**FOCUS**

**POLICY**

Time to Integrate Clinical and Research Informatics

Irene L. Katzan*1 and Richard A. Rudick1,2

Integration of clinical and research informatics can streamline clinical research, patient care, and the building of a learning health care system.

The wide array of clinical data contained in the electronic health record (EHR) has tremendous potential to facilitate comparative effectiveness and outcomes research, which could in turn define effective treatment strategies and inform intelligent allocation of health care resources. But EHRs are dominated by unstructured narrative data that are not available for research or quality improvement efforts. Clinical research databases—which contain well-defined and structured data—are created independently, and clinically relevant data within these research databases are not available for purposes of clinical care. This separation creates inefficiencies in research and patient care as well as barriers to a learning health care system (1). A meaningfully integrated approach to clinical and research data informatics is needed to promote improved health outcomes and more rational allocation of health care resources. Here, we highlight bottlenecks in the integration pathway and modern approaches to defining solutions.

**THE PROBLEM**

The harnessing of clinical information in EHRs and other electronic clinical data sources has the potential to facilitate comparative effectiveness and clinical outcomes research (1, 2), which can then be used to define effective treatment strategies and inform intelligent allocation of scarce health care resources by relating incremental costs to expected outcomes. This so-called learning health system, as envisioned by the Institute of Medicine, is “designed to generate and apply the best evidence for the collaborative health care choices of each patient and provider; to drive the process of discovery as a natural outgrowth of patient care; and to ensure innovation, quality, safety, and value in health care” (3).

However, these laudable goals currently are beyond reach (1, 4). EHRs are primarily document management systems that optimize access to case-specific clinical information. As such, EHRs are designed to capture and retrieve unstructured, narrative data. The discrete data available in EHRs are generally related to billing requirements—diagnoses, procedure codes, or documentation of care—or laboratory test results. The ability to retrieve, manipulate, analyze, and report data within EHRs or to merge data across institutions is limited by the lack of standard data definitions or interoperativity.

Given the transformative potential of clinical data to generate new knowledge, health care systems should make solving these problems a high priority. Such an effort will require coordination at multiple levels, including national bodies that make policy; information technology experts who develop technological infrastructures that can speak to each other; researchers and statisticians who must use, and in some cases develop, appropriate analytical approaches for clinically derived data; and health care providers who will need to change their clinical documentation practices.

**STIMULATING CONVERSATION**

The lack of data standardization is one of the biggest barriers to research and clinical data integration because it prevents interoperability across systems. Several initiatives are in place to address this problem.

Enacted in 2009, the Health Information Technology for Economic and Clinical Health (HITECH) Act focuses on the adoption and meaningful use of health information technology (1). The Healthcare Information Technology Standards Panel has recently defined standards for data flow from EHRs to clinical research databases and registries. International consortia are developing processes for obtaining a multinational consensus on research standards and regulations. For example, the Clinical Data Interchange Standards Consortium—a multidisciplinary, nonprofit standards-developing organization—has been established to create worldwide data standards to streamline medical research and link with health care (5). The EHRs for Clinical Research project, funded by the European Innovative Medicines Initiative, is aimed at providing adaptable, reusable, and scalable data.

---

*Neurological Institute Center for Outcomes Research and Evaluation, Cleveland Clinic, Cleveland, OH 44195, USA. 1Mellen Center for Multiple Sclerosis Neurological Institute, Cleveland Clinic, Cleveland, OH 44195, USA. 2Corresponding author. E-mail: katzan@ccf.org

---

---

Fig. 1. Come together. Integration of clinical and research informatics can speed the building of a learning health care system.
solutions for reusing data from EHR systems for clinical research (6).

Unfortunately, these initiatives will take many years to mature, and results are unpredictable. Even after these initiatives have matured, health care institutions will still need a local solution for accessing clinically derived data for use in their own outcomes and effectiveness research programs.

