Clinical research informatics: a conceptual perspective

Michael G Kahn,¹ Chunhua Weng²

ABSTRACT
Clinical research informatics is the rapidly evolving sub-discipline within biomedical informatics that focuses on developing new informatics theories, tools, and solutions to accelerate the full translational continuum: basic research to clinical trials (T1), clinical trials to academic health center practice (T2), diffusion and implementation to community practice (T3), and ‘real world’ outcomes (T4). Two recent factors accelerating the translation of clinical research work have been (1) the extensive and diverse informatics needs of the NIH Clinical and Translational Sciences Awards (CTSAs),¹⁰⁻¹⁶ and (2) the growing interest in electronic health records (EHRs) to notify patients of medication safety alerting, which leverages the growing use of surveillance systems.¹⁷⁻¹⁹

This 2012 CRI supplement adds 18 new publications to this growing field.

A CONCEPTUAL MODEL OF CLINICAL RESEARCH INFORMATICS
To provide guidance on the CRI innovations represented in this special supplement, we developed the conceptual model in figure 1. This figure illustrates how CRI integrates clinical and translational research workflows in addition to core informatics methodologies and principles into a framework that reflects the unique informatics needs of translational investigators. The model is organized around three conceptual components: workflows, data sources and platforms, and informatics core methods and topics.

The central structure that establishes the unique context for CRI is the informatics-enabled clinical research workflow. The elements and sequence of this workflow should be familiar as it reflects the key phases in the scientific model of knowledge discovery.¹¹⁻¹⁹ Unlike diagrams that appear in traditional research methodology textbooks, figure 1 applies an informatics-centric perspective to each step and contains two translational workflow cycles, which reflect the use of CRI technologies in both early (T1–T2) and later (T3–T4) translational phases.¹¹⁻¹² The ‘outer’ cycle represents translational discoveries within carefully controlled study conditions in a limited number of clinical trial sites. The ‘outer’ cycle represents the later stages of clinical translational research, where implementation and dissemination tasks become more prominent across community practices. The later stages of clinical translational research are represented by implementation-oriented translational activities such as evidence generation and synthesis, personalized evidence application, and population surveillance.

New scientific knowledge, both hypothesis-generating and hypothesis-testing, begins with a research question that drives the investigative process. While previous studies may suggest possible new research questions, ultimately this step reflects the creative insight of a well-trained translational investigator. During the early planning phases, study feasibility assessment and cohort identification are important tasks for ensuring that sufficient study participants and data exist to move the proposed study forward. Eligibility alerting, which leverages the growing use of electronic health records (EHRs) to notify physicians of their patients’ eligibility for clinical trials, is one of the major informatics solutions to address the leading cause of failures in clinical studies—the inability to recruit sufficient study...
Figure 1 A conceptual model for clinical research informatics consisting of an informatics-enabled clinical research workflow, heterogeneous data sources, and a collection of informatics methods and platforms. EHR, electronic health records; IDR, integrated data repositories; PHR, personal health records.

...participants. Obtaining informed consent is a critical step in clinical research recruitment. Advanced interactive human—computer educational systems could reduce the burden for investigators and improve the understanding of risks and benefits by patients. Data collection and analyses follow naturally after patients are enrolled, but are often seen (erroneously) as the sole use of informatics by most investigators. As shown in figure 1, CRI supports the cycle for converting data into knowledge by encompassing data analysis, evidence generation, and evidence synthesis. Population surveillance seeks to discover unmet community-based health needs, which can be used to drive another set of research questions.

Reflecting the expanding scope of data sources that are commonly used to drive clinical and translational research, figure 1 highlights CRI’s emphasis on data integration across EHRs or over time to form integrated longitudinal data repositories, which in turn are integrated across institutions to form multi-institutional federated data networks. A wide range of additional sources of data is reflected in figure 1: personal health records, registries, claims databases, public reports, and social media that contain patient self-reported outcome data. This list is intentionally incomplete—it is intended only to highlight the endless variety of both ‘traditional’ and ‘non-traditional’ data sources, such as in-home continuous monitoring, public and specialized social networks, and geo-location data. Significant CRI research has focused on the challenges of data integration across disparate data sources that may differ in concept specificity (granularity), representation, syntax, and semantics.

Similarly, a large body of informatics research has developed alternative models for data federation across independent data sources, including distributed, federated, and mediator-based architectures. Two of the largest efforts to develop large-scale data integration and distributed data sharing environments specifically directed toward clinical and translational research are caBIG from the National Cancer Institute and BIRN from the National Center for Research Resources (now part of the National Center for Advancing Translational Sciences). Some CRI investigators are adopting and adapting these architectures to meet the needs of multi-institutional data sharing networks.

