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 Section on Clinical Research Informatics
INSERM UM RS 1142, Paris, France

Summary

Objective: To select and summarize key contributions to current research in the field of Clinical Research Informatics (CRI).

Method: A bibliographic search using a combination of MeSH and free terms search over PubMed was performed followed by a blinded review.

Results: The review process resulted in the selection of four papers illustrating various aspects of current research efforts in the area of CRI. The first paper tackles the challenge of extracting accurate phenotypes from Electronic Healthcare Records (EHRs). Privacy protection within shared de-identified, patient-level research databases is the focus of the second selected paper. Two other papers exemplify the growing role of formal representation of clinical data - in metadata repositories - and knowledge – in ontologies - for supporting the process of reusing data for clinical research.

Conclusions: The selected articles demonstrate how concrete platforms are currently achieving interoperability across clinical research and care domains and have reached the evaluation phase. When EHRs linked to genetic data have the potential to shift the research focus from research driven patient recruitment to phenotyping in large population, a key issue is to lower patient re-identification risks for biomedical research databases. Current research illustrates the potential of knowledge engineering to support, in the coming years, the scientific lifecycle of clinical research.

Keywords

Medical informatics, International Medical Informatics Association, yearbook, Biomedical Research, Clinical research, Nursing Research, Pharmacovigilance, Patient Selection, Phenotyping

Yearb Med Inform 2014:224-7
http://dx.doi.org/10.15265/IY-2014-0040
Published online August 15, 2014

Introduction

Clinical Research Informatics (CRI) involves the "use of informatics in the discovery and management of new knowledge relating to health and disease"[1]. It includes management of information related to clinical trials and involves informatics related to secondary use of clinical data for research.

The goal of this section is to provide an overview of research trends and of "best" papers published in the past year that demonstrate excellent CRI relevant research.

The explosion of available data for biomedical research enabled by the rise of genomics and phenomics generated data is now being confronted to a promising new field for data acquisition: the exposome coming from patient generated data via sensors or manual data entry [10]. The equation "Phenotype=Genotype×Environment" poses enormous challenges to current information systems for biomedical research and the question of big data - the special topic of the 2014 Yearbook - is consequently emerging in CRI. Though the use of big data technology and cloud computing is at its infancy in CRI, proofs of concepts from early adopters of map reduce, Hadoop, NoSQL data bases, cloud computing, illustrate how these technologies are likely to be used to improve the performance of health care IT systems very soon [13,17].

The comprehensive review of the CRI section focused on various categories of CRI activity: data and knowledge management, clinical data re-use for research, methods in CRI, policy and perspectives, security, confidentiality, and regulatory issues.

[1] http://www.amia.org/applications-informatics/clinical-research-informatics [Accessed: 24/05/2014].

About the Paper Selection

A comprehensive review of published articles in 2013 addressing a wide range of issues for clinical research informatics was conducted. The selection was performed by querying PubMed with predefined keywords and yielded a total of 898 references. From this original set, a first subset of 781 references that were in the scope of CRI, were blindly reviewed by the two section editors and considered according to their relevancy to the field. The two reviews were merged, yielding 159 references from which the two section editors consensually retained 15 articles to be submitted to peer-review following the IMIA Yearbook process. Table 1 lists the four papers that were finally selected. A content summary of the selected papers can be found in the appendix of this synopsis.

Conclusion and Outlook

Clinical research is on the edge of a new era, in which electronic health records (EHRs) are gaining an important novel supporting role to produce phenotype data [2]. As an increased amount of data is produced in EHRs, the of data gathered in process of providing care for reseach is re-used to facilitate patient identification for new clinical trials [3,7], to collect care data within clinical research setting [5] or to be combined with DNA biorepositories for identifying accurately phenotyped cases and controls for large-scale genomic studies [12] or to conduct pharmacogenetic studies [14,15]. Utilizing the availability of patient data from federated EHR systems in many different sites, as well as in international multilingual

settings is still challenging [2]. Although promising clinical data re-use for research is being enabled through the building of major emerging research infrastructures such as SHARPn [16], i2b2-SHRINE [6,11], EHR4CR [2], limitations and new issues arise [5,8].