EXPANDING ACCESS

Because of the immediate need, several approaches have emerged to make information in EHRs available for comparative effectiveness, clinical outcomes, and translational research. These approaches include EHR data repositories, institutional data warehouses, and scalable frameworks for storing and analyzing data. Many EHRs have developed capacity to store discrete data in large clinical data repositories, such as Epic’s Clarity (Epic Systems, Verona, Wisconsin). The framework of and data retrieval (pulling data out of a registry or a database) from these EHR mechanisms are complex, and use requires the help of specialists, although data extraction—pulling out less-structured data from a system and moving it to another database—is gradually becoming easier.

Institution-based data warehouses contain data from EHRs and other clinical systems such as billing or laboratory systems (7) that are accessible to authenticated users via an automated query tool. Deidentified data are usually immediately available, and identifiable data may be downloaded once institutional review board (IRB) approval and regulatory assurances are obtained. Because data that are accessible through an automated query tool are restricted in content and complexity, queries generally are limited to identifying patients for clinical research projects. The system architecture for these warehouses is often relational and uses internally derived data mappings, which limit scalability outside that institution.

The Knowledge Program (KP)—Cleveland Clinic’s institutional data warehouse—focuses on maximizing both the quantity and quality of EHR-based clinical data available for clinical research (8). Health-status data, patient-reported outcomes, and condition-specific data elements are entered by the patient and provider and stored discretely at each patient visit. Patient-reported data are collected using electronic tablets in the ambulatory clinics of disease-based institutes. Currently, measures of health status are collected on ~26,000 clinical encounters per month through this system. The systematic collection of standardized clinical information on all patients during the flow of patient care distinguishes the KP program from most other known clinical data initiatives. Processes are in place to obtain systematic follow-up information in select patient subgroups that will enhance data completeness and representativeness. Patient-reported data and clinical data from EHRs and other clinical systems are aggregated in the KP data warehouse and are available for review and download through a Web-based query tool. The KP data warehouse uses a relational database model, so scalability is limited.

Scalable informatics frameworks are designed to enable expansion of data within a system and extension of the system to additional sites. Examples include the Cancer Biomedical Informatics Grid (caBIG) and the Informatics for Integrating Biology and the Bedside (i2b2). Both caBIG and i2b2 provide a standardized framework that can be shared across institutions. These large informatics platforms may present opportunities for integrating clinical research and clinical care data because they can link to data elements that are external to EHRs.

The i2b2 initiative is funded by the U.S. National Institutes of Health (NIH) and led by the U.S. National Center for Biomedical Computing. The effort is dedicated to delivering a scalable methodological informatics framework with supporting tools for managing and sharing clinical information that is optimized for genomic, molecular, and disease-based searches in order to promote the goals of translational research. Currently, i2b2 is being used by more than 60 institutions, and functionality is being extended at some sites to manage their custom research registries and EHR-based registries. The i2b2 platform provides a robust infrastructure for sharing important clinical data sets from multiple, disconnected sources and increasing usability through standard ontologies.

The caBIG program, funded by the U.S. National Cancer Institute, was developed to increase researchers’ capacity to use and share biomedical information through the development and sharing of software and utilities. The underlying caGRID provides the platform that connects the caBIG tools within and across institutions using an integrated framework (9). By delivering infrastructure as a service, caBIG allows contributors to share data across localities and domains, using standardized interfaces so that users can express key clinical terms in a common manner. There has been only modest uptake of caBIG tools and its infrastructure among cancer networks, and caBIG’s problems highlight the complexity of large IT initiatives and the challenge of integrating clinical research across sites (10).

Table 1. Characteristics of research data by data source.

<table>
<thead>
<tr>
<th>Differences</th>
<th>Clinically derived data</th>
<th>Traditional research data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data quality</td>
<td>Data in EHRs may be incomplete or incorrect. Data definitions and terminology are not uniform.</td>
<td>Data are reviewed for discrepancies or missing values. Queries may be sent to research teams to resolve questions.</td>
</tr>
<tr>
<td>Availability of clinical data elements</td>
<td>EHR data available for extraction are often limited to demographics, diagnoses, laboratory values, and health care use. Clinical outcome information is limited.</td>
<td>Data forms are customized for each phase of study and contain all necessary variables.</td>
</tr>
<tr>
<td>Follow-up data</td>
<td>Presence and timing of follow-up information on study participants are variable.</td>
<td>Follow-up data are collected at prespecified times. Extensive efforts are made to obtain follow-up information on study participants.</td>
</tr>
<tr>
<td>Representativeness of study cohorts</td>
<td>Study cohorts consist of “real-world” patients. Observational studies of cohorts may not require participant consent. Characteristics of study cohorts derived from EHRs from single institutions may reflect local demographics or physician-referral patterns.</td>
<td>Study participants meet strict inclusion and exclusion criteria and agree to participate in the clinical trial. Management of disease processes is fairly standardized. Study participants often do not reflect the majority of patients managed in clinical practice.</td>
</tr>
</tbody>
</table>
TRANSLATIONAL DIVIDE

