Clinical research informatics: a conceptual perspective

Michael G Kahn,[1] Chunhua Weng[2]

[1]Department of Pediatrics, University of Colorado, Aurora, Colorado, USA
[2]Department 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

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). We present a conceptual model based on an informatics-enabled clinical research workflow, integration across heterogeneous data sources, and core informatics tools and platforms. We use this conceptual model to highlight 18 new articles in the *JAMIA* special issue on clinical research informatics.

Clinical research informatics (CRI) 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[1 2]: 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).[3] 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 (CTSAs),[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 2009[39–41] to four in 2010[14 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 80 100–110] 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

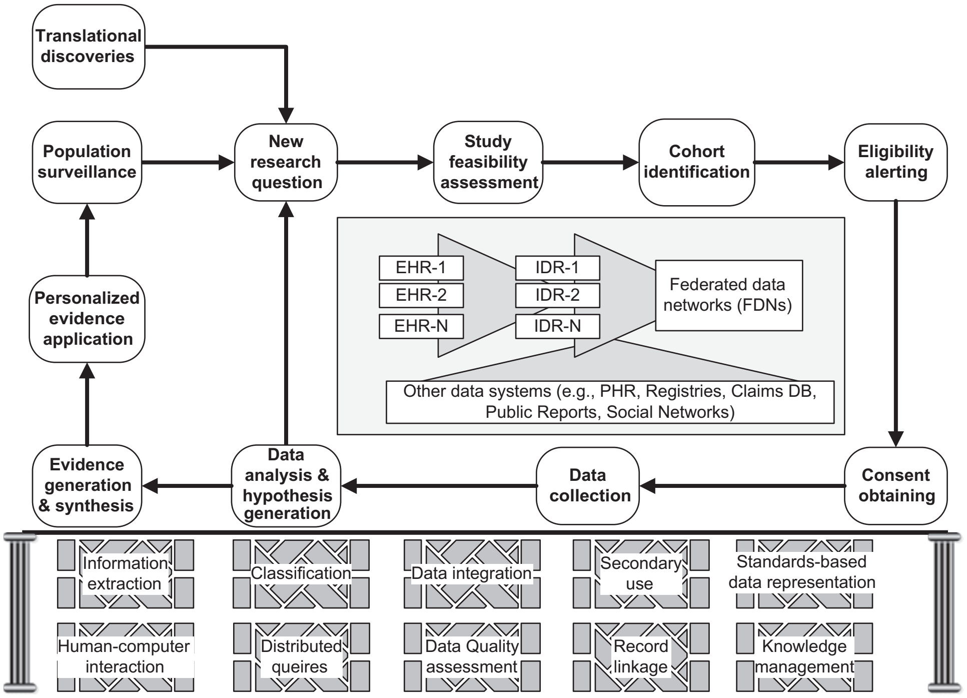
Translational discoveries
Population surveillance
New research question
Study feasibility assessment
Cohort identification
Eligibility alerting
Personalized evidence application
EHR-1
EHR-2
EHR-N
IDR-1
IDR-2
IDR-N
Federated data networks (FDNs)
Other data systems (e.g., PHR, Registries, Claims DB, Public Reports, Social Networks)
Evidence generation & synthesis
Data analysis & hypothesis generation
Data collection
Consent obtaining
Information extraction
Classification
Data integration
Secondary use
Standards-based data representation
Human-computer interaction
Distributed queires
Data Quality assessment
Record linkage
Knowledge management

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.[121 122] 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.[123—128] 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.[8 9 129—132] 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).[31 133—135] 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 CTSAs and the NIH Clinical Center.[136] 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 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.[138] 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[141] (***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 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***)[33] 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.[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'.[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.[150] 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 RR025780 and AHRQ R01HS019908 (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 R01LM010815 from the National Library of Medicine, grant UL1RR024156 from the National Center for Research Resources, and AHRQ grant R01HS019853.

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, Fagnan 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 clinical translational science. *Sci Transl Med* 2012;**4**:118cm2.
5. **Zerhouni EA,** Alving B. Clinical and translational science awards: a framework for a national research agenda. *Transl Res* 2006;**148**:4–5.
6. **Califf 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**:457–62.
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,** Cifuentes M, Valuck RJ, *et al*. An electronic practice-based network for observational comparative effectiveness research. *Ann Intern Med* 2009;**151**:338–40.
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):S45–51.
10. **Chute CG,** Beck 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*. caTIES: 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 collections. *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**:615–16.
