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 CRI research and development efforts are (1) the extensive and diverse informatics needs of the NIH Clinical and Translational Sciences Awards (CTSA),4–6 and (2) the growing interest in sustainable, large-scale, multi-institutional distributed research networks for comparative effectiveness research.7–9 Given the large landscape that comprises translational science, CRI scientists are asked to conceive innovative informatics solutions that span biological, clinical, and population-based research. It is therefore not surprising that the field has simultaneously borrowed from and contributed to many related informatics disciplines.

Paralleling the growth in CRI prominence, JAMIA has received an increasing number of CRI submissions. In 2010, five published articles were completely focused on CRI,10–14 while in 2011 this number rose to 23,15–37 accounting for 11.5% of all JAMIA articles for that year. There was a special section focused on CRI papers in the December 2011 supplement issue. Much of the increase can be attributed to publications from awardees of the CTSA, since publication rate is related to funding.38 JAMIA publications acknowledging CTSA funding rose from three in 200939–41 to four in 201014 42–44 and 15 in 2011.15 17 19 36 45–55 Some of the articles were not exclusively focused on CRI, but were directly related, covering many different topics that are highly relevant to CRI: data models and terminologies,27 56–68 natural language processing (NLP),16 50 61 69–99 surveillance systems,48 65 80 and privacy technology and policy.33 111–117 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.118 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.119 120 The ‘inner’ 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

1Department of Pediatrics, University of Colorado, Aurora, Colorado, USA
2Department of Biomedical Informatics, Columbia University, New York, New York, USA

Correspondence to Dr Michael G Kahn, c/o Children’s Hospital Colorado, 13123 East 16th Avenue, B400, Aurora, CO 80045, USA; michael.kahn@ucdenver.edu

Received 23 March 2012
Accepted 26 March 2012
Published Online First 20 April 2012
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.\(^{13,41,157}\)

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 and Enrollment Research Network (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 e45) 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.\(^{40}\) 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 the analytic data set, the code required to identify the study cohort occupies three full pages, highlighting that the EHR, while providing access to detailed clinical data, requires very complex query logic to ensure that the right patients have been extracted. Their study shows that EHR data can play an important role in monitoring unnecessary test orders and containing healthcare costs.

Li and colleagues (see page e54) 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)22 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)143 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)144 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.145 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’. 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. We also expect more CRI research that is informed by and responsive to patient or population needs. We encourage investigators developing new methods and tools that accelerate clinical and translational research to continue to contribute to the explosive growth in the peer-reviewed literature in clinical research informatics.

Funding Dr Kahn was supported in part by NIH/NCRR Colorado CTSI Grant Number UL1 RR025790 and AHRQ R15 HS019908 (Scalable Architecture for Federated 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 R01LM010185 from the National Library of Medicine, grant UL1RR024156 from the National Center for Research Resources, and AHRQ grant R01HS011853.

Competing interests None.

Provenance and peer review Commissioned; internally peer reviewed.

