Physiol Genomics 39: 131–140, 2009.
First published September 8, 2009; doi:10.1152/physiolgenomics.00050.2009.

Review

Translational informatics: enabling high-throughput research paradigms

Philip R. O. Payne,[1,2] Peter J. Embi,[4,5] and Chandan K. Sen[2,3]

[1]*Department of Biomedical Informatics,* [2]*Center for Clinical and Translational Science, and* [3]*Department of Surgery, The Ohio State University, Columbus; and* [4]*Center for Health Informatics and* [5] *Department of Medicine, University of Cincinnati, Cincinnati, Ohio*

Submitted 11 March 2009; accepted in final form 1 September 2009

Payne PRO, Embi PJ, Sen CK. Translational informatics: enabling high-throughput research paradigms. *Physiol Genomics* 39: 131–140, 2009. First published September 8, 2009; doi:10.1152/physiolgenomics.00050.2009.—A common thread throughout the clinical and translational research domains is the need to collect, manage, integrate, analyze, and disseminate large-scale, heterogeneous biomedical data sets. However, well-established and broadly adopted theoretical and practical frameworks and models intended to address such needs are conspicuously absent in the published literature or other reputable knowledge sources. Instead, the development and execution of multidisciplinary, clinical, or translational studies are significantly limited by the propagation of "silos" of both data and expertise. Motivated by this fundamental challenge, we report upon the current state and evolution of biomedical informatics as it pertains to the conduct of high-throughput clinical and translational research and will present both a conceptual and practical framework for the design and execution of informatics-enabled studies. The objective of presenting such findings and constructs is to provide the clinical and translational research community with a common frame of reference for discussing and expanding upon such models and methodologies.

biomedical research

THE MODERN BIOMEDICAL RESEARCH domain has experienced a fundamental shift toward integrative and translational methodologies and frameworks over the past several years. This shift has been manifested in a number of ways, including the launch of the National Institutes of Health (NIH) Roadmap initiative (82–84), which has resulted in the creation of the Clinical and Translational Science Award (CTSA) program (83), as well as the rapid growth and increasing availability of high-throughput biomolecular technologies and corresponding bio-marker-to-phenotype mapping efforts (11). A commonly reported thread in a broad variety of reports and commentaries concerned with this evolution focuses on the challenges and requirements related to the collection, management, integration, analysis, and dissemination of large-scale, heterogeneous biomedical data sets (19, 25, 58, 72). However, well-established and broadly adopted theoretical and practical frameworks intended to address such needs are still conspicuously lacking in the published literature or other reputable knowledge sources (14, 46, 58). Instead, the development and execution of integrative clinical or translational research are significantly limited by the propagation of "silos" of both data and expertise. Motivated by this fundamental challenge, the remainder of this manuscript will present the findings of a four-phase approach to define the current state and practice of clinical/translational science and its intersection with biomedical informatics.

Address for reprint requests and other correspondence: P. R. O. Payne, The Ohio State Univ., Dept. of Biomedical Informatics, 3190 Graves Hall, 333 W. 10th Ave., Columbus, OH 43210 (e-mail: philip.payne@osumc.edu).

METHODOLOGY

As noted in the introduction and illustrated in Fig. 1, the phases and associated findings of the four phase approach used to develop this manuscript can be broadly divided into the following four categories: *1*) a review of the current state of biomedical informatics as it pertains to the conduct of high-throughput clinical and translational research, with an emphasis on key definitions and critical information management challenges; *2*) the definition of a conceptual framework for translational informatics that is intended to foster greater integration of the biomedical informatics and the clinical or translational research domains, informed by the exemplary experiences of the authors and a number of contributory literature reviews; *3*) the definition of a practical model for the design and implementation of translational informatics projects; and *4*) a synthesis of the preceding research products and an associated set of recommendations concerning how to fully realize the potential benefits afforded by systematic approaches to translational informatics in the contemporary biomedical research environment. Our objectives in presenting these findings are to: *1*) introduce researchers who are new to the clinical and translational science domains to the basic concepts, challenges, and informatics-related tools and methods incumbent to their domain; and *2*) provide experienced clinical, translational, and informatics researchers with a broad framework in which to situate their current work and to identify potentially novel linkages between their efforts and emerging challenges and opportunities in the translational informatics domain. This work is not intended to serve as a comprehensive review of the current state of knowledge in the clinical or translational research informatics domains, an area recently addressed in a

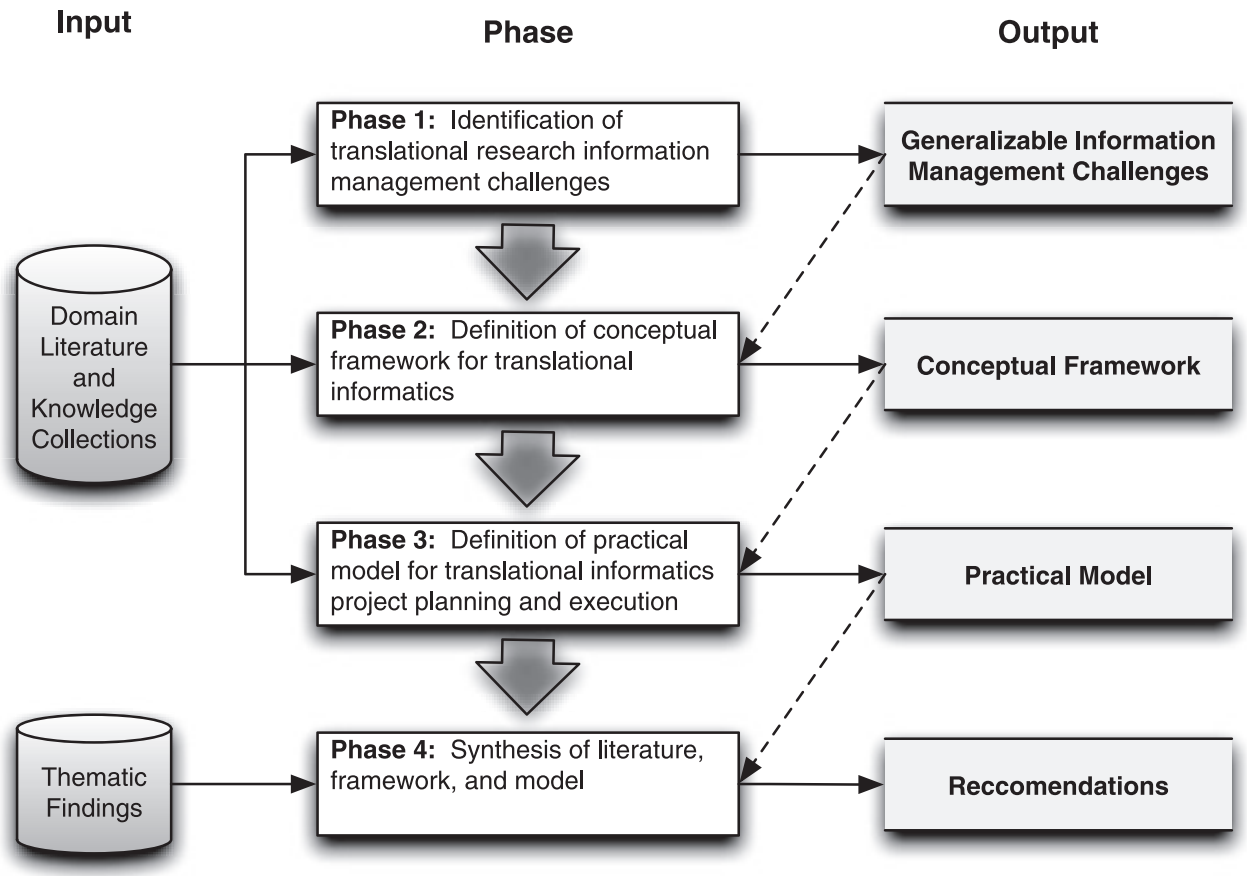

Fig. 1. Overview of our 4-phase approach, highlighting sources of information (input), findings, or knowledge products (output) and the relationships between such components.

number of published reports and manuscripts (8, 19, 25, 58, 63, 64, 72).