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, Oetjens 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*. EliXR: an approach to eligibility criteria extraction and representation. *J Am Med Inform Assoc* 2011;**18**(Suppl 1):i116–24.
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): i157–60.
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):i96–102.
20. **Tait AR,** Voepel-Lewis T, McGonegal M, *et al*. Evaluation of a prototype interactive consent program for pediatric clinical trials: a pilot study. *J Am Med Inform Assoc* 2012;**19**:e43–e45.
21. **Segagni D,** Ferrazzi F, Larizza C, *et al*. R engine cell: integrating R into the i2b2 software infrastructure. *J Am Med Inform Assoc* 2011;**18**:314–17.
22. **Sariyar M,** Borg A, Pommerening K. Missing values in deduplication of electronic patient data. *J Am Med Inform Assoc* 2012;**19**:e76–e82.
23. **Richesson RL,** Nadkarni P. Data standards for clinical research data collection forms: current status and challenges. *J Am Med Inform Assoc* 2011;**18**:341–6.
24. **Rasmussen LV,** Peissig PL, McCarty CA, *et al*. Development of an optical character recognition pipeline for handwritten form fields from an electronic health record. *J Am Med Inform Assoc* 2012;**19**:e90–e95.
25. **Payne PR,** Borlawsky TB, Lele O, *et al*. The TOKEn project: knowledge synthesis for in silico science. *J Am Med Inform Assoc* 2011;**18**(Suppl 1):i125–31.
26. **Pathak J,** Wang J, Kashyap S, *et al*. Mapping clinical phenotype data elements to standardized metadata repositories and controlled terminologies: the eMERGE Network experience. *J Am Med Inform Assoc* 2011;**18**:376–86.
27. **Pathak J,** Chute CG. Further revamping VA's NDF-RT drug terminology for clinical research. *J Am Med Inform Assoc* 2011;**18**:347–8.
28. **Murphy SN,** Gainer V, Mendis M, *et al*. Strategies for maintaining patient privacy in i2b2. *J Am Med Inform Assoc* 2011;**18**(Suppl 1):i103–8.
29. **Kagan JM,** Gupta N, Varghese S, *et al*. The NIAID Division of AIDS enterprise information system: integrated decision support for global clinical research programs. *J Am Med Inform Assoc* 2011;**18**(Suppl 1):i161–5.
30. **Herasevich V,** Pieper MS, Pulido J, *et al*. Enrollment into a time sensitive clinical study in the critical care setting: results from computerized septic shock sniffer implementation. *J Am Med Inform Assoc* 2011;**18**:639–44.
31. **Helmer KG,** Ambite JL, Ames J, *et al*; Biomedical Informatics Research Network. Enabling collaborative research using the biomedical informatics research network (BIRN). *J Am Med Inform Assoc* 2011;**18**:416–22.
32. **Ferranti JM,** Gilbert W, McCall J, *et al*. The design and implementation of an open-source, data-driven cohort recruitment system: the Duke Integrated Subject Cohort and Enrollment Research Network (DISCERN). *J Am Med Inform Assoc* 2012;**19**:e68–e75.
33. **Duvall SL,** Fraser AM, Rowe K, *et al*. Evaluation of record linkage between a large healthcare provider and the Utah Population Database. *J Am Med Inform Assoc* 2012;**19**:e54–e59.
34. **Carroll AE,** Biondich PG, Anand V, *et al*. Targeted screening for pediatric conditions with the CHICA system. *J Am Med Inform Assoc* 2011;**18**:485–90.
35. **Buck MD,** Anane S, Taverna J, *et al*. The Hub Population Health System: distributed ad hoc queries and alerts. *J Am Med Inform Assoc* 2012;**19**:e46–e50.
36. **Borlawsky TB,** Lele O, Jensen D, *et al*. Enabling distributed electronic research data collection for a rural appalachian tobacco cessation study. *J Am Med Inform Assoc* 2011;**18**(Suppl 1):i140–3.
37. **Anderson N,** Abend A, Mandel A, *et al*. Implementation of a deidentified federated data network for population-based cohort discovery. *J Am Med Inform Assoc* 2012;**19**:e60–e67.
38. **Boyack KW,** Jordan P. Metrics associated with NIH funding: a high-level view. *J Am Med Inform Assoc* 2011;**18**:423–31.