REFERENCES
1. Embi PJ, Payne PR. Clinical research informatics: challenges, opportunities and definition for an emerging domain. J Am Med Inform Assoc 2009;16:316–27.
2. Payne PR, Embi PJ, Sen CK. Translational informatics: enabling high-throughput research paradigms. Physiol Genomics 2009;39:131–40.
3. Westfall JM, Mold J, Fagan L. Practice-based research—“Blue Highways” on the NIH roadmap. JAMA 2007;297:403–6.
4. Dilts DM, Rosenblum D, Trochim WM. A virtual national laboratory for reengineering classical translational science. Sci Transl Med 2012;4:118cm2.
5. Zerhouni EA, Akgin B. Clinical and translational science awards: a framework for a national research agenda. Transit Res 2006;148:4–5.
6. Calif RM, Berglund L, Principal Investigators of National Institutes of Health Clinical and Translational Science Awards. Linking scientific discovery and better health for the nation: the first three years of the NIH’s clinical and translational science awards. Acad Med 2010;85:467–62.
7. Toh S, Platt R, Steiner JF, et al. Comparative-effectiveness research in distributed health data networks. Clin Pharmacol Ther 2011;90:883–7.
8. Pace WD, Duflores M, Valulak RJ, et al. An electronic practice-based network for observational comparative effectiveness research. Ann Intern Med 2009;151:335–40.
9. Brown JS, Holmes JH, Shah K, et al. Distributed health data networks: a practical and preferred approach to multi-institutional evaluations of comparative effectiveness, safety, and quality of care. Med Care 2010;48(Suppl 6):S345–51.
10. Chute CG, Steck SA, Fisk TB, et al. The Enterprise Data Trust at Mayo Clinic, a semantically integrated warehouse of biomedical data. J Am Med Inform Assoc 2010;17:131–5.
11. Crowley RS, Castine M, Mitchell K, et al. cTIES: a grid based system for coding and retrieval of surgical pathology reports and tissue specimens in support of translational research. J Am Med Inform Assoc 2010;17:253–64.
12. Johnson SB, Whitney G, McAuliffe M, et al. Using global unique identifiers to link autism cohorts. J Am Med Inform Assoc 2010;17:689–95.
13. Murphy SN, Weber G, Mendis M, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc 2010;17:124–30.
14. Payne PR, Embi PJ, Niland J. Foundational biomedical informatics research in the clinical and translational science era: a call to action. J Am Med Inform Assoc 2010;17:615–8.
15. Zheng K, Mei Q, Hanauer DA. Collaborative search in electronic health records. J Am Med Inform Assoc 2011;18:282–91.
16. Xu H, Jiang M, Detjens M, et al. Facilitating pharmacogenetic studies using electronic health records and natural-language processing: a case study of warfarin. J Am Med Inform Assoc 2011;18:387–91.
17. Weng C, Wu X, Luo Z, et al. EXR: a retrieval approach to eligibility criteria extraction and representation. J Am Med Inform Assoc 2011;18(Suppl 1):116–24.
18. Weber GM, Barnett W, Conlon M, et al. Direct2Experts Collaboration. Direct2Experts: a pilot national network to demonstrate interoperability among research-networking platforms. J Am Med Inform Assoc 2011;18(Suppl 1):157–60.
19. Wade TD, Hum RC, Murphy JR. A Dimensional Bus model for integrating clinical and research data. J Am Med Inform Assoc 2011;18(Suppl 1):98–102.
20. Tait AR, Voepel-Lewis T, McConnel M, et al. Evaluation of a prototype interactive case report program for pediatric clinical trials: a pilot study. J Am Med Inform Assoc 2012;19:e43–8.
21. Segadini G, Ferrari F, Lanzia C, et al. R engine class: integrating R into the i2b2 software infrastructure. J Am Med Inform Assoc 2011;18:314–17.
Perspective

115. El Emam K, Neri E, Jonker E, et al. The inadvertent disclosure of personal health information through peer-to-peer file sharing programs. J Am Med Inform Assoc 2010;17:148–58.

116. Leukides G, Denny JC, Malin B. The disclosure of diagnosis codes can breach research participants’ privacy. J Am Med Inform Assoc 2010;17:322–7.

117. Yeniterzi R, Aberdeen J, Bayer S, et al. Effects of personal identifier resynthesis on clinical text de-identification. J Am Med Inform Assoc 2010;17:559–68.

118. Jacobsen KH. Introduction to Health Research Methods: A Practical Guide. Sudbury, Mass: Jones & Bartlett Learning, 2012.

119. Woolf SH. The meaning of translational research and why it matters. JAMA 2008; 299:211–13.

120. Dougherty D, Slater T, et al. A model-driven approach for biomedical data integration. Stud Health Technol Inform 2008;139:291–7.

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

122. Sullivan J. Subject Recruitment and Retention: Barriers to Success

123. Denny JC, Malin B. The disclosure of diagnosis codes can breach research participants’ privacy. J Am Med Inform Assoc 2010;17:322–7.

124. Carlson D, Farkash A, Timm JT. A model-driven approach for biomedical data integration. Stud Health Technol Inform 2010;160:1164–9.

125. Bodenreider O. Biomedical ontologies in action: role in knowledge management, data integration and decision support. Yearb Med Inform 2008:67–79.