DEFINITIONS AND INFORMATION MANAGEMENT CHALLENGES IN CLINICAL AND TRANSLATIONAL RESEARCH

Working Definitions

To provide sufficient context and scope to our ensuing discussion, we will define translational research per the conventions provided by the NIH as follows:

> "Translational research includes two areas of translation. One is the process of applying discoveries generated during research in the laboratory, and in preclinical studies, to the development of trials and studies in humans. The second area of translation concerns research aimed at enhancing the adoption of best practices in the community. Cost-effectiveness of prevention and treatment strategies is also an important part of translational science." (62)

Several recent publications have defined a translational research cycle, which involves the translational of knowledge and evidence from "the bench" (e.g., laboratory-based discoveries) to "the bedside" (e.g., clinical or public health interventions informed by basic science and clinical research), and reciprocally from the bedside back to the bench (e.g., basic science studies informed by observations from the point-of-care) (72). Within this translational cycle, Sung and colleagues (72) have defined two critical blockages that exist between basic science discovery and the design of prospective clinical studies, and subsequently between the knowledge generated during clinical studies and the provision of evidence-based care in the clinical or public health settings. These are known as the T1 and T2 blocks, respectively (Fig. 2). Much of the work conducted under the auspices of the NIH Roadmap initiative is specifically focused on identifying approaches or policies that can mitigate these T1 and T2 blockages and thus increase the speed and efficiency by which new biomedical knowledge can be realized in terms of improved health and patient outcomes.

In addition to defining translational research and the translational research cycle, we will define Biomedical Informatics per the convention provided by Shortliffe and Cimino (67) as follows:

> "Biomedical Informatics is the scientific field that deals with the storage, retrieval, sharing, and optimal use of biomedical information, data, and knowledge for problem solving and decision making. It touches on all basic and applied fields in biomedical science and is closely tied to modern information technologies, notably in the areas of computing and communication."

For the purposes of this article, we will be focusing primarily upon the role of biomedical informatics relative to information management challenges, frameworks, and practical methods that may be applied to overcome the T1 translational block (e.g., from bench to clinical research).

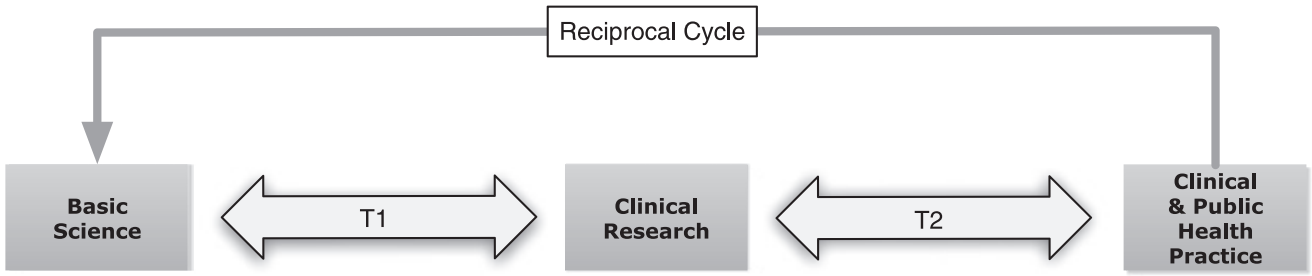

Fig. 2. Translational research cycle, illustrating the T1 and T2 "translational blocks," as described by Sung and colleagues (72).

Information Management Challenges in Clinical and Translational Research

The benefits, challenges, and opportunities afforded by integrating biomedical informatics across the clinical and translational research spectrum are many (11, 19, 58, 72). At a high level, the essential information management challenges to be addressed by such integration can be classified as belonging to one or more of the following categories.

The ability to collect and manage heterogeneous data sets with increasing levels of dimensionality. With the ever-increasing availability of high-value patient-centric phenotypic data sources, such as electronic health records (EHRs), clinical trials management systems (CTMS), as well as biomolecular measurements such as genotypic and proteomic expression profiles fed by a growing suite of instrumentation platforms, the size and complexity of data sets that researchers can collect, store, and retrieve on a regular basis are growing at an exponential rate (11, 14, 34, 42, 46). At the same time, the data management practices currently used in many basic and clinical science research settings rely on the use of conventional databases or individual file-based approaches that are ill-suited to enabling interaction with large-scale data sets (29, 42, 64). Therefore, the dissemination and adoption of advanced information management platforms that allow researchers and their staff to focus on fundamental scientific problems rather than practical informatics needs are critical to reducing the data management burden associated with today's multidimensional data (4, 14, 48, 58). In addition, with the impetus to link high-throughput biomolecular and phenotypic data to better understand potential relationships between variables that could inform novel diagnostic or treatment planning approaches, it is also imperative that the semantics of such data be well understood (59, 63, 64). Such semantic interoperability between data sets is at least if not more important to clinical and translational research efforts as "technical" or "syntactic" interoperability. Achieving semantic interoperability requires the use of informatics-based approaches to map among various data representation schemas and to use those mappings to support data integration or analysis operations (63, 64). For example, if a variable of interest in a genomic data set is encoded with the name of the probe used to measure the expression level of a particular gene, and a phenotypic variable that is concerned with production of a protein encoded by that same gene is labeled using the name of that protein, then it is necessary to use one or more ontologies or terminologies to ascertain that the two variables ultimately resolve to a relationship anchored upon a shared gene.

The need to employ knowledge-anchored methods to discover and test hypotheses concerning linkages between phenotypic and biomolecular variables of interest. Given the high-throughput data sets described in the preceding challenge, a corresponding high-throughput hypothesis discovery and testing challenge also exists. Contemporary approaches to hypothesis discovery and testing primarily rely upon the intuition of the individual investigator or his/her team to pose a question and then carry out testing to validate or refine that question (11, 13). Such an approach is often feasible when working with data sets comprising up to hundreds of variables, where a researcher or team member can reasonably possess enough domain knowledge to understand and hypothesize about them. However, as data sets expand by multiple orders of magnitude to incorporate thousands or even millions of variables, such an approach quickly becomes intractable due to inherent human cognitive limitations (59, 64). At the same time, relevant domain knowledge needed to reason upon and generate hypotheses relative to such data sets is often incorporated into a variety of sources, such as public databases, terminologies, ontologies, and published literature (59). However, tools and applications that allow researchers to access and extract this domain knowledge from such sources, and then use those resulting knowledge extracts to induce large sets of readily testable hypotheses relative to a targeted data set, are still in the very preliminary stages of research and development (59, 64). Therefore, significant additional effort is needed to validate such tools and provide them for regular use by the clinical and translational research community.