39. **Boyd AD,** Saxman PR, Hunscher DA, *et al*. The University of Michigan Honest Broker: a web-based Service for clinical and translational research and practice. *J Am Med Inform Assoc* 2009;**16**:784–91.
40. **Thadani SR,** Weng C, Bigger JT, *et al*. Electronic screening improves efficiency in clinical trial recruitment. *J Am Med Inform Assoc* 2009;**16**:869–73.
41. **Weber GM,** Murphy SN, McMurry AJ, *et al*. The shared health research information network (SHRINE): a prototype federated query tool for clinical data repositories. *J Am Med Inform Assoc* 2009;**16**:624–30.
42. **Kahn MG,** Ranade D. The impact of electronic medical records data sources on an adverse drug event quality measure. *J Am Med Inform Assoc* 2010;**17**:185–91.
43. **Yackel TR,** Embi PJ. Unintended errors with EHR-based result management: a case series. *J Am Med Inform Assoc* 2010;**17**:104–7.
44. **Zheng K,** Haftel HM, Hirschl RB, *et al*. Quantifying the impact of health IT implementations on clinical workflow: a new methodological perspective. *J Am Med Inform Assoc* 2010;**17**:454–61.
45. **Banas CA,** Erskine AR, Sun S, *et al*. Phased implementation of electronic health records through an office of clinical transformation. *J Am Med Inform Assoc* 2011;**18**:721–5.
46. **Foran DJ,** Yang L, Chen W, *et al*. ImageMiner: a software system for comparative analysis of tissue microarrays using content-based image retrieval, high-performance computing, and grid technology. *J Am Med Inform Assoc* 2011;**18**:403–15.
47. **Gadd CS,** Ho YX, Cala CM, *et al*. User perspectives on the usability of a regional health information exchange. *J Am Med Inform Assoc* 2011;**18**:711–16.
48. **Harkema H,** Chapman WW, Saul M, *et al*. Developing a natural language processing application for measuring the quality of colonoscopy procedures. *J Am Med Inform Assoc* 2011;**18**(Suppl 1):i150–6.
49. **Hoeksema LJ,** Bazzy-Asaad A, Lomotan EA, *et al*. Accuracy of a computerized clinical decision-support system for asthma assessment and management. *J Am Med Inform Assoc* 2011;**18**:243–50.
50. **Nadkarni PM,** Ohno-Machado L, Chapman WW. Natural language processing: an introduction. *J Am Med Inform Assoc* 2011;**18**:544–51.
51. **Sarkar IN,** Butte AJ, Lussier YA, *et al*. Translational bioinformatics: linking knowledge across biological and clinical realms. *J Am Med Inform Assoc* 2011;**18**:354–7.
52. **Sarkar U,** Karter AJ, Liu JY, *et al*. Social disparities in internet patient portal use in diabetes: evidence that the digital divide extends beyond access. *J Am Med Inform Assoc* 2011;**18**:318–21.

53. **Zheng K,** Fear K, Chaffee BW, *et al*. Development and validation of a survey instrument for assessing prescribers' perception of computerized drug—drug interaction alerts. *J Am Med Inform Assoc* 2011;**18**(Suppl 1):i51—61.
54. **Zheng K,** Guo MH, Hanauer DA. Using the time and motion method to study clinical work processes and workflow: methodological inconsistencies and a call for standardized research. *J Am Med Inform Assoc* 2011;**18**:704—10.
55. **Zheng K,** Hanauer DA, Padman R, *et al*. Handling anticipated exceptions in clinical care: investigating clinician use of 'exit strategies' in an electronic health records system. *J Am Med Inform Assoc* 2011;**18**:883—9.
56. **Elhanan G,** Perl Y, Geller J. A survey of SNOMED CT direct users, 2010: impressions and preferences regarding content and quality. *J Am Med Inform Assoc* 2011;**18**(Suppl 1):i36—44.
57. **Liu H,** Burkhart Q, Bell DS. Evaluation of the NCPDP Structured and Codified Sig Format for e-prescriptions. *J Am Med Inform Assoc* 2011;**18**:645—51.
58. **Nelson SJ,** Zeng K, Kilbourne J, *et al*. Normalized names for clinical drugs: RxNorm at 6 years. *J Am Med Inform Assoc* 2011;**18**:441—8.
59. **Rector AL,** Brandt S, Schneider T. Getting the foot out of the pelvis: modeling problems affecting use of SNOMED CT hierarchies in practical applications. *J Am Med Inform Assoc* 2011;**18**:432—40.
