Information Technology for Clinical, Translational and Comparative Effectiveness Research

Findings from the Section Clinical Research Informatics

C. Daniel, R. Choquet, Section Editors for the IMIA Yearbook Section on Clinical Research Informatics
INSERM UMRS 872 eq 20, Paris, France

Summary

Objectives: To summarize advances of excellent current research in the new emerging field of Clinical Research Informatics.

Method: Synopsis of four key articles selected for the IMIA Yearbook 2013. The selection was performed by querying PubMed and Web of Science with predefined keywords. From the original set of 590 papers, a first subset of 461 articles which was in the scope of Clinical Research Informatics was refined into a second subset of 79 relevant articles from which 15 articles were retained for peer-review.

Results: The four selected articles exemplify current research efforts conducted in the areas of data representation and management in clinical trials, secondary use of EHR data for clinical research, information technology platforms for translational and comparative effectiveness research and implementation of privacy control.

Conclusions: The selected articles not only illustrate how innovative information technology supports classically organized randomized controlled trials but also demonstrate that the long promised benefits of electronic health care data for research are becoming a reality through concrete platforms and projects.

Keywords

Medical informatics, Biomedical Research, Epidemiologic Study Characteristics as Topic, Evaluation Studies as Topic, Patient Selection

Yearb Med Inform 2013:185-9

Introduction

Clinical research informatics is a new emerging sub-domain in biomedical informatics that addresses the reoccurring demand for tools and methods that can reduce the data management burden and assist investigators as they seek to aggregate and reason upon multi-dimensional and multi-scale data sets for clinical research purposes.

Because research studies are becoming increasingly more extensive and complex, a number of initiatives, including the Human Studies Database (HSDB) [1], LinkedCT [2] and the database for aggregate analysis of ClinicalTrials.gov (AACT) [3] propose access to clinical trials repositories allowing significant information to be re-used across studies. Several papers report efforts for leveraging information management to streamline trial design and optimize production of clinical research deliverables while meeting the requirements enunciated by the US Food and Drug Administration for electronic systems [4].

The adoption rate for IT-supported research recently increased remarkably. Beyond electronic data capture for clinical trials, the development of data repositories for secondary use of clinical data and of integrative platforms now allows researchers and their staff to focus on fundamental scientific problems rather than on practical informatics needs [5]. Deploying a clinical data warehouse in a healthcare organization is a long and methodical process requiring a solid sponsoring from the decision-makers, the motivated staff of various professions and most importantly a strong perceived value by the operational services for the information and its analysis [6]. As an example, the successful Informatics for Integrating Biology and the Bedside (i2b2) solution, adopted by over 60 academic health centers internationally, seeks to provide the instrumentation for using data accumulated through the delivery of health care to conduct discovery research. i2b2 has been used to generate genome-wide studies at less cost and time of conventionally performed studies as well as to identify important risk from commonly used medications [7]. Different methodologies or generic tools are currently being developed to establish clinical data warehouses or to achieve heterogeneous data stores mediation and federation. One of the key challenges for the development of IT infrastructures designed for cross-domain and cross-countries research is the integration of heterogeneous sources of data. Achieving this goal requires the adoption of multiple standards that must be consistent themselves, and cross compatible. In the domain of clinical research, the Clinical Data Information Standards Committee (CDISC) – a non-profit organization - has developed a number of standards for study design, data collection and analysis, and submission to the regulatory bodies. In the domain of patient care, several decade-long and large-scale efforts have focused on specifying both the syntax and the semantics of patient clinical information (CEN/ISO 13606 Reference Model and Archetypes, openEHR, HL7 RIM, HL7 Clinical Document Architecture (CDA) meta-standard and the derived Continuity of Care Document (CCD)). Although CCD was designed for individual communications, it can be effectively be reused for population-based research and public health [8].

The advancing development of IT infrastructures for translational research represents a great opportunity. But at the same time many challenging issues that are not only technical, need to be carefully identified and addressed in collaboration with all the stakeholders to meet their expectations. A key challenge is the definition of a good balance between supporting medical research - by increasing efficient access to specimens and data stored in research platforms - and answering societal demands by securing the rights of specimens and data donors [9, 10]. In addition, giving individuals, whose specimens and data are used for biomedical research, access to the research results and incidental findings is a subject of considerable controversy [11, 12].