The provision of systematic and extensible platforms capable of expediting data integration and analysis workflows. A third challenge in the context of integrating biomedical informatics and translational research pertains to the availability of systematic data-analytic "pipelining" tools 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 (52, 75). The value of such data analysis pipelining is two-fold: *1*) pipelines support the rapid execution of complex data analysis plans that would otherwise require multiple time- and resource-intensive manual processes to collect and manipulate source data and *2*) pipelines enable the collection of meta-data describing the data analysis process being performed. This meta-data can be used both 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 (52, 75). Recent research and development in the

biomedical informatics domain has yielded highly promising technology platforms capable of supporting such data-analytic pipelining, such as the caGrid middleware (52). However, despite the promise of such platforms, their adoption rates are still relatively low, largely owing to a combination of data ownership/security and socio-technical barriers, which will require community-based consensus and improved understanding of contributing human factors to overcome (4, 40).

Dissemination of evidence and knowledge. The fourth and final challenge we will enumerate in the context of integrating clinical or translational research and biomedical informatics is the ability to disseminate the evidence and knowledge gained during the conduct of constituent activities to the intended end-user(s) in a resource efficient and timely manner. It is a well-known phenomenon that the time period required to move a basic science discovery into clinical research and ultimately clinical or public health practice can span in excess of a decade (14, 19, 32, 72, 84). Numerous studies have identified the dissemination or exchange of information between various research and operational settings (e.g., basic science, preclinical investigation, clinical trials, commercialization, implementation in clinical care or public health practice) as one of the most pressing issues contributing to long research, development, and implementation lifecycles (14). Again, as was the case in the context of the preceding challenges, a wide variety of informatics platforms 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 (19, 42, 58, 69). Of note, throughout the clinical, translational science, and informatics communities and their constituent bodies of literature, a lexicon that includes the terms "dissemination" and "exchange" relative to new knowledge and evidence is regularly used, despite the passive nature of such descriptors. By facilitating greater adoption, understanding, and appropriate use of the preceding informatics platforms, this paradigm could shift to emphasize a more active model, which would incorporate approaches more accurately labeled as "teaching" and "learning."

CONCEPTUAL FRAMEWORK

To overcome the previously introduced challenges, particularly in the context of information-intensive, multidisciplinary team-based clinical and translational research projects, a framework should and must be developed to enable the categorization and conceptual integration of the major sources of information and associated information needs involved in such endeavors. Such a framework can ultimately assist in: *1*) identifying major categories of information to be collected, managed, and disseminated throughout the clinical or translational research process and the ways in which they relate to one another, thus enabling the development of integrative platforms capable of addressing such needs in a systematic manner; *2*) providing individual researchers with the ability to understand how their unique activities contribute to a broader goal of generating new knowledge or evidence that spans multiple domains or subdomains, thus increasing awareness of the needs to exchange or disseminate such information in an easily and readily consumable manner; and *3*) supporting the modeling and development of cross-cutting technology and socio-technical approaches that are specifically targeted at achieving high-level, translational integration spanning what are often distinct data, knowledge and/or evidence silos.

Based upon our experiences at both The Ohio State University and the University of Cincinnati, as well as prior surveys of the state of biomedical informatics relative to the clinical and translational science domains as conducted by the authors (25, 58), we have developed a prototype for such a framework, which we will present in the following discussion. Central to this framework are a number of critical information types involved in the conduct of clinical and translational research, as enumerated below and illustrated in Fig. 3.

Individual and/or population phenotype. This information type generally involves data elements and variables that describe human and/or animal-derived characteristics at the individual or population levels that relate to the physiological and behavioral manifestation of healthy and disease states. Examples can include demographics, clinical exam findings, qualitative characteristics (e.g., quality of life, disease-specific performance status/staging), and analytic laboratory testing results such as those commonly employed in clinical care or equivalent activities (13). This information is primarily generated via public health, clinical care, and clinical research operations (11, 42).

Individual and/or population biomarkers. This information type generally involves data elements and variables that describe human and/or animal-derived characteristics at the individual or population levels that relate to the biomolecular manifestation of healthy and disease states. Examples include genomic, proteomic, and metabolomic expression profiles, as well as novel biomolecular assays capable of differentiating normal and abnormal (e.g., diseased) biomolecular structure and function (11, 13). Such information is primarily generated via laboratory-based studies and automated instrumentation (11, 84, 85).

Domain knowledge. This information type comprises community-accepted or otherwise verified and validated (75) sources of biomedical knowledge relevant to a domain of interest. Examples of sources of such knowledge include published literature databases, public or private databases containing experimental results or reference standards for biomolecular or phenotypic measurements, and ontologies or terminologies that serve to formalize the taxonomic and semantic descriptions of a given domain (58, 59, 75). Collectively, these types of domain knowledge may be used to support multiple operations, including: *1*) hypothesis development, *2*) hypothesis testing, *3*) comparative analyses, or *4*) augmentation of experimental data sets with statistical or semantic annotations (46, 63, 75). Such information is primarily generated via the reporting and dissemination of results, models, and data sets from contributing basic science, clinical, and/or translational studies (46, 75).

Biological models and technologies. Such sources of knowledge consist primarily of: *1*) empirically validated system or subsystem level models that serve to define the mechanisms by which biomolecular and phenotypic processes and their markers/indicators interact as a network (11, 38, 78, 84) and *2*) novel technologies that enable the analysis of integrative data sets in light of such models. By their nature these tools