60. **Aronson AR,** Lang FM. An overview of MetaMap: historical perspective and recent advances. *J Am Med Inform Assoc* 2010;**17**:229—36.
61. **D'Avolio LW,** Nguyen TM, Farwell WR, *et al*. Evaluation of a generalizable approach to clinical information retrieval using the automated retrieval console (ARC). *J Am Med Inform Assoc* 2010;**17**:375—82.
62. **Fung KW,** McDonald C, Srinivasan S. The UMLS-CORE project: a study of the problem list terminologies used in large healthcare institutions. *J Am Med Inform Assoc* 2010;**17**:675—80.
63. **Green DL,** Boonstra JA, Bober MA. Use of a codified medication process for documentation of home medications. *J Am Med Inform Assoc* 2010;**17**:608—12.
64. **Nadkarni PM,** Darer JA. Migrating existing clinical content from ICD-9 to SNOMED. *J Am Med Inform Assoc* 2010;**17**:602—7.
65. **Nadkarni PM,** Marenco LA. Implementing description-logic rules for SNOMED-CT attributes through a table-driven approach. *J Am Med Inform Assoc* 2010;**17**:182—4.
66. **Pathak J,** Chute CG. Analyzing categorical information in two publicly available drug terminologies: RxNorm and NDF-RT. *J Am Med Inform Assoc* 2010;**17**:432—9.
67. **Pathak J,** Peters L, Chute CG, *et al*. Comparing and evaluating terminology services application programming interfaces: RxNav, UMLSKS and LexBIG. *J Am Med Inform Assoc* 2010;**17**:714—19.
68. **Stanfill MH,** Williams M, Fenton SH, *et al*. A systematic literature review of automated clinical coding and classification systems. *J Am Med Inform Assoc* 2010;**17**:646—51.
69. **Botsis T,** Nguyen MD, Woo EJ, *et al*. Text mining for the Vaccine Adverse Event Reporting System: medical text classification using informative feature selection. *J Am Med Inform Assoc* 2011;**18**:631—8.
70. **Clark C,** Aberdeen J, Coarr M, *et al*. MITRE system for clinical assertion status classification. *J Am Med Inform Assoc* 2011;**18**:563—7.
71. **D'Avolio LW,** Nguyen TM, Goryachev S, *et al*. Automated concept-level information extraction to reduce the need for custom software and rules development. *J Am Med Inform Assoc* 2011;**18**:607—13.
72. **de Bruijn B,** Cherry C, Kiritchenko S, *et al*. Machine-learned solutions for three stages of clinical information extraction: the state of the art at i2b2 2010. *J Am Med Inform Assoc* 2011;**18**:557—62.
73. **Garla V,** Lo Re V 3rd, Dorey-Stein Z, *et al*. The Yale cTAKES extensions for document classification: architecture and application. *J Am Med Inform Assoc* 2011;**18**:614—20.
74. **Jiang M,** Chen Y, Liu M, *et al*. A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries. *J Am Med Inform Assoc* 2011;**18**:601—6.
75. **Kirchhoff K,** Turner AM, Axelrod A, *et al*. Application of statistical machine translation to public health information: a feasibility study. *J Am Med Inform Assoc* 2011;**18**:473—8.
76. **Minard AL,** Ligozat AL, Ben Abacha A, *et al*. Hybrid methods for improving information access in clinical documents: concept, assertion, and relation identification. *J Am Med Inform Assoc* 2011;**18**:588—93.
77. **Rink B,** Harabagiu S, Roberts K. Automatic extraction of relations between medical concepts in clinical texts. *J Am Med Inform Assoc* 2011;**18**:594—600.
78. **Roberts K,** Harabagiu SM. A flexible framework for deriving assertions from electronic medical records. *J Am Med Inform Assoc* 2011;**18**:568—73.
79. **Savova GK,** Chapman WW, Zheng J, *et al*. Anaphoric relations in the clinical narrative: corpus creation. *J Am Med Inform Assoc* 2011;**18**:459—65.
80. **Sohn S,** Kocher JP, Chute CG, *et al*. Drug side effect extraction from clinical narratives of psychiatry and psychology patients. *J Am Med Inform Assoc* 2011;**18** (Suppl 1):i144—9.
81. **Torii M,** Wagholikar K, Liu H. Using machine learning for concept extraction on clinical documents from multiple data sources. *J Am Med Inform Assoc* 2011;**18**:580—7.
82. **Uzuner O,** South BR, Shen S, *et al*. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *J Am Med Inform Assoc* 2011;**18**:552—6.