The first selected paper illustrates that EHRs are imperfect objects given the challenge to extract accurate phenotypes from them. In the context of the eMERGE network aiming at re-using phenotypic and genotypic data from EMRs for large scale genomic studies, the authors describe the validation of specific phenotype extraction algorithms. The validation process contributes not only to evaluate but also to enhance the accuracy of adaptative data extraction techniques compatible with underlying EHR systems [12]. In [4], the general problems encountered when re-using EHR data are depicted: i) inaccurate or incorrect data, ii) fragmented patient stories, iii) data encoded or transformed for other purposes than research, iv) unprocessable free text data, v) data of unknown provenance, vi) data granularity that does not match research needs. Additional studies focus on the caveats of using billing claims for clinical research or epidemiology [9].

Lowering re-identification risk in research de-identified database is the topic explored by the second selected paper [1].

The other papers focus on data and knowledge representation in Clinical Research. In the third paper of the selection, the authors describe how the Ontology of Clinical Research (OCRe) can represent study protocol information and benefits to CRI applications supporting the entire scientific lifecycle of both interventional or observational human studies [18]. The fourth paper reports on a unique effort to build an innovative semantic meta-repository to encourage the re-use of data elements across patient care and clinical research domains [19]. Metadata repositories are key elements to reduce interoperability burden of EHRs in healthcare.

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

Table 1 Best paper selection of articles for the IMIA Yearbook of Medical Informatics 2014 in the section 'Clinical Research Informatics'. The articles are listed in alphabetical order of the first author's surname.

Section
Clinical Research Informatics
▪ Atreya RV, Smith JC, McCoy AB, Malin B, Miller RA. Reducing patient re-identification risk for laboratory results within research datasets. J Am Med Inform Assoc 2013;20(1):95-101.
▪ Newton KM, Peissig PL, Kho AN, Bielinski SJ, Berg RL, Choudhary V, Basford M, Chute CG, Kullo IJ, Li R, Pacheco JA, Rasmussen LV, Spangler L, Denny JC. Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network. J Am Med Inform Assoc 2013;20(e1):e147-54.
▪ Sim I, Tu SW, Carini S, Lehmann HP, Pollock BH, Peleg M, Wittkowski KM. The Ontology of Clinical Research (OCRe): An informatics foundation for the science of clinical research. J Biomed Inform 2013;S1532-0464(13)00179-2.
▪ Sinaci AA, Laleci Erturkmen GB. A federated semantic metadata registry framework for enabling interoperability across clinical research and care domains. J Biomed Inform 2013;46(5):784-94.