The need to support the above informatics-enabled clinical research workflows and to strengthen the national research capacity have led to new developments in CRI core topics and techniques. Many technologies used to solve CRI needs have been borrowed from other informatics disciplines and adapted to meet CRI requirements. The bottom portion in figure 1 highlights the major core research topics in CRI, including secondary use of clinical data for research, distributed queries, data integration, record linkage, data quality assessment, integrated data models and terminologies, and a set of common informatics methods, including human—computer interaction, knowledge management, NLP, information extraction, and text classification. Each core topic builds upon and extends fundamental informatics theories and methodologies that are implemented and assembled into functioning CRI solutions. This supplement contains 18 articles that focus on various aspects of CRI workflow, applications, or research topics. The articles contribute to either a CRI workflow task or an underlying core CRI technology or platform or both, as illustrated in figure 1.

NEW CONTRIBUTIONS TO THE CLINICAL RESEARCH INFORMATICS KNOWLEDGE BASE

Integrated clinical data repositories or federated data networks are considered a fundamental infrastructure for biomedical and translational research. With the establishment of the US national CTSA consortium, which currently consists of 60 participating institutions, there is a pressing need to develop and share best practices for clinical data integration in support of clinical research. MacKenzie et al (see page e119) conducted a survey among 28 CTSA and the NIH Clinical Center. This study identified several data integration trends among the CTSA programs, such as a growing presence of centralized integrated data repositories and master patient indexing tools. Another key finding is the increasing movement away from homegrown...
solutions to more broadly used integration platforms such as i2b2.13 41 137

Popular applications of integrated data repositories for clinical and translational research include retrospective data analyses and identification of research participants to improve clinical research recruitment,40 but few institutions have leveraged real-time streams to enrich data. Ferranti et al (see page e68) designed and implemented an open-source, data-driven cohort recruitment system called The Duke Integrated Subject Cohort (DISCERN).32 This system combines both retrospective warehouse data and real-time clinical events via Health Level Seven (HL7) messages to immediately alert study personnel of potential recruits as they become eligible. Real-time data feeds are critical when the required clinical findings have not yet been loaded into the warehouse but have been captured contemporaneously during patient care. The use of both retrospective and real time data provides an interesting example of how multiple data sources may be required to capture important details for cohort discovery.

Extending the capacity of a single institutional data repository to support translational studies, Anderson et al (see page e60) used the i2b2 data warehouse software to implement a multi-institutional federated data network for population-based cohort discovery.37 This infrastructure links de-identified data repositories from three CTSA institutions to support federated queries to identify potentially eligible patients for clinical trial studies. This distributed data-sharing network requires a harmonized common data model, value sets, and data access policies across all participating institutions. It demonstrates the ability for a distributed network containing de-identified patient data to provide aggregated patient counts. An important finding is that while multi-institutional cohort discovery allows for queries to interrogate extremely large patient populations, harmonization of inter-institutional policies, semantics, and use cases is perhaps more important and challenging than technical harmonization.

Motivated by a different use case but using a similar approach, Buck (see page e46) leveraged a widely adopted EHR system in New York City to develop a clinical and public health research platform. This research infrastructure participates in a city-wide distributed query network to support population-based data queries with provider-specific alerting and communication capabilities.35 This virtual network aggregates distributed count information and reports, and disseminates shared decision support alerts and secure messaging directly into provider EHR email accounts. This project illustrates how a common EHR system, with common documentation, codes, and standards, can be used to monitor community health and facilitate communications between clinical and public health practitioners.

Both of these articles highlight the importance of using standard software, data models, and data semantics to enable large-scale research infrastructures and to achieve interoperability across organizations.

Recruitment is the primary and most costly barrier to clinical and translational research.158 This supplement contains two articles that contribute to the literature on informatics solutions for boosting recruitment.20 139 Embi and Leonard (see page e145) evaluated the response patterns over time to EHR-based clinical trial alerts using a randomized clinical trial.139 The authors observed that responses to clinical trial alerts declined gradually over prolonged exposure. However, recruitment performance remained higher than baseline despite this decline in responsiveness to trial alerts over time. The authors found that, while there were no differences in the loss of performance between specialists and generalists, there was a significantly bigger loss of alert responsiveness in community-based practitioners compared to academic practitioners. This study is another reminder that one person’s critical alert is another person’s disruptive annoyance.

Obtaining informed consent remains a labor-intensive step in clinical research recruitment. The study from Tait et al (see page e43) proposed a novel interactive consent program that enables patients to specify their preferences to participate in pediatric clinical trials.20 The interactive computer program contains both child- and parent-appropriate animations of a clinical trial of asthma and shows that innovative technologies can open new possibilities for eliminating workflow barriers in translational research. The improved understanding of key clinical trial concepts by both children and adults indicates that this approach should be explored in more depth as more powerful hand-held tablet devices become widely available.