83. **Wright A,** Pang J, Feblowitz JC, *et al*. A method and knowledge base for automated inference of patient problems from structured data in an electronic medical record. *J Am Med Inform Assoc* 2011;**18**:859—67.
84. **Agarwal S,** Yu H. Biomedical negation scope detection with conditional random fields. *J Am Med Inform Assoc* 2010;**17**:696—701.
85. **Deleger L,** Grouin C, Zweigenbaum P. Extracting medical information from narrative patient records: the case of medication-related information. *J Am Med Inform Assoc* 2010;**17**:555—8.
86. **Denny JC,** Peterson JF, Choma NN, *et al*. Extracting timing and status descriptors for colonoscopy testing from electronic medical records. *J Am Med Inform Assoc* 2010;**17**:383—8.
87. **Doan S,** Bastarache L, Klimkowski S, *et al*. Integrating existing natural language processing tools for medication extraction from discharge summaries. *J Am Med Inform Assoc* 2010;**17**:528—31.
88. **Hamon T,** Grabar N. Linguistic approach for identification of medication names and related information in clinical narratives. *J Am Med Inform Assoc* 2010;**17**:549—54.
89. **Li Z,** Liu F, Antieau L, *et al*. Lancet: a high precision medication event extraction system for clinical text. *J Am Med Inform Assoc* 2010;**17**:563—7.
90. **Meystre SM,** Thibault J, Shen S, *et al*. Textractor: a hybrid system for medications and reason for their prescription extraction from clinical text documents. *J Am Med Inform Assoc* 2010;**17**:559—62.
91. **Mork JG,** Bodenreider O, Demner-Fushman D, *et al*. Extracting Rx information from clinical narrative. *J Am Med Inform Assoc* 2010;**17**:536—9.
92. **Nguyen AN,** Lawley MJ, Hansen DP, *et al*. Symbolic rule-based classification of lung cancer stages from free-text pathology reports. *J Am Med Inform Assoc* 2010;**17**:440—5.
93. **Patrick J,** Li M. High accuracy information extraction of medication information from clinical notes: 2009 i2b2 medication extraction challenge. *J Am Med Inform Assoc* 2010;**17**:524—7.
94. **Savova GK,** Masanz JJ, Ogren PV, *et al*. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J Am Med Inform Assoc* 2010;**17**:507—13.
95. **Spasic I,** Sarafraz F, Keane JA, *et al*. Medication information extraction with linguistic pattern matching and semantic rules. *J Am Med Inform Assoc* 2010;**17**:532—5.
96. **Tikk D,** Solt I. Improving textual medication extraction using combined conditional random fields and rule-based systems. *J Am Med Inform Assoc* 2010;**17**:540—4.
97. **Uzuner O,** Solti I, Cadag E. Extracting medication information from clinical text. *J Am Med Inform Assoc* 2010;**17**:514—18.
98. **Xu H,** Stenner SP, Doan S, *et al*. MedEx: a medication information extraction system for clinical narratives. *J Am Med Inform Assoc* 2010;**17**:19—24.
99. **Yang H.** Automatic extraction of medication information from medical discharge summaries. *J Am Med Inform Assoc* 2010;**17**:545—8.
100. **Carnevale RJ,** Talbot TR, Schaffner W, *et al*. Evaluating the utility of syndromic surveillance algorithms for screening to detect potentially clonal hospital infection outbreaks. *J Am Med Inform Assoc* 2011;**18**:466—72.
101. **Hasan S,** Duncan GT, Neill DB, *et al*. Automatic detection of omissions in medication lists. *J Am Med Inform Assoc* 2011;**18**:449—58.
102. **Jonikas MA,** Mandl KD. Surveillance of medication use: early identification of poor adherence. *J Am Med Inform Assoc*. Published Online First: 19 November 2011. doi:10.1136/amiajnl-2011-000416
103. **Li Z,** Lai S, Buckeridge DL, *et al*. Adjusting outbreak detection algorithms for surveillance during epidemic and non-epidemic periods. *J Am Med Inform Assoc* 2012;**19**:e51—e53.
104. **Strom BL,** Schinnar R, Jones J, *et al*. Detecting pregnancy use of non-hormonal category X medications in electronic medical records. *J Am Med Inform Assoc* 2011;**18**(Suppl 1):i81—6.