References

1. Atreya RV, Smith JC, McCoy AB, Malin B, Miller RA. Reducing patient re-identification risk for laboratory results within research datasets. J Am Med Inform Assoc 2013;20(1):95−101.
2. Coorevits P, Sundgren M, Klein GO, Bahr A, Claerhout B, Daniel C, et al. Electronic health records: new opportunities for clinical research. J Intern Med 2013;274(6):547−60.
3. Fernández-Breis JT, Maldonado JA, Marcos M, Legaz-García MDC, Moner D, Torres-Sospedra J, et al. Leveraging electronic healthcare record standards and semantic web technologies for the identification of patient cohorts. J Am Med Inform Assoc 2013;20(e2):e288−96.
4. Hersh WR, Weiner MG, Embi PJ, Logan JR, Payne PRO, Bernstam EV, et al. Caveats for the use of operational electronic health record data in comparative effectiveness research. Med Care 2013;51(8 Suppl 3):S30−7.
5. Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. J Am Med Inform Assoc 2013;20(1):117−21.
6. Klann JG, Murphy SN. Computing health quality measures using Informatics for Integrating Biology and the Bedside. J Med Internet Res 2013;15(4):e75.
7. Köpcke F, Kraus S, Scholler A, Nau C, Schüttler J, Prokosch H-U, et al. Secondary use of routinely collected patient data in a clinical trial: an evaluation of the effects on patient recruitment and data acquisition. Int J Med Inform 2013;82(3):185−92.
8. Kukafka R, Allegrante JP, Khan S, Bigger JT, Johnson SB. Understanding facilitators and barriers to reengineering the clinical research enterprise in community-based practice settings. Contemp Clin Trials 2013;36(1):166−74.
9. Li AH, Kim SJ, Rangrej J, Scales DC, Shariff S, Redelmeier DA, et al. Validity of physician billing claims to identify deceased organ donors in large healthcare databases. PLoS One 2013;8(8):e70825.
10. Martin Sanchez F, Gray K, Bellazzi R, Lopez-Campos G. Exposome informatics: considerations for the design of future biomedical research information systems. J Am Med Inform Assoc 2014 May-Jun;21(3):386-90.
11. McMurry AJ, Murphy SN, MacFadden D, Weber G, Simons WW, Orechia J, et al. SHRINE: enabling nationally scalable multi-site disease studies. PLoS One 2013;8(3):e55811.
12. Newton KM, Peissig PL, Kho AN, Bielinski SJ, Berg RL, Choudhary V, et al. Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network. J Am Med Inform Assoc 2013;20(e1):e147−54.
13. O'Driscoll A, Daugelaite J, Sleator RD. « Big data », Hadoop and cloud computing in genomics. J Biomed Inform 2013;46(5):774−81.
14. O'Meara H, Carr DF, Evely J, Hobbs M, McCann G, van Staa T, et al. Electronic Health Records For Biological Sample Collection: Feasibility Study Of Statin-Induced Myopathy Using The Clinical Practice Research Datalink. Br J Clin Pharmacol 2014 May;77(5):831-8.
15. Patel VN, Kaelber DC. Using aggregated, de-identified electronic health record data for multivariate pharmacosurveillance: A case study of azathioprine. J Biomed Inform 2013 Oct 28. pii: S1532-0464(13)00161-5.
16. Pathak J, Bailey KR, Beebe CE, Bethard S, Carrell DC, Chen PJ, et al. Normalization and standardization of electronic health records for high-throughput phenotyping: the SHARPn consortium. J Am Med Inform Assoc 2013;20(e2):e341−8.
17. Sahoo SS, Jayapandian C, Garg G, Kaffashi F, Chung S, Bozorgi A, et al. Heart beats in the cloud: distributed analysis of electrophysiological « big data » using cloud computing for epilepsy clinical research. J Am Med Inform Assoc 2014 Mar-Apr;21(2):263-71.
18. Sim I, Tu SW, Carini S, Lehmann HP, Pollock BH, Peleg M, et al. The Ontology of Clinical Research (OCRe): An informatics foundation for the science of clinical research. J Biomed Inform. 2013 Nov 13. pii: S1532-0464(13)00179-2.
19. Sinaci AA, Laleci Erturkmen GB. A federated semantic metadata registry framework for enabling interoperability across clinical research and care domains. J Biomed Inform 2013;46(5):784−94.
20. Toh S, Gagne JJ, Rassen JA, Fireman BH, Kulldorff M, Brown JS. Confounding adjustment in comparative effectiveness research conducted within distributed research networks. Med Care 2013;51(8 Suppl 3):S4−10.

Correspondence to:
Christel Daniel, MD, PhD
INSERM UMRS 1142
CCS Patient – Assistance Publique – Hôpitaux de Paris
05 rue Santerre - 75 012 PARIS
Tel: 33 1 48 04 20 29
E-mail: christel.daniel@crc.jussieu.fr

Appendix: Content Summaries of Selected Best Papers for the IMIA Yearbook 2014, Section 'Clinical Research Informatics'

Newton KM, Peissig PL, Kho AN, Bielinski SJ, Berg RL, Choudhary V, Basford M, Chute CG, Kullo IJ, Li R, Pacheco JA, Rasmussen LV, Spangler L, Denny JC

Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network

J Am Med Inform Assoc 2013;20(e1):e147-54

The aim of the Electronic Medical Records and Genomics (eMERGE) network, created and funded by the National Human Genome Research Institute is to develop, disseminate, and apply approaches to combine DNA biorepositories with EMR systems for large-scale genomic studies. A successful eMERGE project demonstrated that Electronic medical records (EMRs) hold large numbers of clinical phenotypes such as disease (cases) and nondisease (controls), and quantitative traits of medical importance, with sufficient validity to power genome-wide association studies (GWAS) and other emerging types of genetic studies. However, EMRs, being primarily designed to document patient–provider interactions and to generate billing documentation, are imperfect instruments for extracting accurate phenotypes from them and, therefore, phenotype validation across multiple EMR systems, in different institutions is a critical step in characterizing the types of phenotypes that the EMR can reliably provide, and establishing the utility of the EMR for genetic studies.

The authors report the creation and validation of 13 EMR-based phenotypes from the eMERGE studies that were executed within five sites of the network.