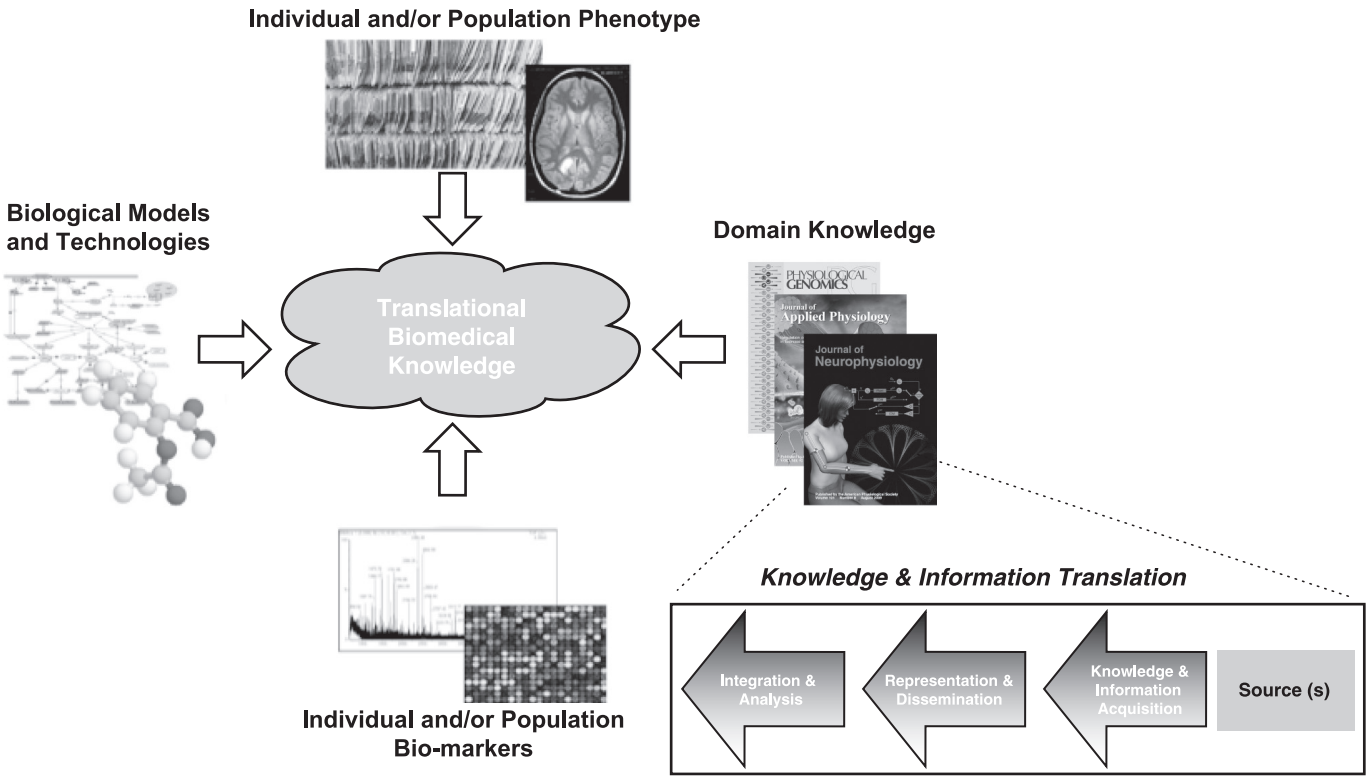

Fig. 3. Conceptual framework incorporating: *1*) major information types and needs involved in the creation of translational biomedical knowledge and *2*) a high-level overview of the 3-phase process by which knowledge and information from each such source is utilized to enable the generation of such knowledge.

include algorithmic or embedded knowledge sources (38, 78). These types of knowledge are primarily leveraged to reason upon or otherwise analyze the preceding three knowledge types (phenotypes, biomarkers, and domain knowledge). Such information is usually generated via the reporting and dissemination of results, models, and data sets from prior system-level modeling efforts, and their empirical verification and validation (19, 38, 84, 85).

Translational biomedical knowledge. Translational biomedical knowledge represents a subtype of general biomedical knowledge that is concerned with a systems-level synthesis (i.e., incorporate quantitative, qualitative, and semantic annotations) of pathophysiological or biophysical processes or functions of interest (e.g., pharmacokinetics, pharmacodynamics, bionutrition, etc.), and the markers or other indicators that can be used to instrument and evaluate such models. This type of knowledge is most commonly used to inform novel diagnostic, therapeutic, or population-level interventions or measurements that share a common goal of decreasing disease and/or increasing quality of life (38, 78).

Given the preceding definitions, we can define Translational Informatics as the informatics subdiscipline that is primarily concerned with the development and application of biomedical informatics theories, methods, and best practices intended to support: *1*) the acquisition of knowledge and information from the preceding sources; *2*) the representation of such knowledge and information in an actionable format (e.g., readily consumed and analyzed, usually through the use of computational tools and applications), and the subsequent dissemination of that knowledge or information to targeted end-users or analytical platforms; and *3*) the semantic integration of disparate data sources to support the discovery and verification/validation of complex bio-marker-to-phenotype relationships that may collectively define a translational biomedical knowledge model.

Our prototypical framework (as illustrated in Fig. 3) postulates that the four knowledge types defined previously can serve to categorize the constituent information needs and analytical requirements of the translational research cycle. Furthermore, the role of Biomedical Informatics, and more specifically Translational Informatics, in this framework is to address the four major information management challenges enumerated earlier to generate Translational Biomedical Knowledge, namely: *1*) the collection and management of high-throughput, multidimensional phenotypic and biomolecular data; *2*) the generation and testing of knowledge-anchored hypotheses relative to such data sets; *3*) the provision of reproducible and extensible data analytic pipelines; and *4*) the dissemination of knowledge and evidence generated by such translational research activities. Using this framework, one can evaluate the information needs of specific translational studies and plan for and address such needs in a manner consistent with a broader context of information and knowledge generation, integration, analysis, and dissemination spanning the complete translational research spectrum.

DESIGN AND EXECUTION OF TRANSLATIONAL INFORMATICS PROJECTS

Building upon the conceptual framework introduced in the preceding section, we are then able to present a practical model

for the design and execution of translational informatics projects, broadly informed by the prevailing methods and best practices being used in the National Cancer Institute (NCI)-sponsored Cancer Biomedical Informatics Grid (caBIG) initiative, as well as the National Center for Research Resources-sponsored CTSA consortium (9, 79, 82–85).

Translational Informatics Project Phases

The design and execution of a translational informatics project can be broadly divided into four major phases of an overall translational informatics cycle, as described below and illustrated in Fig. 4. For each such phase, a number of critical inputs and outputs are required or generated, and we will provide exemplary cases of such components for each phase.

Stakeholder engagement and knowledge acquisition. During this initial phase of a project, the key research and operational stakeholders who will be involved in the collection, management, analysis, and dissemination of project-specific data and knowledge are identified and engaged in a process of formal and informal knowledge acquisition, with the ultimate goal of defining the essential workflows, processes, and data sources (including their semantics) that will be involved in addressing a hypothesis or set of hypotheses. Such engagement and knowledge acquisition usually involve the use of ethnographic, cognitive science, workflow modeling, and conceptual knowledge acquisition techniques (59). The results of such efforts are usually recorded using a set of qualitative or thematic narratives (21, 31, 56) and formalized workflow or process artifacts (36, 37, 56). Major challenges that are encountered at this stage include the identification of appropriate stakeholders and the provision of incentives to encourage their engagement (4, 40), the ability of such stakeholders to adequately articulate their requirements and existing workflows/processes and/or data resources (53–55), and the expertise of project staff as it relates to the ability to execute such knowledge gathering and representation activities (59). In some cases, it is necessary to engage domain-specific subject matter experts who are not directly involved in a given project to augment available stakeholder generated knowledge or to validate the artifacts and knowledge generated during this phase (56, 59).