About the Paper Selection

A comprehensive review of published articles in 2012 addressing a wide range of issues for clinical research informatics was performed. The selection was performed by querying PubMed and Web of Science with predefined keywords. From the original set of 590 papers, a first subset of 461 articles which was in the scope of Clinical Research Informatics was refined into a second subset of 79 relevant articles from which 15 articles were retained for peer-review.

Table 1 lists the four selected papers from international peer reviewed journals in the fields of medicine and medical informatics that exemplify current research efforts done in the areas of data representation and management in clinical trials, secondary use of EHR data for clinical research, information technology platforms for translational and comparative effectiveness research, and security, confidentiality and regulatory issues.

The focus of the first paper presented in this selection is the implementation of privacy control in biobanks. The authors present solutions that comply with researchers' needs and patient protection [9]. Mining electronic health records (EHRs) for research is the topic explored by the second selected paper. The authors propose a general review of the use of analytical methods of varying complexity (clustering, classification, association discovery, causality analysis, etc.) [13].

In the third selected paper, the authors compare how six informatics platforms providing access to electronic clinical data and the governance infrastructure required for inter-institutional studies have been set up or extended in order to answer specific questions of comparative effectiveness research [14]. The fourth paper reports on a unique effort to prepare and maintain a publicly accessible dataset derived from ClinicalTrials.gov content - the Aggregate Analysis of ClinicalTrials (AACT) database - and to extend its utility by means of an associated clinical specialty taxonomy designed to support research policy analyses [3].

A brief content summary of the four selected papers can be found in the appendix of this synopsis.

Conclusion and Outlook

In the comings years, current research efforts in the field of clinical research informatics are likely to continue in order to support, beyond clinical research, the full cycle of translational research as well as comparative effectiveness research by addressing i) standard-based cross domain and cross country data integration, ii) repurposing of EHR clinical data for building and mining large scale in silico cohorts and next generation registries, iii) ethical issues related to privacy and potential return to participants.

Following previous work in the United States, the International Medical Informatics Association convened the 2012 European Summit on Trustworthy Reuse of Health Data. Over 100 delegates from 21 countries representing national governments, academia, patient groups, industry, and the European Commission contributed to a white paper exploring a wide range of perspectives on **trustworthy reuse of health data** [15]. Institutional national or international programs support the use of the electronic health record (EHR) for secondary purposes. In US, the Strategic Health IT Advanced Research Projects (SHARP) team is developing open source services and components to support the ubiquitous exchange and reuse of operational clinical data stored in electronic health record [16]. In Europe, the Innovative Medicines Initiative (IMI), a joint undertaking between the European Union and the pharmaceutical industry association EFPIA, funds projects aiming to speed up the development of better and safer medicines for patients. IMI's primary areas of focus include **Knowledge Management** and efforts are under way to establish synergies between the IMI Knowledge Management projects (e.g. EHR4CR, Open PHACTS (IMI Call 2), EMIF, eTRIKS (IMI Call 4)) and to capitalize on them. IMI promotes the development of Knowledge Manage-

Table 1 Best paper selection of articles for the IMIA Yearbook of Medical Informatics 2013 in the section 'Clinical Research Informatics'. The articles are listed in alphabetical order of the first author's surname.

Section
Clinical Research Informatics

- Eder J, Gottweis H, Zatloukal K. IT solutions for privacy protection in biobanking. Public Health Genomics 2012;15(5):254–62.
- Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. Nat Rev Genet 2012 Jun;13(6):395–405.
- Sittig DF, Hazlehurst BL, Brown J, Murphy S, Rosenman M, Tarczy-Hornoch P, Wilcox AB. A survey of informatics platforms that enable distributed comparative effectiveness research using multi-institutional heterogeneous clinical data. Med Care 2012 Jul;50 Suppl:S49–59.
- Tasneem A, Aberle L, Ananth H, Chakraborty S, Chiswell K, McCourt BJ, Pietrobon R. The database for aggregate analysis of ClinicalTrials.gov (AACT) and subsequent regrouping by clinical specialty. Plos One 2012;7(3):e33677.

ment platforms for interoperability issues between data repositories, for retrieving project data, and ensuring sustainability.

The goal of semantic interoperability is to be able to recognize and process semantically equivalent information even if data is heterogeneously represented with great variety by using different combinations of information models and/or terminologies/ontologies. A key challenge is to adopt a systematic methodology to develop a set of semantically unambiguous and context-neutral (to enable reuse) ontological models for clinical data while leveraging existing information model standards (e.g. 13606, openEHR, HL7 and CDISC), and biomedical terminologies and ontologies.