The first phenotypes were selected based on the investigators' expertise and interests, the scientific importance of GWAS for the phenotype, and the feasibility of clearly identifying the phenotype within the EMR. They are related to different disease areas such as cataract, dementia, diabetes, retinopathy, hypertension, peripheral arterial disease, primary hypothtroidism, etc.

To create phenotype definitions with a reasonable likelihood of success, the underlying logic used, including constituent data elements (encoded variables e.g. ICD-9-CM and Current Procedural Terminology (CPT)-4 codes) nested in Boolean logic associated to temporal operators, was carefully analyzed and represented as «pseudocode» providing a detailed map for data extraction to the sites.

Validation reviews were accomplished via manual record review of paper or electronic records to confirm the correctness of the variables used to create the phenotype algorithm. The performance of a phenotype algorithm was measured by its capability of identifying cases and controls meeting eligibility and quantified by the positive predictive value (PV+) for being a case, for being a control or for meeting algorithm eligibility criteria. Among 51 algorithm reviews across five sites, almost three quarters of the reviews yielded PV+ values of 90% or greater, and only three reviews yielded PV+ values less than 80%. Phenotype algorithms with validation metrics are publicly available at http://www.PheKB.org.

This work is a key contribution to the community adressing the challenge of the deployment of EMR-derived phenotype algorithms. The authors demonstrate that phenotype validation is a worthwhile process that not only measures algorithms performance but also strengthens their definition and accuracy and enhances their inter-institutional sharing. The authors also share lessons learned and provide useful guidance for the set up of the validation process.

Atreya RV, Smith JC, McCoy AB, Malin B, Miller RA

Reducing patient re-identification risk for laboratory results within research datasets

J Am Med Inform Assoc 2013;20(1):95-101

Routinely collected health data are becoming useful to researchers and public authorities for building surveillance networks, epidemiology studies, or pharmacovigilance. EHR data re-use faces several issues, such as data quality or data misinterpretation as data is context dependent. Preserving patient privacy is always an issue, even when data are not presenting any personal information such as sex, age, or birth city. We often refer to de-identified datasets. The more de-identified datasets are available, the more risks of re-identification occurs but current regulations do not mandate complete elimination of re-identification risk.

The paper presents an original method to lower re-identification risk in a laboratory result set by applying a privacy model on additive random noise to preserve the clinical meaning of the data. The approach consisted in i) evaluating the uniquely distinguishing nature of the laboratory dataset and elaborating a threat model and ii) evaluating perturbation methods to lower re-identification risks while minimizing alterations in the clinical meaning of the data.

During the first phase, authors simulated an attack upon a NIH-funded Vanderbilt TIME database including 61 280 adult inpatients hospitalizations and comprising 8,5 million laboratory results sets. The study assessed whether the attack could re-identify patients' pseudo-identifiers. They could identify unique patients for sequences of five to six results (glucose, calcium, LymAbs, cholesterol, SGPT and creatine kinase) for more than 95% of cases.

The authors then implemented a simple perturbation algorithm to the dataset and an expert driven perturbation algorithm. Whereas the simple perturbation algorithm applies random offsets to blur the results, the expert driven algorithm constrain the offsets to the same clinical meaning. For example, patients with severe infections or leukemia can have values of white blood cell results > 50 000.

The perturbation algorithms were then compared using a rank-based re-identification algorithm as previously used during the first phase. The expert based perturbation algorithm is showing a slight re-identification risk increase compared to the simple algorithm. However, it was much better in maintaining the meaning of the result. The approach shows new possibilities to enable the constitution of «open for research» patient datasets and

represents an important contribution in the field. Such expert driven perturbation algorithms are key to facilitate the opening of large de-identified health datasets for public health and research.

Sim I, Tu SW, Carini S, Lehmann HP, Pollock BH, Peleg M, Wittkowski KM

The Ontology of Clinical Research (OCRe): An informatics foundation for the science of clinical research

J Biomed Inform 2013;S1532-0464(13)00179-2

Clinical research informatics (CRI) aims at supporting the scientific lifecycle of clinical research including ***design phase*** - review and interpretation of results of previous studies to refine a scientific question and design of a new study ; ***execution phase*** - study execution and results reporting ; and ***application phase*** - interpretation and application of the results to clinical care or policy.