Data identification and modeling. Building upon the artifacts and knowledge generated in the prior phase, the next step in the design and execution of a translational informatics project is the identification of specific, pertinent data sources and the creation of models that encapsulate the physical and semantic representation of such data. In most if not all cases, the identification of existing or new data sources in informed by the prior phase (stakeholder engagement and knowledge

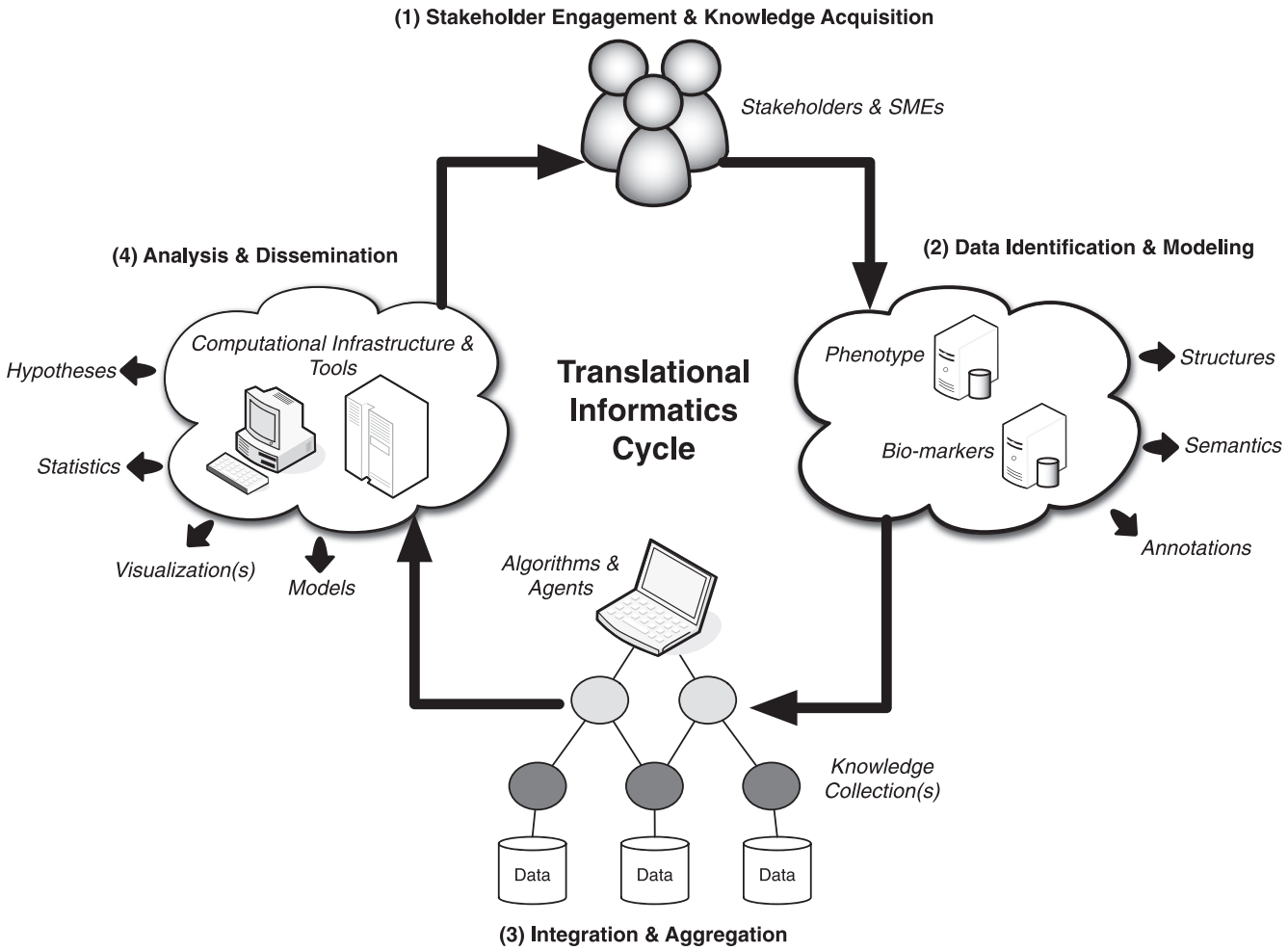

Fig. 4. Practical model for the design and execution of translational informatics projects, illustrating major phases and exemplary input or output resources and data sets.

acquisition). However, in some cases, additional research must be conducted to identify local or externally located data sources pertinent to a project's aims or hypotheses. Once such data sources have been identified and catalogued, it is then necessary to model their contents in an implementation-independent manner, for example using a model-driven architecture approach (1, 61, 68, 71, 74). Such approaches generate data structure artifacts, often using standards such as those found in the Unified Modeling Language (5, 24, 26, 76). In addition, during this phase, the semantic or domain-specific annotation of such data structures, using locally relevant and/or standardized terminologies and ontologies, provides even greater conceptual depth to the resulting knowledge or data models (26, 39, 41). Furthermore, these annotations can be later leveraged to support the interoperability of a diverse set of data sources (41) and to support the automated generation of hypotheses relative to a given data set and its semantics (57).

Integration and aggregation. As has been described in preceding sections, the primary objective of translational studies is to generate and analyze integrative biomedical knowledge that spans multiple levels of granularity (from biomolecules to tissues and systems) and sources (e.g., public databases, research databases, clinical systems, high-throughput instrumentation, etc.). Central to the ability to perform such analyses is the ability to integrate and aggregate such disparate data and knowledge sources. A common theme throughout contemporary approaches to this problem is the use of technology-agnostic domain or data models, incorporating semantic annotations, to perform either federated queries (15) or to transform data and load it into a shared data repository or warehouse (3, 7, 23). Current technologies and tools being used to support such operations include the caGrid data-sharing framework and associated tools/applications (in the case of federated queries) or the i2b2 data-warehousing platform (in the case of the transformation and loading of data into a shared repository) (18, 33). Given this premise, during this specific phase of the translational informatics cycle, the artifacts, data structures, data models, and semantic annotations generated in prior phases, are used to create such a data "mash up" (6, 16, 17, 49, 65, 66, 81). Such mash ups are often created using a variety of freely available and open-source reasoners and semantic web technologies (6, 16, 17, 49, 65, 66, 81).

Analysis and dissemination. In this fourth phase of the translational informatics cycle, the integrated data set or sets generated in the preceding phase can be subject to analysis and dissemination. Often, owing to the large and multidimensional nature of such data, it is necessary to use highly specialized data analysis platforms, such as the geWorkbench application that has been developed as part of the NCI's caBIG initiative to support biomolecular expression profiling and motif detection (12) or pattern detection and cluster analysis packages implemented within prevailing statistical packages such as R or SAS, to identify potentially novel or informative bio-marker-to-phenotype correlations (47, 73, 80). Of course, these are but a few examples of a rapidly growing constellation of such tools being generated by public and private sector contributors. In the majority of instances, these types of analytical tools are applied to address questions pertaining to one or more of the following four generalizable patterns: *1*) to generate testable hypotheses concerning relationships or patterns that may define important correlations between phenotypic and biomolecular markers (which can be derived from multiple measurement or detection modalities, including laboratory instrumentation, clinical data sets, and imaging) (57); *2*) to evaluate the validity of such hypotheses and the strength of their constituent relationships or patterns of interest using community-accepted statistical methodologies (47, 80); *3*) to visualize the interrelationships between and physical or quantitative distribution of variables or entities of interest within a given data set to leverage inherent human cognitive strengths in the areas of pattern detection and matching, thus augmenting naïve or partially guided computational pattern detection techniques (2, 22, 27); and *4*) to impute or otherwise extract reproducible and computationally tractable models capable of representing the phenomena of interest identified via the preceding interrogative pattern detection techniques (51, 77). The data and knowledge generated during this phase can then be appropriately validated (often employing a combination of quantitative and qualitative methods) and disseminated via multiple mechanisms, including the release of raw or synthesized data sets or knowledge collections to public repositories and/or the generation of publications and other human-accessible variants of experimental outcomes (50, 69, 70). It is important to note that these disseminated data and knowledge sources can in turn inform domain-appropriate stakeholders and catalyze their engagement in the initial phase of the cycle, thus creating a feedback loop (25, 58).

