental paradigm shift the challenges preventing advances in translational informatics will continue unabated, and the necessary technical, organizational and cultural changes mentioned above will remain elusive, at best. Let us explore the basis for this shift.

10.3.2 Traditional Model

10.3.2.1 Unidirectional Translation of Knowledge from Research into Practice (EBM)

The traditional research-practice paradigm upon which our healthcare system is built defines research and practice as entirely distinct endeavors that share a unidirectional relationship, with research findings being applied to practice, ideally via an Evidence Based Medicine (EBM) approach. While this prevailing paradigm is well suited to individual healthcare delivery, it is at odds with and often frustrates the very research and healthcare improvement initiatives in which we are investing. Guided by this current research-practice paradigm, mal-aligned organizational, financial, and policy decisions impede the integration of research and practice, frustrating ongoing efforts [22]. The prevailing paradigm even feeds into problematic perceptions at the individual level, such as the view by clinicians and patients alike that engaging in research activities is not part of what should happen at the bedside.

10.3.2.2 Limited Feedback Between Activities

A major and significant consequence of the prevailing paradigm as it relates to transitional informatics is the limited amount of information flowing feedback between clinical and research activities. Indeed, the challenges facing biomedical informatics professionals working at translational junctures are evident on a daily basis, and they expose the flaws in the current paradigm. Unfortunately, the notion that simply digitizing medical information will automatically enable downstream research and evidence generation flawed [23]. Well data collected at the point of Karen Tesdal form is certainly more helpful than that collected otherwise such that are often incomplete or inaccurate for the needs of researchers. Moreover current regulatory and policy frameworks limit the feedback loops between the clinical and research environments due to concerns such as those related to privacy and the reuse of data. Groups are beginning to report on such concerns and suggest options for reconciling the importance of privacy and advancing research simultaneously [11, 24].

10.3.3 New Model

10.3.3.1 Enabling a Virtuous Cycle Between Research and Practice to Create a Learning Health System

As previously mentioned, the IOM has recognized the need for a "learning healthcare system" that will systematically improve based on evidence gleaned in the course of healthcare delivery [8]. What the above noted challenge make clear is that in order to achieve such a vision, a fundamental paradigm shift in the relationship between research and practice is needed. Rather than the current unidirectional relationship between research and practice, the new paradigm recognizes the essential connection between practice and research. The end result is the cyclical generation, application, and refinement of high-quality evidence (via Evidence Based Medicine) with the insertion of Evidence Generating Medicine (EGM) as the missing concept needed to create such a virtuous evidence cycle (Fig. 10.2).

As defined in our recent publication on the topic, EGM involves: "the systematic consideration and incorporation of evidence generating activities into the organization and practice of healthcare to advance biomedical discovery and improve the health of individuals and populations" [25].

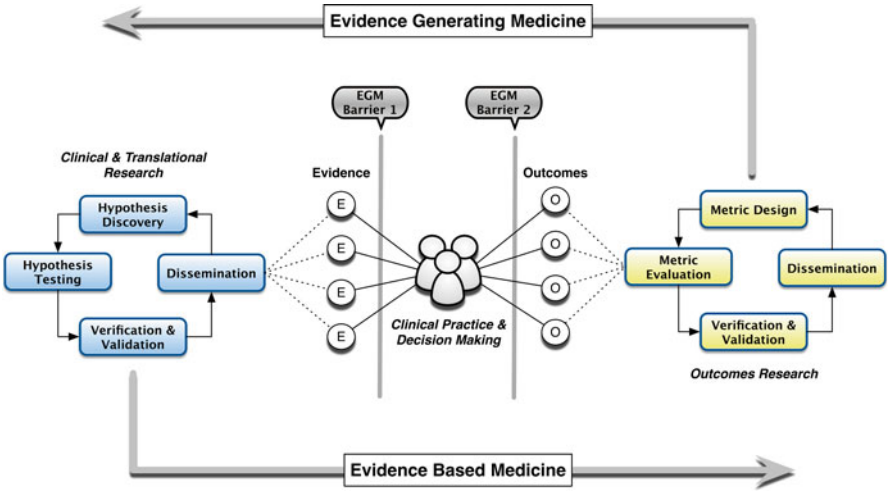

Fig. 10.2 The "Evidence Cycle" that emerges with a paradigm shift from a uni-directional model of the research-practice relationship to one that is a "virtuous cycle" of evidence, including both the generation of evidence through practice-enabled research (i.e. via Evidence Generating Medicine) and the application of research-derived evidence to practice (i.e. Evidence Based Medicine)

10.3.3.2 Key Stakeholders and Factors Essential to Operationalizing EGM

By advancing EGM we achieve and evidence cycle that is essential to altering the research–practice paradigm. This new paradigm informs the work of various stakeholders, from individual researchers, practitioners, and the public, to institutions, government agencies and private-sector concerns. Once recognized as the prevailing paradigm, all such stakeholders can begin to see they have a role in advancing the evidence cycle – including research and translation.