105. **Tinoco A,** Evans RS, Staes CJ, *et al*. Comparison of computerized surveillance and manual chart review for adverse events. *J Am Med Inform Assoc* 2011;**18**:491—7.
106. **Wilcox AB,** Chen YH, Hripcsak G. Minimizing electronic health record patient-note mismatches. *J Am Med Inform Assoc* 2011;**18**:511—14.
107. **Chapman WW,** Dowling JN, Baer A, *et al*. Developing syndrome definitions based on consensus and current use. *J Am Med Inform Assoc* 2010;**17**:595—601.
108. **Fine AM,** Reis BY, Nigrovic LE, *et al*. Use of population health data to refine diagnostic decision-making for pertussis. *J Am Med Inform Assoc* 2010;**17**:85—90.
109. **Hota B,** Lin M, Doherty JA, *et al*. Formulation of a model for automating infection surveillance: algorithmic detection of central-line associated bloodstream infection. *J Am Med Inform Assoc* 2010;**17**:42—8.
110. **Reisinger SJ,** Ryan PB, O'Hara DJ, *et al*. Development and evaluation of a common data model enabling active drug safety surveillance using disparate healthcare databases. *J Am Med Inform Assoc* 2010;**17**:652—62.
111. **Boxwala AA,** Kim J, Grillo JM, *et al*. Using statistical and machine learning to help institutions detect suspicious access to electronic health records. *J Am Med Inform Assoc* 2011;**18**:498—505.
112. **El Emam K,** Hu J, Mercer J, *et al*. A secure protocol for protecting the identity of providers when disclosing data for disease surveillance. *J Am Med Inform Assoc* 2011;**18**:212—17.
113. **Malin B,** Benitez K, Masys D. Never too old for anonymity: a statistical standard for demographic data sharing via the HIPAA Privacy Rule. *J Am Med Inform Assoc* 2011;**18**:3—10.
114. **Benitez K,** Malin B. Evaluating re-identification risks with respect to the HIPAA privacy rule. *J Am Med Inform Assoc* 2010;**17**:169—77.

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. **Loukides 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**:159—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,** Conway PH. The "3T's" road map to transform US health care: the "how" of high-quality care. *JAMA* 2008;**299**:2319—21.
121. **Clinical Trials News.** *Subject Recruitment by Far Biggest Clinical Trial Concern*. Applied Clinical Trials, 2004.
122. **Sullivan J.** *Subject Recruitment and Retention: Barriers to Success*. Applied Clinical Trials, 2004:50—4.
123. **Timm J,** Renly S, Farkash A. Large scale healthcare data integration and analysis using the semantic web. *Stud Health Technol Inform* 2011;**169**:729—33.
124. **Carlson D,** Farkash A, Timm JT. A model-driven approach for biomedical data integration. *Stud Health Technol Inform* 2010;**160**:1164—8.
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. *AMIA Annu Symp Proc* 2003:125—9. PMCID: 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 interoperability mediator. *Stud Health Technol Inform* 2010;**155**:101—7.
131. **Grethe JS,** Ross E, Little D, *et al*. Mediator infrastructure for information integration and semantic data integration environment for biomedical research. *Methods Mol Biol* 2009;**569**:33—53.
132. **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.
133. **Kakazu KK,** 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.
134. **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.
135. **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.
136. **MacKenzie S,** Wyatt M, Schuff 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.
137. **Abend A,** Housman D, Johnson B. Integrating clinical data into the i2b2 repository. *Summit on Translat Bioinforma* 2009;**2009**:1—5.
138. **Clark A,** Hanna KE, Parrish G. Still thinking research: strategies to advance the use of electronic health records to bridge patient care and research. 2011.
139. **Embi PJ,** Leonard AC. Evaluating Alert Fatigue Over Time to EHR-based Clinical Trial Alerts: Findings from a Randomized, Controlled Study. *J Am Med Informat Assoc* 2012;**19**:e145—e148.
140. **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 Informat Assoc* 2012;**19**:e137—e144.
141. **Mathias 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.
142. **McGarvey 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.
143. **Jiang G,** Solbrig 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.
144. **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.
145. **Savova GK,** Olson 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**:e83—e89.
146. **Chapman WW,** Cohen KB. Current issues in biomedical text mining and natural language processing. *J Biomed Inform* 2009;**42**:757—9.
147. **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.
148. **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.
149. **Hood L,** Perlmutter RM. The impact of systems approaches on biological problems in drug discovery. *Nat Biotechnol* 2004;**22**:1215—17.
150. **Ohno-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.