Challenges and Opportunities Associated with the Translational Informatics Cycle

Investigators and thought-leaders in the fields of clinical and translational science and biomedical informatics will quickly realize that the preceding description of the translational informatics cycle and project planning/execution process is highly idealized, a position shared by the authors. Any number of challenges exist within a given setting or project that may affect the efficacy and tractability of the methods and patterns we have introduced. However, such challenges are exceptionally context-specific and thus not generalizable. That being said, based upon our review of the literature and our own experiences, we believe that such impediments can be broadly associated with one or more of the following three factors:

1) The inappropriate designation of purely basic science or clinical research projects and aims as being translational, leading to a lack of consistency between project aims and methods and the generalized cycle and patterns introduced previously. In reality, it is a minority of projects that are translational in nature (20, 28, 43).

2) A lack of appropriate or necessary engagement of all key stakeholders, including individuals with the necessary skills and expertise, in a truly team-science construct that is mobilized and incentivized to address a targeted translational research problem (4, 72).

3) The absence of sufficiently robust data sets, knowledge sources, or analytical tools capable of addressing the requirements of a specific research question, hypothesis, or aim (10, 30, 35, 44, 45, 60).

We believe that these challenges, instead of being obstacles, represent an opportunity for the clinical and translational science and biomedical informatics communities to mobilize themselves to address core requirements in the

areas of clinical and translational research informatics training, methodology development, and application design, as well as to realign institutional and broader incentivizing structures and funding models to overcome such challenges. While a full description of such activities and their objectives is beyond the scope of this specific report, several recent reviews and position papers have argued the preceding points convincingly (10, 25, 35, 44, 58).

DISCUSSION

The holistic and systematic approach to both a conceptual and practical understanding of translational informatics and its role in supporting clinical or translational research, as presented in the preceding sections, is critical to both: *1*) allow informaticians to appropriately plan for, deploy, and utilize comprehensive and integrative informatics platforms capable of meeting the complex information management needs of the translational research domain; and *2*) allow clinical or translational researchers to better understand the broader context of their information needs and the available informatics capabilities, thus allowing them to serve as active participants in the development and use of such platforms. Providing such a common basis for understanding this domain is particularly important when we consider critical characteristics of the interactions between informaticians and clinical/translational scientists, which can directly affect project outcomes and success. For example, when clinical or translational researchers serve in a purely consumer role with respect to informatics tools due to a lack of familiarity with the domain of translational informatics, there is an associated and undesirable lack of input and guidance in terms of their development and verification/validation. Similarly, when informaticians serve in a purely technical role with respect to the design and implementation of research platforms, without understanding and leveraging a contextual basis for their utilization, less than optimal outcomes are usually realized. It is our position that a bidirectional collaboration between translational informaticians and clinical/translational researchers, informed by a shared conceptual and practical understanding of translational informatics, is both highly necessary and desirable relative to the ultimate goals of translational research and its ultimate impact on human health and wellness.

However, when one is considering the frameworks, models, and other findings presented in this article, in support of the preceding position statement, a number of critical limitations should be considered: *1*) this report does not constitute a self-contained and comprehensive literature review and relies on the domain coverage and knowledge content of a set of contributing reviews as we have previously cited; *2*) the conceptual framework and practical model set forth in the preceding sections are informed by the experiences and contributing domain-specific literature reviews of the authors, thus introducing the potential for bias or incomplete support for our positions; and *3*) a systematic evaluation of the efficacy and validity of our framework and model has not yet been performed. However, even with such limitations, we believe that a preponderance of peer-reviewed and anecdotal evidence supports the generalizable nature of our findings and assertions and that further systematic evaluation of our framework and model, while beyond the scope of this specific report, will only serve to further refine and enhance the basic concepts and methodological approaches presented in this study. Furthermore, we believe that a community dialogue concerning such issues is critical to the advancement of our collective understanding of such critical issues. The authors suggest that readers interested in conducting such a dialogue consult the online forums provides by many scientific organizations, such as the American Association for the Advancement of Science (http://www.aaas.org), American Medical Informatics Association (http://www.amia.org), or researchinformatics.org (http://www.researchinformatics.org).

CONCLUSIONS

We have presented a broad overview of the critical information challenges that exist in the clinical and translational research domain and a framework for translational informatics that can serve to provide greater context and a conceptual understanding of the role and practice of translational informatics relative to those challenges. We have also provided a practical model for the planning and execution of translational informatics projects, building upon the preceding conceptual model. Placing this framework and model in the context of our review of the current state of translational informatics knowledge and practice that can be synthesized from the published literature, we believe a number of additional measures are needed at the local, regional, and national levels to fully realize the benefits of employing such a systematic and multidisciplinary approach to translational informatics, specifically:

1) The establishment of incentives, extramural funding models, organizational forums, and career development pathways that can catalyze the creation of multidisciplinary translational research teams including informaticians, basic scientists, and clinical researchers;

2) The creation of knowledge and information exchange media for all members of the clinical and translational research community, with particular emphasis on increasing understanding of informatics capabilities and best practices in a manner consistent with a broad variety of expertise levels;

3) The provision of institutional fiscal support mechanisms for shared translational research infrastructure and platforms that are too costly or difficult to deploy at the individual investigator or laboratory level; and

4) The execution of studies and the generation of reports relative to translational studies that generate novel evidence or knowledge and that were made possible through the use of advance informatics tools and techniques.

In reporting upon the preceding framework, models, and recommendations, we intend to provide both translational researchers and biomedical informaticians with a mechanism of analyzing and addressing these challenges in a collaborative and constructive manner that can ultimately advance and support the conduct of high-throughput, translational research paradigms.

GRANTS

Dr. Payne's efforts in the preparation of this manuscript were supported in part by NIH Grants P01-CA-081534, R01-CA-134232, and UL1-RR-025755. Dr. Embi's efforts in the preparation of this manuscript were supported in part by NIH Grants R01-LM-009533 and UL1-RR-026314. Dr. Sen's efforts in the preparation of this manuscript were supported in part by NIH Grants UL1-RR-025755, R01-GM-077185, R01-GM-069589, and R01-HL-073087.