Another important issue is documentation of data quality. Although promising emerging projects lead to interesting results; expecting a goldmine of discoveries from existing federated EHRs and CDWs data is, at least today, premature since the difficulty in preparing this data for serious use in discovery is often largely underestimated. It remains challenging to assess data quality within electronic healthcare records prior to its storage and use within data warehouses, registries, or IT platforms dedicated to translational or comparative effectiveness research.

Acknowledgement

I would like to acknowledge the support of Martina Hutter and the reviewers in the selection process of the IMIA Yearbook.

References

1. Sim I, Carini S, Tu SW, Detwiler LT, Brinkley J, Mollah SA, et al. Ontology-based federated data access to human studies information. AMIA Annu Symp Proc 2012:856–65.
2. LinkedCT. LinkedCT Databrowse [Internet]. [cited 2013 May 5]. Available from: http://data.linkedct.org/
3. Tasneem A, Aberle L, Ananth H, Chakraborty S, Chiswell K, McCourt BJ, et al. The database for aggregate analysis of ClinicalTrials.gov (AACT) and subsequent regrouping by clinical specialty. Plos One 2012;7(3):e33677.
4. Bansal A, Chamberlain R, Karr S, Kwasa S, McLaughlin B, Nguyen B, et al. A 21 CFR Part 11 compliant graphically based electronic system for clinical research documentation. J Med Syst 2012 Jun;36(3):1661–72.
5. Murphy SN, Dubey A, Embi PJ, Harris PA, Richter BG, Turisco F, et al. Current state of information technologies for the clinical research enterprise across academic medical centers. Clin Transl Sci 2012 Jun;5(3):281–4.
6. Bettencourt-Silva J, De La Iglesia B, Donell S, Rayward-Smith V. On creating a patient-centric database from multiple Hospital Information Systems. Methods Inf Med 2012;51(3):210–20.
7. Kohane IS, Churchill SE, Murphy SN. A translational engine at the national scale: informatics for integrating biology and the bedside. J Am Med Inform Assoc 2012 Apr;19(2):181–5.
8. D'Amore JD, Sittig DF, Ness RB. How the continuity of care document can advance medical research and public health. Am J Public Health 2012 May;102(5):e1–4.
9. Eder J, Gottweis H, Zatloukal K. IT solutions for privacy protection in biobanking. Public Health Genomics 2012;15(5):254–62.
10. King T, Brankovic L, Gillard P. Perspectives of Australian adults about protecting the privacy of their health information in statistical databases. Int J Med Inform 2012 Apr;81(4):279–89.
11. Bemmels HR, Wolf SM, Van Ness B. Mapping the inputs, analyses, and outputs of biobank research systems to identify sources of incidental findings and individual research results for potential return to participants. Genet Med 2012 Apr;14(4):385–92.
12. Bledsoe MJ, Grizzle WE, Clark BJ, Zeps N. Practical implementation issues and challenges for biobanks in the return of individual research results. Genet Med 2012 Apr;14(4):478–83.
13. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. Nat Rev Genet 2012 Jun;13(6):395–405.
14. Sittig DF, Hazlehurst BL, Brown J, Murphy S, Rosenman M, Tarczy-Hornoch P, et al. A survey of informatics platforms that enable distributed comparative effectiveness research using multi-institutional heterogenous clinical data. Med Care 2012 Jul;50 Suppl:S49–59.
15. Geissbuhler A, Safran C, Buchan I, Bellazzi R, Labkoff S, Eilenberg K, et al. Trustworthy reuse of health data: A transnational perspective. Int J Med Inform 2013 Jan;82(1):1–9.
16. Rea S, Pathak J, Savova G, Oniki TA, Westberg L, Beebe CE, et al. Building a robust, scalable and standards-driven infrastructure for secondary use of EHR data: the SHARPn project. J Biomed Inform 2012 Aug;45(4):763–71.