Besides the use of clinical data to facilitate clinical trial recruitment, broadened secondary use of clinical data has been on the rise. Secondary data use requirements have resulted in the development of new approaches to deriving actionable knowledge from the mass of patient data in structured fields, unstructured text, and handwritten notes.103 140 141 For example, adapting the results of large-scale clinical studies to individual patients remains challenging. Jiang et al (see page e137) investigated model adaptation challenges in risk prediction for individual patients and developed a patient-driven adaptive prediction technique (ADAPT) to improve personalized risk estimation for clinical decision support.140 This method selects the best risk estimation model from a set of models for an individual patient. The technique examines individualized confidence intervals based on an individual’s data to select the ‘best’ risk prediction. This very simple, computationally inexpensive approach shows better performance using receiver operating characteristic (ROC) and goodness-of-fit tests compared to alternative model-selection approaches.

Mathias, Gossett, and Baker (see page e96) describe a retrospective study using EHR data to estimate the incidence of inappropriate use of cervical cancer screening. Using manual chart review to validate the accuracy of their electronic query, they were able to determine that most low-risk women were receiving Pap tests more frequently than recommended. Of particular interest, Mathias provides the actual query logic used to identify study participants. Excluding the lines that generate data from the mass of patient data in structured fields, unstructured text, and handwritten notes.103 140 141 For example, adapting the results of large-scale clinical studies to individual patients remains challenging. Jiang et al (see page e137) investigated model adaptation challenges in risk prediction for individual patients and developed a patient-driven adaptive prediction technique (ADAPT) to improve personalized risk estimation for clinical decision support.140 This method selects the best risk estimation model from a set of models for an individual patient. The technique examines individualized confidence intervals based on an individual’s data to select the ‘best’ risk prediction. This very simple, computationally inexpensive approach shows better performance using receiver operating characteristic (ROC) and goodness-of-fit tests compared to alternative model-selection approaches.

Li and colleagues (see page e51) describe the use of seasonally adjusted alerting thresholds in a disease surveillance system to obtain improved outbreak detection performance during epidemic and non-epidemic seasons of hand-foot-and-mouth disease.103 Their conclusions indicate that, for diseases with known seasonal variability, different thresholds may be most appropriate for optimizing high sensitivity and low false alarm rates without reducing the time to outbreak detection.

A patient’s data is often scattered in data repositories from multiple organizations. Therefore, record linkage is a critical step in integrating data about patients obtained from different data sources. To address information fragmentation and incompleteness problems that are common to many data repository developers, Duvall and colleagues (see page e54) describe their
experience performing record linkage between a large institutional enterprise data warehouse and a statewide (Utah) population database. The results of record linkage were then validated using a state cancer registry. They developed a Master Subject Index, which has become an increasing popular method to identify the same person in multiple data sources to support linked data discovery. The project used a commercial record linkage tool based on probabilistic record matching. An analysis of their findings indicated the strong negative impact of missing values in fields used in the record linkage algorithm.

A common concern related to secondary use of clinical data is data quality. In this supplement, three articles present different methods for data quality assurance: the use of imputation; rule-based error detection; and knowledge-based approaches leveraging semantic web and UMLS' semantic network knowledge. Sariyar, Borg, and Pommerening (see page e76) focus on systematic approaches for dealing with missing values that occur in fields that are used to perform record linkage. Their ‘measure of success’ for alternative approaches is the accuracy of record linkage following the application of alternative methods. Using both real and simulated data and four alternative linkage scoring methods based on classification and regression trees (CART), they show that assuming that a missing value always represents a non-match is a computationally efficient heuristic with only a small loss in accuracy compared to alternative algorithms that are substantially more complex.

Rather than using imputation, McGarvey and colleagues (see page e125) describe a multi-faceted approach to improving data completeness and quality in a multi-center breast and colon cancer family registry.142 The authors implemented a rule-based validation system that facilitates error detection and correction for research data centers. Evaluation over a 2-year period showed a decrease in the numbers of errors per patient in the database and a concurrent increase in data consistency and accuracy. While their approach improved efficiency and operational effectiveness, an important finding is the need to establish data-quality governance that explicitly acknowledges the shared responsibilities between members of the data coordinating center and the data collection sites in improving the overall quality of research data. As additional data validation routines were implemented, their findings highlight the oft-stated observation that ‘you cannot improve what you do not measure.’