REFERENCES

1. **Aksit M, Kurtev I.** Elsevier special issue on foundations and applications of model driven architecture. *Science Computer Programming* 73: 1–2, 2008.
2. **Ardekani AM, Akhondi MM, Sadeghi MR.** Application of genomic and proteomic technologies to early detection of cancer. *Archives Iranian Medicine* 11: 427–434, 2008.
3. **Ariyachandra T, Watson HJ.** Which data warehouse architecture is best? *Commun ACM* 51: 146–147, 2008.
4. **Ash JS, Anderson NR, and Tarczy-Hornoch P.** People and organizational issues in research systems implementation. *J Am Med Inform Assoc* 15: 283–289, 2008.
5. **Batra D.** Unified modeling language (UML) topics: the past, the problems, and the prospects. *J Database Management* 19: I–VII, 2008.
6. **Belleau F, Nolin MA, Tourigny N, Rigault P, Morissette J.** Bio2RDF: Towards a mashup to build bioinformatics knowledge systems. *J Biomed Inform* 41: 706–716, 2008.
7. **Braa J.** A data warehouse approach can manage multiple data sets. *Bull World Health Organ* 83: 638–639, 2005.
8. **Brandt CA, Argraves S, Money R, Ananth G, Trocky NM, Nadkarni PM.** Informatics tools to improve clinical research study implementation. *Contemp Clin Trials* 27: 112–122, 2006.
9. **Buetow KH, Niederhuber J.** Infrastructure for a learning health care system: CaBIG. *Health Aff (Millwood)* 28: 923–924, 2009.
10. **Butler D.** Translational research: crossing the valley of death. *Nature* 453: 840–842, 2008.
11. **Butte AJ.** Medicine. The ultimate model organism. *Science* 320: 325–327, 2008.
12. **Califano A.** geWorkbench National Cancer Institute. https://cabig.nci.nih.gov/tools/geWorkbench (July 29, 2009).
13. **Campbell P.** Nature Glossary Nature Publishing Group. www.nature.com (November 15, 2008).
14. **Casey K, Elwell K, Friedman J, Gibbons D, Goggin M, Leshan T, Stoddard R, Tate D, Vick P, Vincent J.** A Broken Pipeline: Flat Funding of the NIH puts a Generation of Science at Risk. Bethesda, MD: National Institutes of Health, 2008; http://www.brokenpipeline.org/.
15. **Chakravarthy S, Whang WK, Navathe SB.** A logic-based approach to query-processing in federated databases. *Information Sciences* 79: 1–28, 1994.
16. **Cheung KH, Kashyap V, Luciano JS, Chen HJ, Wang YM, Stephens S.** Semantic mashup of biomedical data. *J Biomed Inform* 41: 683–686, 2008.
17. **Cheung KH, Yip KY, Townsend JP, Scotch M.** HCLS 2.0/3.0: health care and life sciences data mashup using Web 2.0/30. *J Biomed Inform* 41: 694–705, 2008.
18. **Childs LC, Enelow R, Simonsen L, Heintzelman NH, Kowalski KM, Taylor RJ.** Description of a rule-based system for the i2b2 challenge in natural language processing for clinical data. *J Am Med Inform Assoc* 16: 571–575, 2009.
19. **Chung TK, Kukafka R, Johnson SB.** Reengineering clinical research with informatics. *J Investig Med* 54: 327–333, 2006.
20. **Contopoulos-Ioannidis DG, Alexiou GA, Gouvias TC, Ioannidis JP.** Medicine. Life cycle of translational research for medical interventions. *Science* 321: 1298–1299, 2008.
21. **Crabtree BF, Miller WL.** *Doing Qualitative Research*. Newbury Park, CA: Sage, 1992.
22. **De Fonzo V, Aluffi-Pentini F, Parisi V.** Hidden Markov models in bioinformatics. *Curr Bioinform* 2: 49–61, 2007.
23. **DeWitt JG, Hampton PM.** Development of a data warehouse at an academic health system: Knowing a place for the first time. *Acad Med* 80: 1019–1025, 2005.
24. **Dobing B, Parsons J.** Dimensions of UML diagram use: a survey of practitioners. *J Database Management* 19: 1–18, 2008.
25. **Embi PJ, Payne PR.** Clinical research informatics: challenges, opportunities and definition for an emerging domain. *J Am Med Inform Assoc* 16: 316–327, 2009.
26. **Erickson J.** A decade and more of UML: An overview of UML semantic and structural issues and UML field use. *J Database Management* 19: I–VII, 2008.
27. **Feng J, Naiman DQ, Cooper B.** Probability-based pattern recognition and statistical framework for randomization: modeling tandem mass spectrum/peptide sequence false match frequencies. *Bioinformatics* 23: 2210–2217, 2007.
28. **Finkelstein JB.** Translational research going mainstream. *J Natl Cancer Inst* 100: 1430–1431, 2008.
29. **Fridsma DB, Evans J, Hastak S, Mead CN.** The BRIDG project: a technical report. *J Am Med Inform Assoc* 15: 130–137, 2008.
30. **Funder JW.** Translational research goes both ways: lessons from clinical studies. *Clin Exp Pharmacol Physiol* 35: 526–529, 2008.
31. **Glaser B, Strauss A.** *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Piscataway, NJ: Aldine Transaction, 1967, p. 271.
32. **Haines A, Jones R.** Implementing findings of research. *BMJ* 308: 1488–1492, 1994.
33. **Heinze DT, Morsch ML, Potter BC, Sheffer RE.** Medical i2b2 NLP smoking challenge: the A-life system architecture and methodology. *J Am Med Inform Assoc* 15: 40–43, 2008.
34. **Kaiser J.** US budget 2009. NIH hopes for more mileage from roadmap. *Science* 319: 716, 2008.
35. **Keramaris NC, Kanakaris NK, Tzioupis C, Kontakis G, Giannoudis PV.** Translational research: from benchside to bedside. *Injury* 39: 643–650, 2008.
36. **Khan SA, Kukafka R, Payne PR, Bigger JT, Johnson SB.** A day in the life of a clinical research coordinator: observations from community practice settings. *Medinfo* 12: 247–251, 2007.
37. **Khan SA, Payne PR, Johnson SB, Bigger JT, Kukafka R.** Modeling clinical trials workflow in community practice settings. *AMIA Annu Symp Proc*: 419–423, 2006.
38. **Knaup P, Ammenwerth E, Brandner R, Brigl B, Fischer G, Garde S, Lang E, Pilgram R, Ruderich F, Singer R, Wolff AC, Haux R, Kulikowski C.** Towards clinical bioinformatics: advancing genomic medicine with informatics methods and tools. *Methods Inf Med* 43: 302–307, 2004.
39. **Komatsoulis GA, Warzel DB, Hartel FW, Shanbhag K, Chilukuri R, Fragoso G, de Coronado S, Reeves DM, Hadfield JB, Ludet C, Covitz PA.** caCORE version 3: Implementation of a model driven, service-oriented architecture for semantic interoperability. *J Biomed Inform* 41: 106–123, 2008.
40. **Kukafka R, Johnson SB, Linfante A, Allegrante JP.** Grounding a new information technology implementation framework in behavioral science: a systematic analysis of the literature on IT use. *J Biomed Inform* 36: 218–227, 2003.
41. **Kunz I, Lin MC, Frey L.** Metadata mapping and reuse in caBIG. *BMC Bioinformatics* 10, *Suppl* 2: S4, 2009.
42. **Kush RD, Helton E, Rockhold FW, Hardison CD.** Electronic health records, medical research, and the Tower of Babel. *N Engl J Med* 358: 1738–1740, 2008.
43. **Lean ME, Mann JI, Hoek JA, Elliot RM, Schofield G.** Translational research. *BMJ* 337: a863, 2008.
44. **Ledford H.** Translational research: the full cycle. *Nature* 453: 843–845, 2008.
45. **Lehmann CU, Altuwaijri MM, Li YC, Ball MJ, Haux R.** Translational research in medical informatics or from theory to practice. A call for an applied informatics journal. *Methods Inf Med* 47: 1–3, 2008.
46. **Levy D, Dondero R, Veronneau P.** Research Rewired: Merging Care and Research Information to Improve Knowledge Discovery. New York: Price Waterhouse Coopers, 2008.
47. **Mansmann U.** Genomic profiling. Interplay between clinical epidemiology, Bioinformatics and biostatistics. *Methods Inf Med* 44: 454–460, 2005.
48. **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* 45: 180–185, 2006.
49. **Marks P.** 'Mashup' websites are a dream come true for hackers. *New Scientist* 190: 28–29, 2006.
50. **Moja LP, Moschetti I, Nurbhai M, Compagnoni A, Liberati A, Grimshaw JM, Chan AW, Dickersin K, Krleza-Jeric K, Moher D, Sim I, Volmink J.** Compliance of clinical trial registries with the World Health Organization minimum data set: a survey. *Trials* 10: 56, 2009.
51. **Oehmen CS, Straatsma TP, Anderson GA, Orr G, Webb-Robertson BJM, Taylor RC, Mooney RW, Baxter DJ, Jones DR, Dixon DA.** New challenges facing integrative biological science in the post-genomic era. *J Biol Sys* 14: 275–293, 2006.
52. **Oster S, Langella S, Hastings S, Ervin D, Madduri R, Phillips J, Kurc T, Siebenlist F, Covitz P, Shanbhag K, Foster I, Saltz J.** caGrid 1.0: an enterprise Grid infrastructure for biomedical research. *J Am Med Inform Assoc* 15: 138–149, 2008.