Gaining access to clinical data for research is only the beginning. Researchers must be cognizant of the fundamental differences between data collected as part of clinical care and more rigid protocol-based data collected as part of a clinical research project (Table 1).

**Data errors.** The quality of the data available in clinically derived data sets is more variable than in data sets developed specifically for clinical research. Documentation in the EHR is affected by the complexities of the clinical encounter, time pressures of providers, varied approaches to care and documentation styles, and the uncertainties inherent in medicine. Nonrandom errors in the data usually result in systematic bias, which can lead to invalid conclusions. Careful data collection done as part of a traditional research clinical trial can reduce this bias by using standardized definitions and increasing the completeness of data collection. This approach can be difficult to do with some clinically derived data sets, and statistical approaches such as imputation or the use of algorithms limit some of the potential effect of bias on study results.

**Availability of clinical information in data sets.** Limitations in the availability of clinical information restrict the types of research questions that can be addressed. Clinically derived data are, by definition, obtained as part of the patient care process in which the main focuses are documentation of the patient's clinical characteristics and therapeutic plan and the necessary information for accurate billing. Most information is entered as free text except for items required for billing, such as the encounter diagnoses. Data elements that define clinical course, disease severity, and, importantly, patient outcomes of care are generally not entered in a uniform fashion in EHRs. Thus analyses are often limited to relatively crude outcome measures such as readmission and mortality or surrogate outcomes such as changes in laboratory test values.

For many diseases, there are no validated outcomes of care available in the EHR. In traditional clinical research studies, data collection forms, which are typically dutifully completed, prevent this problem. A primary focus of the KP is to provide patient-reported outcomes such as health status, which greatly enhances its potential for use in clinical research. But this practice requires effort from patients and their health care teams. The HITECH act may improve the discrete collection of clinical information by providing financial incentives for institutions and physicians to implement EHRs that meet criteria for “meaningful use (1).” Some of the meaningful use requirements set by HITECH include use of Systematized Nomenclature of Medicine (SNOMED) as a terminology tool and increased use of “Problem Lists,” in which indications and comorbid conditions are entered as discrete electronic fields.

**Follow-up information.** If EHRs contain any follow-up data, the clinically derived data sets typically are garnered at variable time points. The pattern of follow-up is typically nonrandom, which can lead to systematic bias and inaccurate conclusions. To minimize bias from lack of follow-up data in selected patient populations in the KP, patients receive a telephone call if they do not return for follow-up within 30 days. Traditional research trials, in contrast, have prespecified time points for patient follow-up, and every effort is made to ensure that these processes are completed.

**Representativeness of the patient samples.** Sampling bias occurs when patient and disease characteristics differ from the represented patient population. Sampling bias commonly occurs in EHR-derived databases from single institutions, as patient populations reflect the local socioeconomic environment or specialty interests of the hospital. Although statistical approaches sometimes can be used to reduce this problem, investigators must often limit the study conclusions. Subjects enrolled in clinical trials may not be representative of target population either; they meet specific inclusion and exclusion criteria that often do not match the characteristics of patients seen in real-world clinical practice. In addition, subjects who consent to enroll in a clinical trial may differ from patients who would prefer not to participate, and they may receive closer follow-up or more structured management.