To operationalize EGM, there are a range of enabling factors that span these stakeholders. These include policy and organizational factors, fiscal and administrative factors, and Informatics and Health IT factors [25].

10.3.3.3 Implications of Adopting and Operationalizing EGM

Adopting EGM and creating the evidence cycle enables us to address the challenges facing our research and healthcare enterprises. Indeed with EGM in mind the creation of a learning health system becomes not only something that would be nice to have but something that is essential to an effective and efficient health system. Moreover, in order to create an evidence cycle as required by such a paradigm, the various goals of technological, organizational, and cultural changes must take place. As a result, changes to align policies and resources should be made that will improve the efforts of those working to advance translational informatics and create learning health systems.

10.4 Conclusion

Realizing the promise of translational informatics requires a significant amount of coordinated effort and change. In addition to advancing technologies changes to organizational structures regulatory and policy frameworks as well as cultural changes must take place. However the momentum to do so is clearly in place and we have never been better position to realize the vision of translational informatics and knowledge driven healthcare as we are today. By following the frameworks and paradigms laid out above the promise of translational informatics to significantly improve personalized and precision medicine is within our grasp.

Discussion Points

- How can the theories and methods that comprise Translational Informatics contribute to the creation of a learning health system?
- What challenges and opportunities exist relative to aligning technical capabilities with motivating problems for designing an informatics “ecosystem”?
- What important socio-cultural barriers need to be overcome in order to operationalize science, rapid cycle translation, and systems thinking?

- How can we go about creating organizational settings that support and enable TI and knowledge-driven healthcare?
- What are the critical issues that impact our ability to deliver on the vision of EGM – including stakeholder alignment, as well as policy, fiscal, administrative and technological factors at the local and broader levels?

References

1. Chung TK, Kukafka R, Johnson SB. Reengineering clinical research with informatics. J Investig Med. 2006;54(6):327–33. PubMed PMID: 17134616.
2. 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. PubMed PMID: 15974245.
3. Sung NS, Crowley Jr WF, Genel M, Salber P, Sandy L, Sherwood LM, et al. Central challenges facing the national clinical research enterprise. JAMA. 2003;289(10):1278–87. PubMed PMID: 12633190, Epub 2003/03/14. eng.
4. Payne PR, Embi PJ, Sen CK. Translational informatics: enabling high-throughput research paradigms. Physiol Genomics. 2009;39(3):131–40. PubMed PMID: 19737991, Pubmed Central PMCID: 2789669.
5. Balas EA, Suzanne AB. Managing clinical knowledge for health care improvement. Yearb Med Inform. 2000;(2000):65–70.
6. Kaiser J. Research funding. Health bill backs evidence-based medicine, new drug studies. Science. 2010;327(5973):1562.
7. Steinbrook R. The NIH, stimulus–the recovery act and biomedical research. N Engl J Med. 2009;360(15):1479–81. PubMed PMID: 19357402, Epub 2009/04/10. eng.
8. Olsen L, Aisner D, Michael McGinnis J. The learning healthcare system: workshop summary. Washington, DC: The National Academies Press; 2007. 374 p.
9. Smith M, Halvorson G, Kaplan G. What's needed is a health care system that learns: recommendations from an IOM report. JAMA. 2012;308(16):1637–8.
10. Payne PR, Embi PJ, Niland J. Foundational biomedical informatics research in the clinical and translational science era: a call to action. J Am Med Inf Assoc. 2010;17(6):615–6. PubMed PMID: 20962120, Pubmed Central PMCID: 3000758.
11. Weiner MG, Embi PJ. Toward reuse of clinical data for research and quality improvement: the end of the beginning? Ann Intern Med. 2009;151(5):359–60. PubMed PMID: 19638404.
12. Friedman CP, Wong AK, Blumenthal D. Achieving a nationwide learning health system. Sci Transl Med. 2010;2(57):57cm29. PubMed PMID: 21068440.
13. McAfee A, Brynjolfsson E. Big data: the management revolution. Harv Bus Rev. 2012;90(10): 60–6. 8, 128, PubMed PMID: 23074865.
14. Welch BM, Kawamoto K. Clinical decision support for genetically guided personalized medicine: a systematic review. J Am Med Inf Assoc. 2013;20(2):388–400. PubMed PMID: 22922173. Pubmed Central PMCID: 3638177. Epub 2012/08/28. eng.
15. Murphy SN, Dubey A, Embi PJ, Harris PA, Richter BG, Turisco F, et al. Current state of information technologies for the clinical research enterprise across academic medical centers. Clin Transl Sci. 2012;5(3):281–4. PubMed PMID: 22686207.
16. Ohno-Machado L. NIH's big data to knowledge initiative and the advancement of biomedical informatics. J Am Med Inform Assoc. 2014;21(2):193. PubMed PMID: 24509598. Pubmed Central PMCID: 3932475.
17. Richesson RL, Nadkarni P. Data standards for clinical research data collection forms: current status and challenges. J Am Med Inf Assoc. 2011;18(3):341–6. English.
18. O'Carroll PW. Public health informatics and information systems. New York: Springer; 2003. xxvii, 790 p.