Common data elements (CDEs) have emerged as an effective way to represent reusable, semantically defined data collection items. Jiang et al (see page e129) evaluated the semantic consistency of CDE value sets contained in the NCI caDSR repository. This paper presents a new methodology for assessing the quality of value set terms using a clever mapping between CDEs and the UMLS semantic network’s 15 semantic groups and 133 semantic types.143 Elements in a value set were considered inconsistent if a member of the value set mapped to a different type or group in the UMLS semantic network. This effort highlights the critical need to constantly evaluate the very large body of CDEs to ensure that these elements, which are critical to future data sharing efforts, are themselves consistent and correct.

The previous articles focused on the reuse of structured data elements. Another common challenge to reusing clinical data for clinical research is to extract information from unstructured data sources, such as text and images. Therefore, various methods for NLP, text classification, information extraction, and optical character recognition (OCR) have been developed to address this challenge. This supplement includes three articles providing examples of the above methods.24 144 145

NLP has emerged as a critical technology in large-scale clinical research.146 Savova (see page e83) describes the use of NLP to extract drug treatment information from breast cancer therapy notes.145 Extracted information was combined with structured information from an electronic prescribing system and integrated into a common treatment timeline. This work shows how integration of information from both structured and unstructured data sources can result in data sets that are richer in content than can be provided by either data source alone. Although not a focus of this paper, it is striking to note that the NLP pipeline required 12 different computational processes to annotate the text, most of which are part of the OpenNLP toolset, and numerous public-domain coding systems.

Rasmussen et al (see page e90) extended conventional information extraction tasks from data fields or electronic text to scanned handwritten forms using an OCR processing pipeline.24 The proposed pipeline leverages the capabilities of existing third-party OCR engines and provides the flexibility offered by a modular system. Pipeline-based architectures are common in NLP solutions, as illustrated by the Savova article described previously. Rasmussen’s results show that the OCR pipeline significantly reduces human effort on chart abstraction. Rasmussen’s focus on OCR reminds us that an enormous body of historical medical information exists in handwritten text notes. Informatics tools that can eliminate or reduce manual chart abstraction would make these data more accessible for clinical research.

Many studies use manual chart reviews to classify patients. Manual methods are not just time-consuming: they are prone to classification bias. Using adverse event reports, Ong, Magrabi, and Coiera (see page e110) show how statistical classification methods can be used to classify extreme risk (Severity Code Assessment level one) reports with high accuracy. As seen in other uses of statistical classifiers, performance was better when the training set consisted of a narrow set of conditions (specifically, patient misidentification errors) rather than a diverse population of events.

An important resource for information retrieval in clinical data is the wide range of semantic knowledge resources such as UMLS and SNOMED-CT. Given the importance of data models and semantic knowledge for CRI, much work has been focused on improving the quality of these critical knowledge resources. López-García (see page e102) describes a usability-driven pruning technique to study the modularity of SNOMED-CT.147 This study concludes that graph-traversal strategies and frequency data from an authoritative source can prune large biomedical ontologies and produce useful segmentations that still exhibit acceptable coverage for annotating clinical data. Similarly, Wu et al (see page e149) investigate the frequency of UMLS terms in clinical notes across multiple institutions’ clinical data warehouses.148 The authors found that only 3.56% of UMLS terms were empirically attested in clinical notes, implying that a lightweight lexicon could be developed to improve the efficiency of NLP systems for clinical notes.

LOOKING FORWARD
From all the diversity of workflow applications, methods, and knowledge resources that we see represented in this special issue, we not only identify a steadily growing literature in classic CRI topics such as data integration or federation, information retrieval, and data analysis, but we also note some emerging new areas, such as interactive consenting and individualized decision support. We expect the CRI research agenda will continue to evolve to become more precise, predictive, preemptive, and
participatory, in parallel with the development of ‘4P medicine’.\textsuperscript{149} We anticipate more patient-centered research decision support and innovative consent programs to strengthen patient participation and participation, including specifying how an individual’s research data will be used and by whom.\textsuperscript{150} We also expect more CRI research that is informed by and responsive to Commissioned; internally peer reviewed.

Competing interests None.

Provenance and peer review Commissioned; internally peer reviewed.

REFERENCES

22. Payne PR, Borlawsky TB, Lele O, et al. The TOKE project: knowledge synthesis for translational Inquiries Network) and AHRQ R21 HS19726-01A. The contents are the authors’ sole responsibility and do not necessarily represent official NIH views. Dr Weng was support by grants R01LM009886 and R01LM010815 from the National Library of Medicine, grant UL1RR024156 from the National Center for Research Resources, and AHRQ grant R01HS018963.

Downloaded from http://jamia.oxfordjournals.org/ by guest on January 29, 2016


121. Clinical Trials News. Subject Recruitment by Far Biggest Clinical Trial Concern