Correspondence to:
Christel Daniel, MD, PhD
INSERM UMRS 872 équipe 20
CCS Patient – Assistance Publique – Hôpitaux de Paris
05 rue Santerre - 75 012 Paris
Tel: +33 1 48 04 20 47
E-mail: christel.daniel@crc.jussieu.fr

Appendix: Content Summaries of Selected Best Papers for the IMIA Yearbook 2013, Section 'Clinical Research Informatics'[1]

Eder J, Gottweis H, Zatloukal K

IT solutions for privacy protection in biobanking

Public Health Genomics 2012;15(5):254-62

The efficient access to biobank samples and data is nowadays critical for medical research. The BBMRI EU FP7 project aims to provide the largest European biobank infrastructure network for medical research. The major challenge faced is providing interoperability for both sample and biobank data. Interoperability of legal and ethical requirements is also an open research question. The paper presents the inadequacy between privacy needs (anonymisation) and long term research needs related to biobanking data. The lack of standardization of ethical and legal requirements across European countries is also a recurring issue.

The paper first focuses on describing the overall problem, at the European scale, of the question of privacy of medical data and shows that the more people know about biobank research, the more they are keen to participate in such research. The study also shows that unique global informed consent is not enough to cover all citizens' needs.

At the second stage, the paper describes how the access to biobank data can be integrated into the workflow that complies with researcher needs and patient protection. Authors then elaborate on innovative IT solutions to maintain the security of personal data and to increase the efficiency of access to biobank data by introducing disclosure filters within their data interoperability framework. The disclosure model describes the proprieties of the requester (institutions,

[1] The complete papers can be accessed in the Yearbook's full electronic version, provided that the article is freely accessible of that your institution has access to the respective journal.

countries, certifications), of the accessible item (information item, records, statistics, material) and of the provider (biobank, organization, network). The disclosure filter is a software component that uses the disclosure model and a set of rules to help in answering the question: *Who* is allowed to receive *what* from *whom* under *which* circumstances and *how*? The proposition should enable us to better adapt patient privacy to research needs. To be widely adopted, the approach should propose standardize workflows and models.

Jensen PB, Jensen LJ, Brunak S

Mining electronic health records: towards better research applications and clinical care

Nat Rev Genet 2012;13(6):395-405

Despite the great potential of EHR captured data to describe fine-grained patient phenotypes, researchers, who wish to analyze these patient data, are still faced with technical challenges of integrating heterogeneous data despite many standards and initiatives as presented in this review paper.

The authors first depict a general, yet exhaustive, view on the data being captured and stored in EHR systems today in the US: Pharmacy, laboratory, radiology and narrative information along with administrative data. Various NLP techniques are then elaborated that help in extracting phenotype data for research. Then, several use cases are presented as to use EHR data for clinical research, such as comorbidity analysis, correlating clinical features, pharmacovigilance, prediction of patient outcomes when data correlations are possible. Large population epidemiological databases can be coupled with EHR datasets to replicate and thus validate discoveries in smaller EHR cohorts.

Much clinical and genetic research critically depends on the identification and recruitment of large, phenotypically restrained case cohorts. Therefore, phenotype querying of structured data and NLP-encoded text in de-identified EHR databases is helpful. The i2b2 framework, eMERGE network, and EHR4CR are significant initiatives in that field.

The authors then elaborate on the potential outcomes of associating EHR enabled data with biobanking data, for example associating genetic variants with an increased risk of thromboembolism in patients with breast cancer treated with Tamoxifen. Pharmacogenomics also benefits from EHR integration of data by better adapting therapeutic choices since drug efficacy is influenced by genetic variation. But there are still limiting factors to address before being able to integrate data between bioinformatics, system biology, and medical informatics. Main issues identified are related to privacy and consent as well as interoperability.

Context specific generation of clinically actionable knowledge from EHR data offers great promise for better research applications and clinical care, but EHR data is still not sufficient for researchers as a single source of information to study cohorts or populations. Citizens' initiatives are more and more valuable in providing research with large scale data as they are becoming more engaged. In the meantime, genetic sequencing techniques evolve rapidly, and fine grained data are now captured in hospitals. All stakeholders have a joint objective to find better ways to help medical research progresses.