19. Embi PJ, Tachinardi U, Lussier Y, Starren J, Silverstein J. Integrating governance of research informatics and health care IT across an enterprise: experiences from the trenches. AMIA Jt Summits Transl Sci Proc. 2013;2013:60–2. PubMed PMID: 24303236, Pubmed Central PMCID: 3845750.
20. Embi PJ, Tsevat J. Commentary: the relative research unit: providing incentives for clinician participation in research activities. Acad Med. 2012;87(1):11–4. PubMed PMID: 22201633, Pubmed Central PMCID: 3914136.
21. The Ohio State University-AMIA 10x10 program in Clinical Research Informatics [July 14, 2011]. Available from: http://www.amia.org/education/academic-and-training-programs/10x10-ohio-state-university.
22. 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. PubMed PMID: 19261934. Pubmed Central PMCID: 2732242.
23. Hersh WR, Weiner MG, Embi PJ, Logan JR, Payne PR, Bernstam EV, et al. Caveats for the use of operational electronic health record data in comparative effectiveness research. Med Care. 2013;51(8 Suppl 3):S30–7. PubMed PMID: 23774517, Pubmed Central PMCID: 3748381.
24. Hripcsak G, Bloomrosen M, FlatelyBrennan P, Chute CG, Cimino J, Detmer DE, et al. Health data use, stewardship, and governance: ongoing gaps and challenges: a report from AMIA's 2012 health policy meeting. J Am Med Inform Assoc. 2014;21(2):204–11. PubMed PMID: 24169275. Pubmed Central PMCID: 3932468.
25. Embi PJ, Payne PR. Evidence generating medicine: redefining the research-practice relationship to complete the evidence cycle. Med Care. 2013;51(8 Suppl 3):S87–91. PubMed PMID: 23793052.

Additional Reading

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. PubMed PMID: 19261934. Pubmed Central PMCID: 2732242.

Embi PJ, Payne PR. Evidence generating medicine: redefining the research-practice relationship to complete the evidence cycle. Med Care. 2013;51(8 Suppl 3):S87–91. PubMed PMID: 23793052.

Hersh WR, Weiner MG, Embi PJ, Logan JR, Payne PR, Bernstam EV, et al. Caveats for the use of operational electronic health record data in comparative effectiveness research. Med Care. 2013;51(8 Suppl 3):S30–7. PubMed PMID: 23774517, Pubmed Central PMCID: 3748381.

Hripcsak G, Bloomrosen M, FlatelyBrennan P, Chute CG, Cimino J, Detmer DE, et al. Health data use, stewardship, and governance: ongoing gaps and challenges: a report from AMIA's 2012 Health Policy Meeting. J Am Med Inform Assoc. 2014;21(2):204–11. PubMed PMID: 24169275. Pubmed Central PMCID: 3932468.

Murphy SN, Dubey A, Embi PJ, Harris PA, Richter BG, Turisco F, et al. Current state of information technologies for the clinical research enterprise across academic medical centers. Clin Transl Sci. 2012;5(3):281–4. PubMed PMID: 22686207.

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. PubMed PMID: 15974245.

Payne PR, Embi PJ, Sen CK. Translational informatics: enabling high-throughput research paradigms. Physiol Genomics. 2009;39(3):131–40. PubMed PMID: 19737991, Pubmed Central PMCID: 2789669.

Sung NS, Crowley Jr WF, Genel M, Salber P, Sandy L, Sherwood LM, et al. Central challenges facing the national clinical research enterprise. JAMA. 2003;289(10):1278–87. PubMed PMID: 12633190, Epub 2003/03/14. eng.

Index

P.R.O. Payne, P.J. Embi (eds.), *Translational Informatics: Realizing the Promise of Knowledge-Driven Healthcare*, Health Informatics,
DOI 10.1007/978-1-4471-4646-9, © Springer-Verlag London 2015