53. **Patel VL, Arocha JF, Kaufman DR.** A primer on aspects of cognition for medical informatics. *J Am Med Inform Assoc* 8: 324–343, 2001.
54. **Patel VL, Glaser R, Arocha JF.** Cognition and expertise: acquisition of medical competence. *Clin Invest Med* 23: 256–260, 2000.
55. **Patel VL, Kaufman DR, Arocha JF.** Emerging paradigms of cognition in medical decision-making. *J Biomed Inform* 35: 52–75, 2002.
56. **Patton MQ.** *Qualitative Research & Evaluation Methods.* New York: Sage Publications, 2001, p. 688.
57. **Payne PR, Borlawsky T, Kwok A, Greaves A.** Supporting the design of translational clinical studies through the generation and verification of conceptual knowledge-anchored hypotheses. *AMIA Annu Symp Proc*: 566–570, 2008.
58. **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* 53: 192–200, 2005.
59. **Payne PR, Mendonca EA, Johnson SB, Starren JB.** Conceptual knowledge acquisition in biomedicine: a methodological review. *J Biomed Inform* 40: 582–602, 2007.
60. **Pearson H.** Translational research: a case history. *Nature* 453: 846–849, 2008.
61. **Rayhupathi W, Umar A.** Exploring a model-driven architecture (MDA) approach to health care information systems development. *Int J Med Inform* 77: 305–314, 2008.
62. **Research NDsPoC.** NIH Director's Panel on Clinical Research Report Bethesda, MD: National Institutes of Health, 1997.
63. **Richesson RL, Krischer J.** Data standards in clinical research: gaps, overlaps, challenges and future directions. *J Am Med Inform Assoc* 14: 687–696, 2007.
64. **Ruttenberg A, Clark T, Bug W, Samwald M, Bodenreider O, Chen H, Doherty D, Forsberg K, Gao Y, Kashyap V, Kinoshita J, Luciano J, Marshall MS, Ogbuji C, Rees J, Stephens S, Wong GT, Wu E, Zaccagnini D, Hongsermeier T, Neumann E, Herman I, Cheung KH.** Advancing translational research with the Semantic Web. *BMC Bioinformatics* 8, *Suppl* 3: S2, 2007.
65. **Sahoo SS, Bodenreider O, Rutter JL, Skinner KJ, Sheth AP.** An ontology-driven semantic mashup of gene and biological pathway information: application to the domain of nicotine dependence. *J Biomed Inform* 41: 752–765, 2008.
66. **Scotch M, Yip KY, Cheung KH.** Development of grid-like applications for public health using web 2.0 mashup techniques. *J Am Med Inform Assoc* 15: 783–786, 2008.
67. **Shortliffe EH, Cimino JJ.** *Biomedical Informatics: Computer Applications in Health Care and Biomedicine*. New York: Springer, 2006, p. 1037.
68. **Shurville S.** Model driven architecture and ontology development. *Interactive Learning Environments* 15: 96–99, 2007.
69. **Sim I.** Trial registration for public trust: making the case for medical devices. *J Gen Intern Med* 23, *Suppl* 1: 64–68, 2008.
70. **Sim I, Chute CG, Lehmann H, Nagarajan R, Nahm M, Scheuermann RH.** Keeping raw data in context. *Science* 323: 713, 2009.
71. **Soley RM.** Model driven architecture: the evolution of object-oriented systems? *Object-Oriented Information Systems* 2817: 2–2, 2003.
72. **Sung NS, Crowley WF Jr, Genel M, Salber P, Sandy L, Sherwood LM, Johnson SB, Catanese V, Tilson H, Getz K, Larson EL, Scheinberg D, Reece EA, Slavkin H, Dobs A, Grebb J, Martinez RA, Korn A, Rimoin D.** Central challenges facing the national clinical research enterprise. *JAMA* 289: 1278–1287, 2003.
73. **Tiwari A, Sekhar AKT.** Workflow based framework for life science informatics. *Comput Biol Chem* 31: 305–319, 2007.
74. **Uhl A.** Model driven architecture is ready for prime time. *Ieee Software* 20: 70–73, 2003.
75. **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* 75: 257–267, 2006.
76. **Vanderperren Y, Mueller W, Dehaene W.** UML for electronic systems design: a comprehensive overview. *Design Automation for Embedded Systems* 12: 261–292, 2008.
77. **Way JC, Silver PA.** Systems engineering without an engineer: why we need systems biology. *Complexity* 13: 22–29, 2007.
78. **Webb CP, Pass HI.** Translation research: from accurate diagnosis to appropriate treatment. *J Transl Med* 2: 35, 2004.
79. **Wolfson W.** caBIG: Seeking cancer cures by bits and bytes. *Chem Biol* 15: 521–522, 2008.
80. **Xu R, Wunsch D.** Survey of clustering algorithms. *Ieee Transactions on Neural Networks* 16: 645–678, 2005.
81. **Yu J, Benatallah B, Casati F, Daniel F.** Understanding mashup development. *Ieee Internet Computing* 12: 44–52, 2008.
82. **Zerhouni E.** Medicine. The NIH Roadmap. *Science* 302: 63–72, 2003.
83. **Zerhouni EA.** Clinical research at a crossroads: the NIH roadmap. *J Investig Med* 54: 171–173, 2006.
84. **Zerhouni EA.** Translational and clinical science–time for a new vision. *N Engl J Med* 353: 1621–1623, 2005.
85. **Zerhouni EA.** US biomedical research: basic, translational, and clinical sciences. *JAMA* 294: 1352–1358, 2005.