This paper describes how the Ontology of Clinical Research (OCRe) can capture detailed abstract study protocol information and yield numerous benefits along the entire scientific lifecycle of human both interventional or observational studies. OCRe models the entities and relationships of study designs to serve as a common semantics for computational approaches to the design and analysis of human studies.

The authors used an hybrid modeling approach. They first built OCRe using the Web Ontology Language (OWL) for capturing the formal semantics of the core structures of a study protocol including the design typology, design features, subgroups in the study, and interventions and exposures. Then, regarding some aspects of study design – eligibility criteria in particular – that are difficult to formulate ontologically, they defined an information model called ERGO Annotation that allows annotating the criteria more simply but still capturing their essential meaning. The ERGO annotations are data structures that allow the application of natural language processing (NLP) methods to automate the recognition of coded concepts and relations in free-text eligibility criteria and that can be transformed into computable formats.

The second part of the paper descibes how CRI applications grounded on OCRe can support the different clinical research activities.

For the ***design phase***, OCRe-based models covers a broader scope than existing published study design taxonomies, including the Cochrane Collaboration's and Hartling's taxonomies. In order to support critical appraisal of the study design, OCRe provides a formal representation of study validity, comparability of comparison groups, existence and nature of follow-up bias. OCRe based descriptions of the design of past and ongoing studies can ensure the completeness and internal coherence of study instances and facilitates their classification in term of their scientific and design features. OCRe can therefore serve as a well-formed semantics for federated data querying of study designs within or between institutions as well as of trial registers and other trial databases (e.g., AHRQ's Systematic Review Data Repository, Cochrane's Central Register of Studies, Human Studies Database project). Investigators may pose a scientific question, review the results of previous studies to refine a scientific question, identify biases or potential clinical confounders that should be taken into account at the design phase, and visualize past recruitment patterns for sample size calculations.

In addition, a computable representation of eligibility criteria that can be then matched against EHR data is useful, not only for cohort identification and eligibility determination implemented in clinical trial recruitment support systems during the ***execution phase,*** but also for determining the applicability of studies to a target patient or population during the ***application phase***.

In contrast to other major foundational models for clinical research informatics (e.g., BRIDG, CDISC, OBX, ClinicalTrials.gov XSD), OCRe takes a logic-oriented ontological modeling approach. Moreover, in terms of scope, unlike BRIDG and CDISC SDTM which address operational and administrative needs, OCRe also attempt to model study validity, confounding, and bias needed for assessing study design strength.

Sinaci AA, Laleci Erturkmen GB

A federated semantic metadata registry framework for enabling interoperability across clinical research and care domains

J Biomed Inform 2013;46(5):784-94

Despite the wider use of standards to represent and code data in medical databases, the bridge between those standards that are often domain dependent is still missing and real interoperability between care and clinical research domains is still an issue. EHRs are now presenting a growing potential for exploitation in clinical research setting. Unfortunately, EHR standards are not equally implemented, and bridges between clinical data representations for EHR (HL7, openEHR) are barely linked to clinical research ones (CDISC). Moreover, electronic data capture (EDC) softwares are now much deployed for registries and cohorts and data are often not standardized more than by defining common data elements between them when possible (CDEs).

Efforts are made to try to harmonize data representations (BRIDG model between HL7 and CDISC) but they do not cover the broad scope of interoperability problems. Coordination for new datasets creation is needed and the paper present a novel architecture of semantic metadata registry by enabling a better formalisation of data elements described in metadata repositories. The goal of the framework proposed is to limit the creation of known data elements across several metadata repositories by proposing a mechanism to link data elements between metadata repositories through Linked Data principles.

The proposition rely on ISO/IEC 11179 standard to represent metadata upon which an extension is built to achieve federated metadata management. Each CDE is uniquely identified by a URI, it is dereferenceable and has a RDF definition. Each CDE can be linked to another MDR CDE through some relation defined in a OWL ontology based on the ISO/IEC 11179 meta-model created by the authors.

This work is currently used in the EU FP7 SALUS project. It enables the population of data coming from HL7 Continuity of Care Documents (CCD) based content models through the Federated Query Service to CDISC (Operational Data Model) ODM document annotated with CDISC Study Data Tabulation Model (SDTM) CDEs.

Metadata repositories are important tools to encourage data elements re-use across both EHRs and EDC systems used in clinical research or epidemiology. But many are now built and now require linking them together to further enable data elements re-use. The proposed method is innovative and should foster interoperability between care and research domains.