CLOSING THE GAP

Although the differences between clinically derived data and structured research protocol–based data raise challenges, the logistical and financial advantages of using clinical data for research are tremendous. The ability to evaluate interventions in real-world settings provided by clinical medicine increases the likelihood that an intervention’s efficacy in a study can be translated into effectiveness in clinical practice. Large data networks such as i2b2 provide more representative patient populations for analysis. The estimated cost of running a traditional clinical trial in the United States ranges from $20 million for a phase 1 trial to $100 million for a phase 3 trial (http://www.appliedclinicaltrialsonline.com/appliedclinicaltrials/Online+Extras/The-Seven-Elements-of-a-Clinical-Collaboration-Pla/ArticleStandard/Article/detail/698117). These costs make it impossible to evaluate many clinical questions in randomized controlled trials. Clinically derived data are being collected for purposes of clinical care, and dual use of these data for research would drastically reduce costs and enable studies that otherwise would be cost-prohibitive.

Integrating clinical and research data is crucial to the goals of clinical research—to apply new knowledge emanating from the nation’s biomedical research investment to patients and populations; to efficiently conduct clinical trials of drugs and devices to generate new therapies; and to define and compare the value of different health care interventions. Potentially analyzable data are already being collected during the course of patient care, and much of this will be housed within electronic formats in the near future. Consequently, intensive work is occurring on multiple fronts to coordinate data systems to better integrate the EHR data with research applications. Large efforts in the United States such as caBIG and i2b2 may provide practical solutions, allowing integration of genomic and imaging data with clinical data in massive data systems. In addition to the ability to integrate across institutions, these systems may allow identification of patients for multicenter research projects and discovery of new relationships among genomic markers, family history, and disease expression. These initiatives will reduce the problem of patient representativeness and extend the ability to perform population-level analyses.

However, these initiatives will not solve problems associated with data completeness or the availability of patient-centered outcomes information. A partial solution is the use of natural language processing (NLP) and similar tools to extract important clinical data elements from free-text narratives in the medical record. Future advances in NLP might allow data extraction on more complicated constructs such as severity of illness. However, data extraction will always be limited by the amount of clinical detail provided in the medical record.
If we are to take advantage of the opportunities that EHRs hold for improvement in patient care, research, quality improvement, and cost analysis, it will be essential to develop ways to generate, store, and retrieve clinical information such that it is easily accessible to the end user for research. With such a system, large practical clinical trials could eventually supplant the need for traditional randomized clinical trials, and consenting patients who are already receiving medical care could populate the studies. By definition, subjects in these studies would be more representative of patients seen in real-world practices. Large practical trials have been suggested as a potential means of moving medical care closer to a learning health care system. But without integrating research and clinical informatics, large practical trials on a wide scale are also impractical. Integration of research and clinical informatics is essential not only for academic medicine but also for the entire health care system. At a time of rapid growth in health care expenditures and anticipated constraints on research budgets, integrating research and clinical informatics should be considered an urgent national and international priority.

REFERENCES


Time to Integrate Clinical and Research Informatics
Irene L. Katzan and Richard A. Rudick

Sci Transl Med 4, 162fs41
DOI: 10.1126/scitranslmed.3004583

ARTICLE TOOLS http://stm.sciencemag.org/content/4/162/162fs41

RELATED CONTENT http://stm.sciencemag.org/content/scitransmed/3/79/79re1.full
http://stm.sciencemag.org/content/scitransmed/2/18/18cm6.full
http://stm.sciencemag.org/content/scitransmed/2/57/57cm29.full
http://stm.sciencemag.org/content/scitransmed/5/179/179re1.full
http://stm.sciencemag.org/content/scitransmed/5/179/179lr1.full
http://stm.sciencemag.org/content/scitransmed/1/44cm4.full
http://stm.sciencemag.org/content/scitransmed/4/149/149fs32.full
http://stm.sciencemag.org/content/scitransmed/5/206/206ra140.full
http://stm.sciencemag.org/content/scitransmed/6/224/224ed4.full
http://stm.sciencemag.org/content/scitransmed/7/318/318fs49.full

REFERENCES This article cites 8 articles, 4 of which you can access for free
http://stm.sciencemag.org/content/4/162/162fs41#BIBL

PERMISSIONS http://www.sciencemag.org/help/reprints-and-permissions

Use of this article is subject to the Terms of Service