Sittig DF, Hazlehurst BL, Brown J, Murphy S, Rosenman M, Tarczy-Hornoch P, Wilcox AB

A survey of informatics platforms that enable distributed comparative effectiveness research using multi-institutional heterogeneous clinical data

Med Care 2012;50 Suppl:S49-59

The goal of comparative effectiveness research (CER) is to generate new evidence on the potential effectiveness, benefits, and harms of different treatments, diagnostics, preventions, and care models under "real world" conditions. CER requires aggregation and analysis of disparate data held by different institutions, each with its own representation of relevant events and accountabilities for protecting data as a matter of patient confidentiality and business operations. The authors offer a great review of 6 large-scale specific CER projects that are either developing informatics platforms or extending existing ones. The authors provide a comprehensive 8-dimension, socio-technical model for comparing the informatics platforms that are under development or in use in 6 large CER efforts. Six generic steps necessary in any distributed, multi-institutional CER project were identified: 1) De-identification of relevant data within health care transaction systems, 2) Extraction to a local data warehouse for staging, 3) Modeling data to enable common representations across multiple health systems, 4) Aggregation of data according to this common data model, 5) Analysis of data to address research questions, 6) Dissemination of study results. The authors concluded that CER requires data from inpatient and outpatient EHRs not only collected from billing and ancillary systems such as laboratory, pharmacy, and radiology but also from the text narrative of clinical encounters. In addition data documenting actual delivery of care that was ordered are critical in the context of CER. Moreover, CER informatics platforms must be able to extract and collect data from many different organizations - such as long-term care facilities, home and public health agencies - to compile a complete view of conditions, treatments, and individuals including attempts to reliably ascertain patients' socioeconomic status on a widespread basis. Such efforts require a community-wide master patient index. In addition, the most challenging issue in any project of multi-institutional CER platform is to provide solutions for mapping data types to standardized clinical representations. At last, the social, legal, ethical, and political challenges involved in setting up and conducting large, multi-institutional CER projects must not be underestimated. Therefore, in addition to providing the technical infrastructure required to collect, standardize, normalize, and analyze disparate data, informatics platforms must conform to local organizations' internal governance and IRBs' rules and regulations as well as existing state and country guidelines.

Tasneem A, Aberle L, Ananth H, Chakraborty S, Chiswell K, McCourt BJ, Pietrobon R

The database for aggregate analysis of ClinicalTrials.gov (AACT) and subsequent regrouping by clinical specialty

PloS One 2012;7(3):e33677

ClinicalTrials.gov (www.ClinicalTrials.gov) is the largest registry of human clinical research studies. Hosted by the National Library of Medicine (NLM) at the National Institutes of Health (NIH) in collaboration with the U.S. Food and Drug Administration (FDA) this registry currently contains over 100,000 research studies covering the full range of clinical trials conducted in more than 170 countries and includes a broad group of trial sponsors. ClinicalTrials.gov has a regulatory mandate and is increasingly used both by medical professionals and the public.

This paper reports on a unique effort to prepare and maintain a publicly accessible analysis dataset derived from ClinicalTrials.gov content - the Aggregate Analysis of ClinicalTrials (AACT) database - and to extend its utility by means of associated clinical specialty taxonomy designed to support research policy analyses. The development of the (AACT) database is based on a core physical data model designed to accommodate both data management and curating needs with implementation of quality metrics. In addition the linking of study metadata to additional resources, such as the Medical Subject Headings (MeSH) thesaurus enhances the retrieval and analysis processes. A dataset comprising 96,346 clinical studies registered in the ClinicalTrials.gov registry from September 27, 2007 to September 27, 2010 was downloaded. Data submitters are requested to provide condition and keywords data as MeSH terms when registering a study and, additionally, an NLM algorithm evaluates studies and applies MeSH terms to clinical trials. The authors also report on the methodology they developed to use MeSH terms associated with each study to regroup studies from ClinicalTrials.gov by clinical specialties as designated by the Department of Health and Human Services. A total of 18,491 MeSH disease condition IDs from interventional studies associated with 9,031 MeSH terms were reviewed for specialty classification. The authors evaluated the accuracy of the classification algorithm in cardiology, oncology, and mental health trials by comparing it with classifications provided by two independent clinicians from each relevant specialty. The authors concluded that the development of a version of ClinicalTrials.gov data optimized for aggregate analysis and the associated clinical specialty taxonomy enables the comparative evaluation across specialties of a number of trials, the resulting publications, and distribution of evidence levels. More generally speaking the AACT database may support the definition of evidence-based strategic plans in relation to national and regional biomedical research policies. The authors plan to enhance the time and resource intensive process for development and curating of the clinical specialty by potentially distributing this task across a large, open curating community. They also, interestingly, intend to represent the curated database and the accompanying clinical specialty taxonomy using the Resource Description Framework (RDF) to enable merging with the linkedCT version of the Clinical- Trials.gov dataset and other RDF resources.