126. Slater T, Bouton C, Huang ES. Beyond data integration. Drug Discov Today 2008;13:584–9.

127. Gardner SP. Ontologies and semantic data integration. Drug Discov Today 2005;10:1001–7.

128. Cantor MN, Lussier YA. Putting data integration into practice: using biomedical terminologies to add structure to existing data sources. JAMA Arch Symp Proc 2003:125–9. PMID: PMC1480054.

129. Maro JC, Platt R, Holmes JH, et al. Design of a national distributed health data network. Ann Intern Med 2009;151:341–4.

130. Kuo MH, Kushniruk AW, Borycki EM. Design and implementation of a health data integration tool to answer diverse biologic queries. Stud Health Technol Inform 2004;107:768–72.

131. Donelson L, Tarczy-Hornoch P, Mork P, et al. The BioMediator system as a data integration tool to answer diverse biologic queries. Stud Health Technol Inform 2004;107:768–72.

132. Kakazu K, Cheung LW, Lynne W. The Cancer Biomedical Informatics Grid (caBIG): pioneering an expansive network of information and tools for collaborative cancer research. Hawaii Med J 2004;63:273–5.

133. caBIG Strategic Planning Workspace. The Cancer Biomedical Informatics Grid (caBIG): infrastructure and applications for a worldwide research community. Stud Health Technol Inform 2007;129:330–4.

134. Keator DB, Grethe JS, Marcus D, et al. A national human neuroimaging collaboratory enabled by the Biomedical Informatics Research Network (BIRN). IEEE Trans Inf Technol Biomed 2008;12:162–72.

135. MacKenzie S, Wyatt M, Schiff R, et al. Practices and Perspectives on Building Integrated Data Repositories: Results from a 2010 CTSA Survey. J Am Med Inform Assoc 2012;19:e119–e124.

136. Abend A, Housman D, Johnson B. Integrating clinical data into the i2b2 repository. Summit on Translat Bioinforma 2009;2009:1–5.

137. Clark A, Hansa KE, Farnish G. Still thinking research: strategies to advance the use of electronic health records to bridge patient care and research. 2011.

138. Embi PJ, Leonard AC. Evaluating Alert Fatigue Over Time to EHR-based Clinical Trial Alerts: Findings from a Randomized, Controlled Study. J Am Med Inform Assoc 2012;19:e145–e148.

139. Jiang X, Boxwala A, El-Kareh R, et al. A patient-driven adaptive prediction technique (ADAPT) to improve personalized risk estimation for clinical decision support. J Am Med Inform Assoc 2012;19:e137–e144.

140. Matthias JS, Gossett D, Baker DW. Use of electronic health record data to evaluate overuse of cervical cancer screening. J Am Med Inform Assoc 2012;19:e96–e101.

141. McCartney PB, Ladwa S, Oberti M, et al. Informatics and data quality at collaborative multicenter breast and colon cancer family registries. J Am Med Inform Assoc 2012;19:e125–e128.

142. Jiang G, Slobirg HR, Chute CG. Quality evaluation of value sets from cancer study common data elements using the UMLS semantic groups. J Am Med Inform Assoc 2012;19:e129–e136.

143. Ong MS, Magrabi F, Coiera E. Automated identification of extreme-risk events in clinical incident reports. J Am Med Inform Assoc 2012;19:e110–e118.

144. Savova GK, Olsson JE, Murphy SP, et al. Automated discovery of drug treatment patterns for endocrine therapy of breast cancer within an electronic medical record. J Am Med Inform Assoc 2012;19:e63–e69.

145. Chapman WW, Cohen KB. Current issues in biomedical text mining and natural language processing. J Biomed Inform 2009;42:757–9.

146. López-García P, Boeker M, Illarramendi A, et al. Usability-driven pruning of large ontologies: the case of SNOMED CT. J Am Med Inform Assoc 2012;19:e102–e109.

147. Wu S, Liu H, Li D, et al. UMLS Term occurrences in clinical notes: a large-scale corpus analysis. J Am Med Inform Assoc 2012;19:e149–e156.

148. Hoof L, Perlmutter RM. The impact of systems approaches on biological problems in drug discovery. Nat Biotechnol 2004;22:1215–17.

149. Onno-Machado L, Bafna V, Boxwala AA, et al. IDaSH team. iDASH: integrating data for analysis, anonymization, and sharing. J Am Med Inform Assoc 2012;19:196–201.